Algunas Aplicaciones Machine Learning en Ciencias Sociales

Germán Rosati (CONICET-UNSAM, PIMSA)

Sobre mí

Sociología, Machine Learning, Métodos Cuantitativos

Ahora…

- Data Scientist en QSocialNow
- Profesor “Métodos Cuantitativos” UNSAM
- Investigador Asistente CONICET (Esperando el milagro)

Antes...

- Coordinador Data Science / IA Digital House
- Becario doctoral (CONICET)
- Investigador invitado Freie Universität Berlin
- Analista Experto de Datos (MTEySS)
- Data Scientist FreeLance (BID, Banco Mundial, PNUD, OIT, Universidades, consultoras)

Algunos proyectos

ISSN: 1852-4222

Saberes

REVISTA DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

INICIO ACERCA DE INICIAR SESIÓN REGISTRARSE BUSCAR ACTUAL ARCHIVOS AVISOS

Inicio > Vol. 9, Núm. 1 (2017) > **Rosati**

Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning. Aplicación en la Encuesta Permanente de Hogares (EPH)

Germán Federico Rosati

Resumen

El presente documento se propone exponer los avances realizados en la construcción de un modelo de imputación de valores perdidos y sin respuesta para las variables de ingreso en encuestas a hogares. Se presentará la propuesta metodológica general y los resultados de las pruebas realizadas. Se evalúan dos tipos de modelos de imputación de datos perdidos: 1) el método hot-deck (ampliamente utilizado por relevamientos importantes en el Sistema Estadístico Nacional, tales como la Encuesta Permanente de Hogares y la Encuesta Anual de Hogares de la Ciudad de Buenos Aires) y 2) un ensamble de modelos de regresión LASSO (Least Absolute Shrinkage and Selection Operator). El mismo se basa en la generación de múltiples modelos de regresión LASSO a través del algoritmo bagging y de su agregación para la generación de la imputación final. En la primera y segunda parte del documento plantea el problema de forma más específica y se pasa revista a los principales mecanismos de generación de los valores perdidos y las implicancias que los mismos tienen al momento de generar modelos de imputación. En el tercer apartado se reseñan los métodos de imputación más habitualmente utilizados, enfatizando sus ventajas y limitaciones. En la cuarta parte, se desarrollan los fundamentos teóricos y metodológicos de las dos técnicas de imputación propuestas. Finalmente, en la quinta sección, se presentan algunos resultados de la aplicación de los métodos propuestos a datos de la Encuesta Permanente de Hogares.

SciELO latindex NB REDIB Red Iberoamericana Dialnet

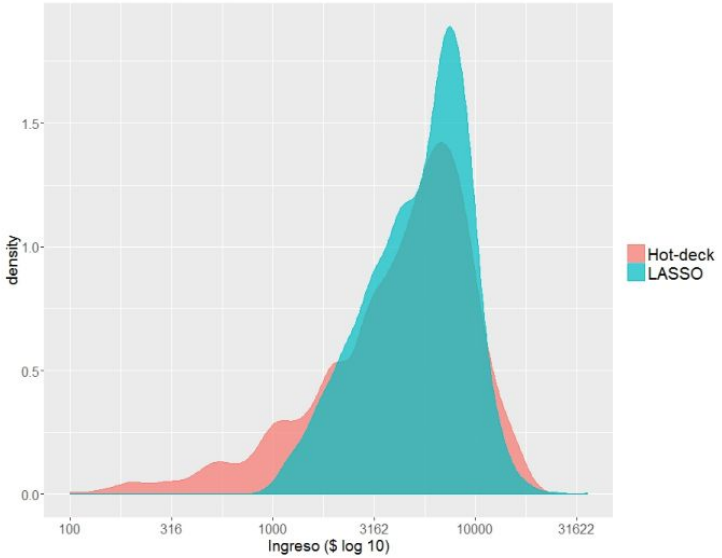
Ensemble Learning para imputación de datos perdidos

[Rosati, 2017]

Algunos proyectos

Ensemble Learning para imputación de datos perdidos en variables de ingreso

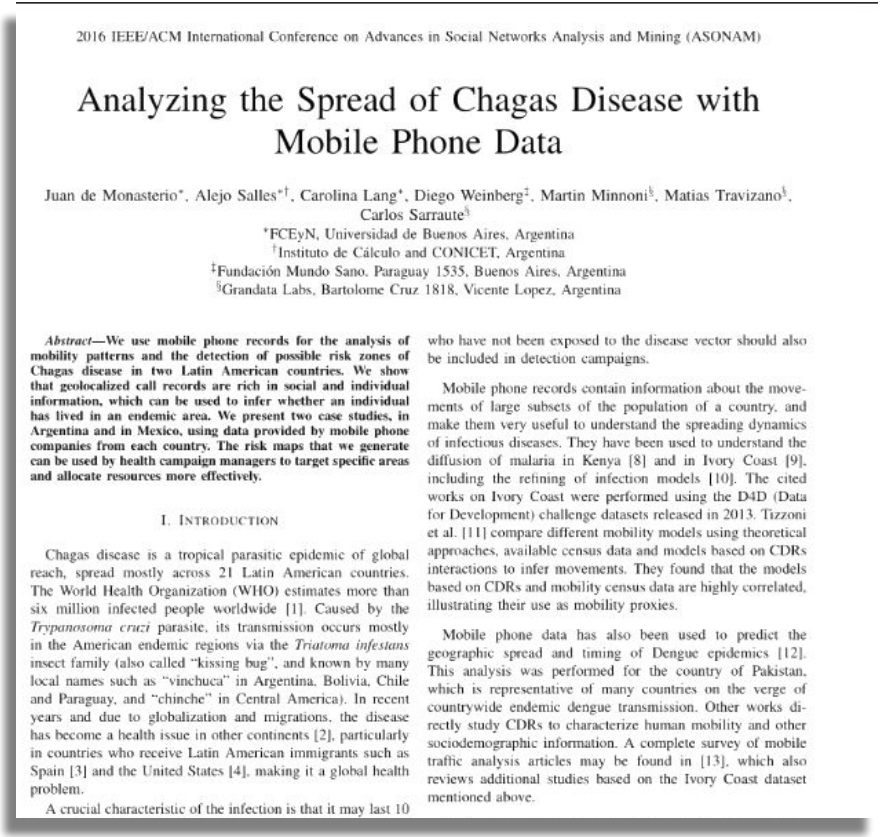
- Mejorar performance de los métodos de imputación “tradicionales” (aka Hot-Deck)
- Ejercicio 1. Bagging de regresiones regularizadas por LASSO
- Mejora del 20% en RMSE respecto a Hot-Deck
- Exploración de otros métodos...



Algunos proyectos

Big Data y Chagas

- Fundación Bunge y Born
- GranData
- Mundo Sano



2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Analyzing the Spread of Chagas Disease with Mobile Phone Data

Juan de Monasterio*, Alejo Salles*†, Carolina Lang*, Diego Weinberg‡, Martin Minnoni§, Matias Travizano§, Carlos Sarraute§
*FCEyN, Universidad de Buenos Aires, Argentina
†Instituto de Cálculo and CONICET, Argentina
‡Fundación Mundo Sano, Paraguay 1535, Buenos Aires, Argentina
§Grandata Labs, Bartolome Cruz 1818, Vicente Lopez, Argentina

***Abstract*—We use mobile phone records for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. We show that geolocalized call records are rich in social and individual information, which can be used to infer whether an individual has lived in an endemic area. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country. The risk maps that we generate can be used by health campaign managers to target specific areas and allocate resources more effectively.**

I. Introduction

Chagas disease is a tropical parasitic epidemic of global reach, spread mostly across 21 Latin American countries. The World Health Organization (WHO) estimates more than six million infected people worldwide [1]. Caused by the *Trypanosoma cruzi* parasite, its transmission occurs mostly in the American endemic regions via the *Triatoma infestans* insect family (also called "kissing bug", and known by many local names such as "vinchuca" in Argentina, Bolivia, Chile and Paraguay, and "chinche" in Central America). In recent years and due to globalization and migrations, the disease has become a health issue in other continents [2], particularly in countries who receive Latin American immigrants such as Spain [3] and the United States [4], making it a global health problem.

A crucial characteristic of the infection is that it may last 10 who have not been exposed to the disease vector should also be included in detection campaigns.

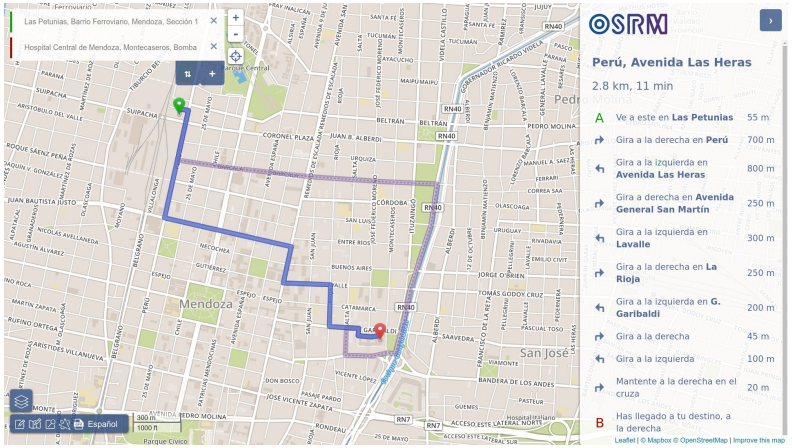
Mobile phone records contain information about the movements of large subsets of the population of a country, and make them very useful to understand the spreading dynamics of infectious diseases. They have been used to understand the diffusion of malaria in Kenya [8] and in Ivory Coast [9], including the refining of infection models [10]. The cited works on Ivory Coast were performed using the D4D (Data for Development) challenge datasets released in 2013. Tizzoni et al. [11] compare different mobility models using theoretical approaches, available census data and models based on CDRs interactions to infer movements. They found that the models based on CDRs and mobility census data are highly correlated, illustrating their use as mobility proxies.

Mobile phone data has also been used to predict the geographic spread and timing of Dengue epidemics [12]. This analysis was performed for the country of Pakistan, which is representative of many countries on the verge of countrywide endemic dengue transmission. Other works directly study CDRs to characterize human mobility and other sociodemographic information. A complete survey of mobile traffic analysis articles may be found in [13], which also reviews additional studies based on the Ivory Coast dataset mentioned above.

Algunos proyectos

Big Data y Chagas

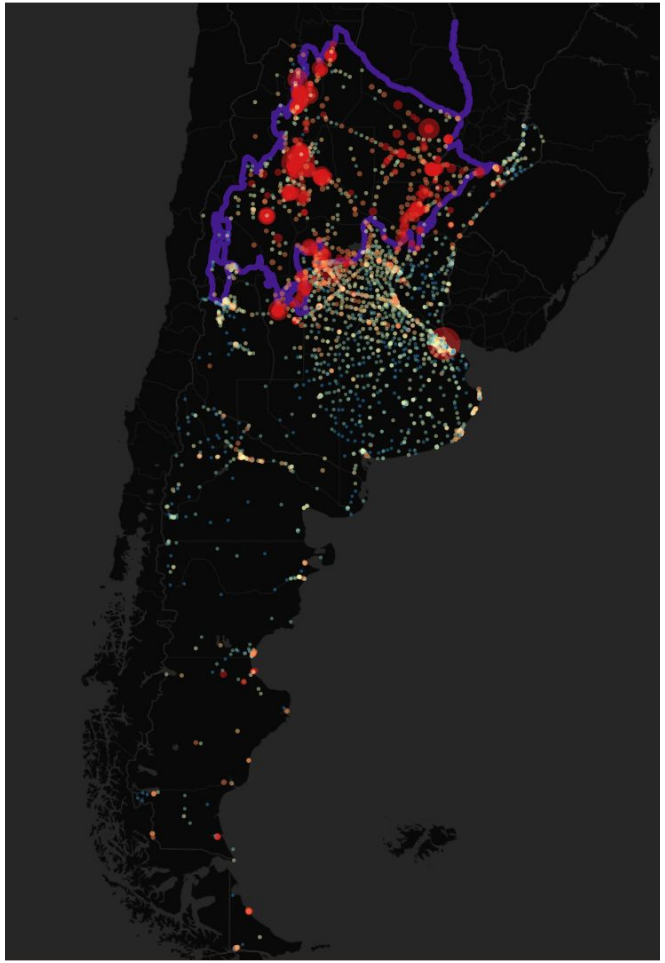
- Mapa de Vulnerabilidad Sanitaria
    - NSE (NBI, Nivel educativo, CALMAT => autoencoder)
    - Accesibilidad a Centros de Salud
        - Dataset centros de salud
        - Distancias



Algunos proyectos

Big Data y Chagas

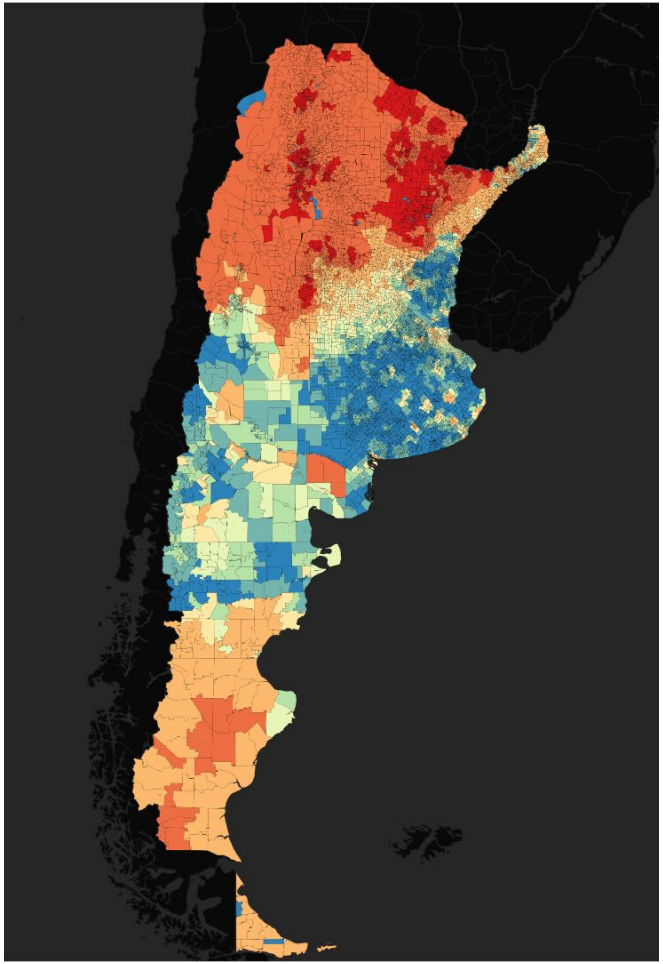
- Mapa de afinidad con región endémica
  - CDR’s
  - Calidad de Vivienda



Algunos proyectos

Big Data y Chagas

- Mapa combinado: Vulnerabilidad Sanitaria + Afinidad

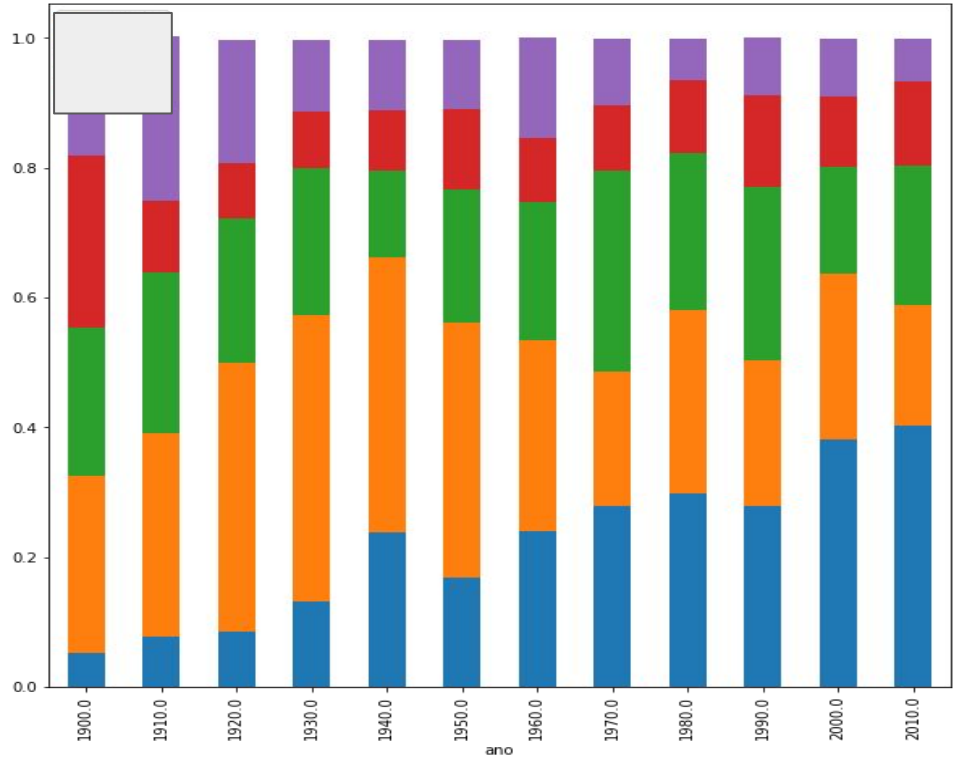


Algunos proyectos

Detección de tópicos en letras de tango

Evolución de los tópicos (mediana de la composición de los documentos) 1880-2010

Topic 1
Topic 2
Topic 3
Topic 4
Topic 5



Algunos proyectos

Detección de tópicos en letras de tango

Composición de tópicos en Malena (Manzi)

Título	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Malena	0.14	**0.55**	0	**0.3**	0

Malena canta el tango como ninguna
y en cada verso pone su corazón.
A yuyo del suburbio su voz perfuma,
Malena tiene pena de bandoneón.
Tal vez allá en la infancia su voz de alondra
tomó ese tono oscuro de callejón,
o acaso aquel romance que sólo nombra
cuando se pone triste con el alcohol.
Malena canta el tango con voz de sombra,
Malena tiene pena de bandoneón.
(...)
Tus ojos son oscuros como el olvido,
tus labios apretados como el rencor,
tus manos dos palomas que sienten frío,
tus venas tienen sangre de bandoneón.
(…)

symmetry topics follow +

Illustration by Sandbox Studio, Chicago with Corinne Mucha

Machine learning proliferates in particle physics

08/01/18 | By Manuel Gnida

A new review in *Nature* chronicles the many ways machine learning is popping up in particle physics research.

Experiments at the Large Hadron Collider produce about a million gigabytes of data every second. Even after reduction and compression, the data amassed in just one hour at the LHC is

nature.com > nature > technology features > article

MENU nature International Journal of science

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

TECHNOLOGY FEATURE · 20 FEBRUARY 2018 · CORRECTION 07 MARCH 2018

Deep learning for biology

A popular artificial-intelligence method provides a powerful tool for surveying and classifying biological data. But for the uninitiated, the technology poses significant difficulties.

nature.com > nature > news > article

MENU nature International Journal of science

Search E-alert

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

NEWS · 28 MARCH 2018

Need to make a molecule? Ask this AI for instructions

Artificial-intelligence tool that has digested nearly every reaction ever performed could transform chemistry.

Machine Learning y Ciencias Sociales

- Avances relevantes para las Ciencias Sociales
  - Interpretable ML
    - Cajas Negras
    - Sesgo Algorítimico
    - Posibilidad de Interpretar resultados
  - Masificación de APIs “amigables”
  - Aprendizaje No Supervisado
  - Text Mining, Natural Language Processing

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como usuarios o consumidores**
  - =~ a la que tenemos con la estadística
  - Usuarios de métodos, APIS, etc.

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como productores**
  - Planteo de nuevos problemas relevantes
  - Reformulación de nuevos métodos en base a problema

¿Preguntas?

@Crst_C

german.rosati@gmail.com

https://gefero.github.io/