Imputación de datos perdidos mediante técnicas de *Machine Learning*

Un experimento usando la Encuesta Permanente de Hogares

Germán Rosati
german.rosati@gmail.com

CONICET/ IDAES-UNSAM / PIMSA / UNTREF

05 de Septiembre de 2019

Hoja de ruta

1. ¿Qué es y como se genera un dato perdido?
2. ¿Cómo lidiar con los datos perdidos?
   - Técnicas tradicionales (imputación simpe)
   - Técnicas basadas en *Machine Learning*
3. Metodología de imputación utilizada
4. Resultados y discusión

¿Qué es un valor perdido?

- Valor del que se carece una dato válido en la variable observada
- Problema generalizado en investigaciones por encuestas
- Problema cada vez más frecuente en investigaciones que usan registros administrativos o datos de redes sociales, aplicaciones, etc.
- ¿Cómo se generan esos datos perdidos?

Procesos de generación de valores perdidos

Ejemplos

MCAR	
X1	Y
0	NA
0	1
0	1
1	1
1	NA
2	NA
2	1
2	1
3	1
3	1
3	NA
3	1
4	1
4	NA
4	1
4	NA
4	1
4	1

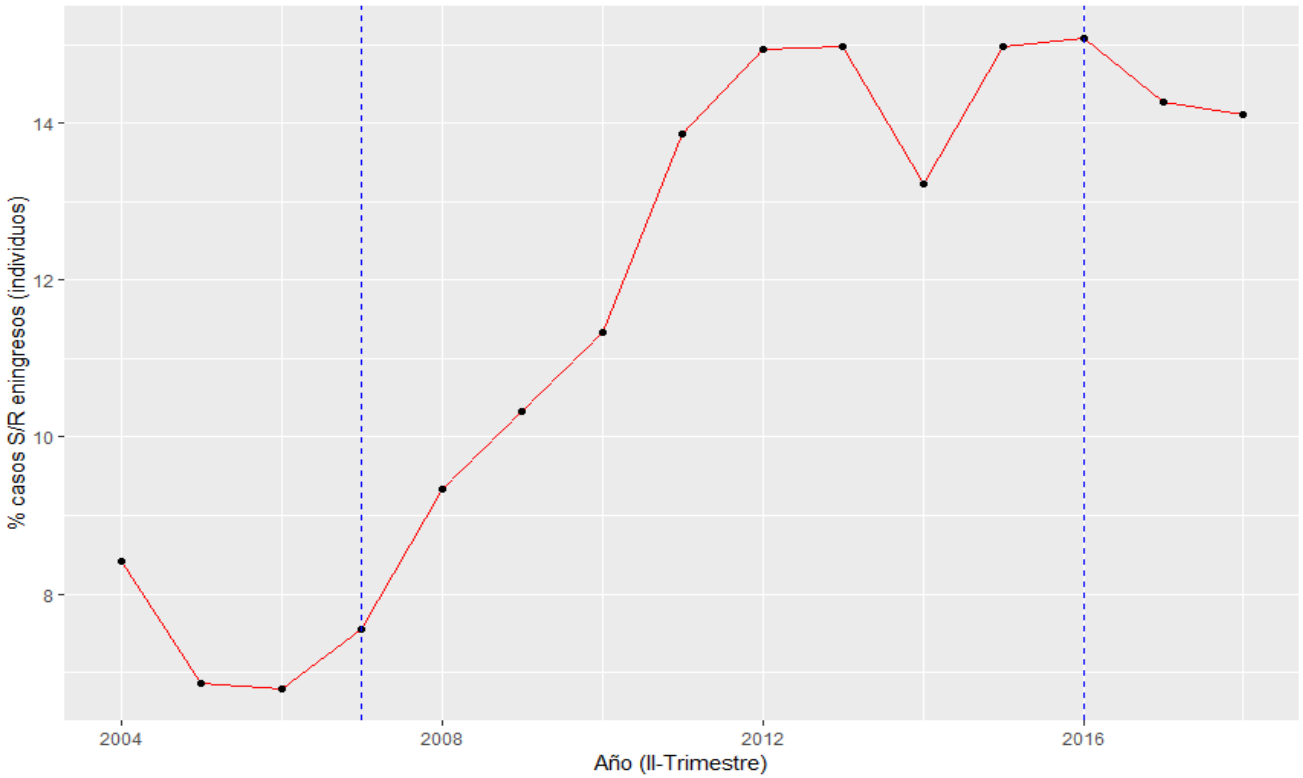
MAR	
X1	Y
0	1
0	1
0	1
1	1
1	1
2	1
2	1
2	1
3	1
3	NA
3	NA
3	1
4	1
4	NA
4	1
4	NA
4	1
4	NA

MAR	
X1	Y
0	1
0	1
0	1
1	1
1	1
2	1
2	1
2	1
3	1
3	1
3	1
3	1
4	1
4	NA
4	1
4	NA
4	NA
4	NA

¿Por qué es importante imputar datos?

Un ejemplo: EPH

Proporción de casos imputados (sin datos en alguna variable de ingresos) en EPH. Total de aglomerados urbanos, 2003-2018 (II-Trimestre de cada año)



¿Cómo lidiar con valores perdidos?

Imputación simple

- Exclusión de casos $\rightarrow$ se achica el dataset
- Reemplazo por la media o alguna otra medida $\rightarrow$ intervalos de confianza más estrechos de forma artificial
- Reponderación $\rightarrow$ es incómodo trabajar con varios sets de pesos.

¿Cómo lidiar con valores perdidos?

Hot Deck

- Método ampliamente usado. INDEC -hasta 2015- y Dirección de Estadística de la Ciudad para realizar imputaciones en EPH y EAH
- Reemplaza valores faltantes de un no respondente (¨receptor") con los valores observados de un respondente (¨donante") que es similar al receptor.

¿Cómo lidiar con valores perdidos?

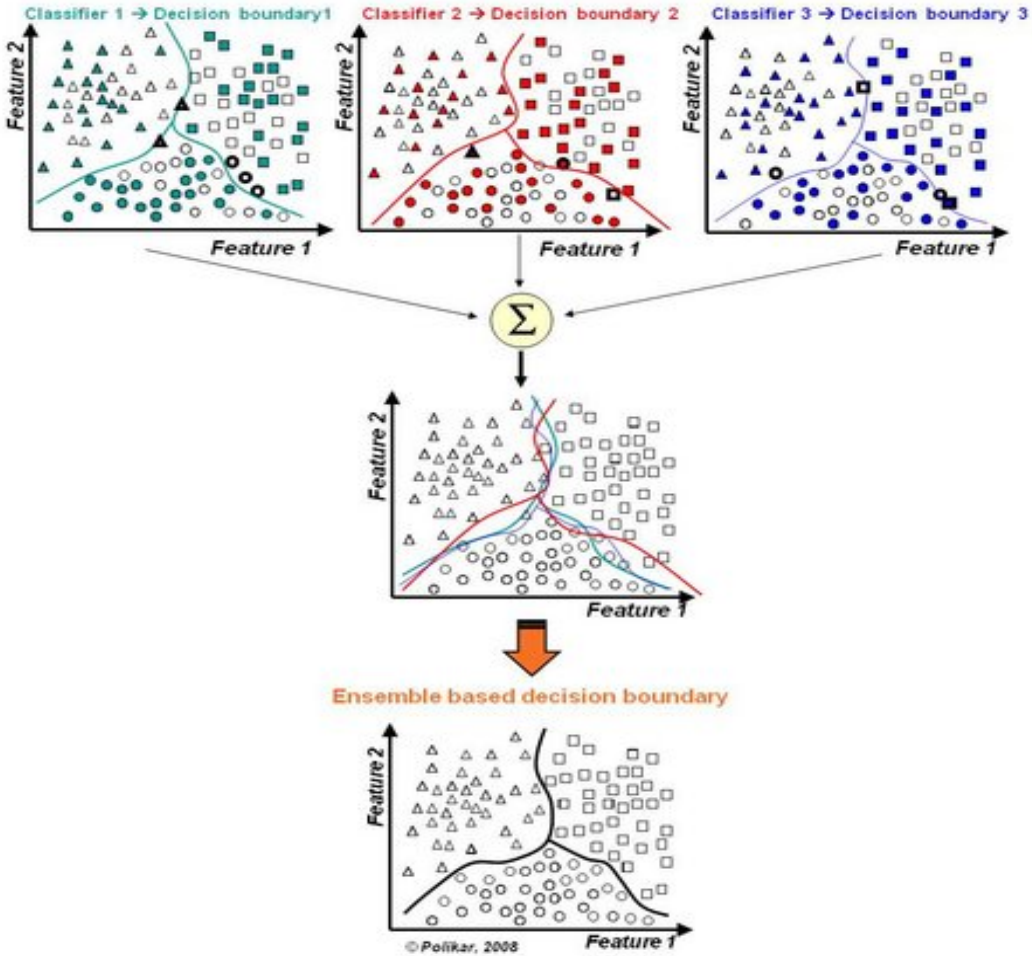
Hot Deck

- **Problema 1**: selección de la métrica de similitud entre los casos
- **Problema 2:** selección de los donantes. El donante es seleccionado aleatoriamente de un set de potenciales donantes hot-deck aleatorio- o bien se selecciona un solo caso donante, generalmente a partir de un algoritmo de v̈ecinos cercanosüsando alguna métrica -hot-deck determinístico-.

¿Cómo lidiar con valores perdidos?

Ensamble Learning

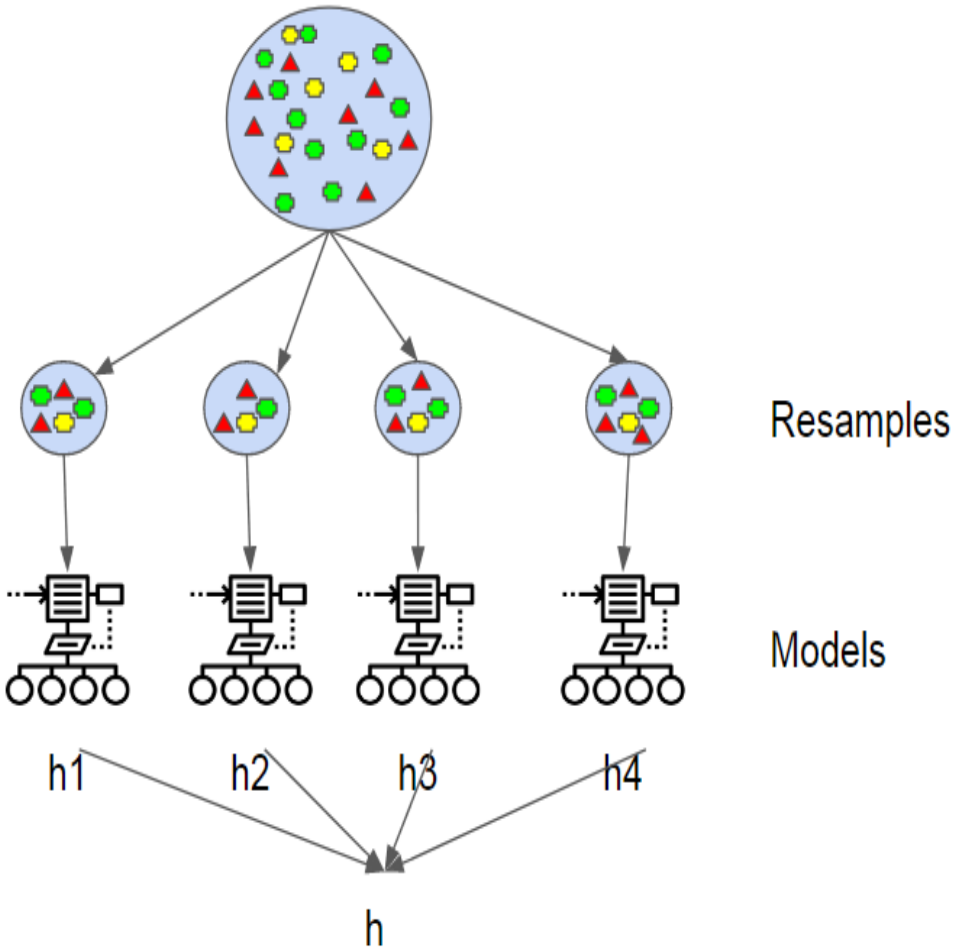
- Técnicas de aprendizaje supervisado donde se combinan varios modelos base.
- Ampliar el espacio de hipótesis posibles para mejorar la precisión predictiva del modelo combinado resultante.
- Los ensambles suelen ser mucho más precisos que los modelos base que los componen.



¿Cómo lidiar con valores perdidos?

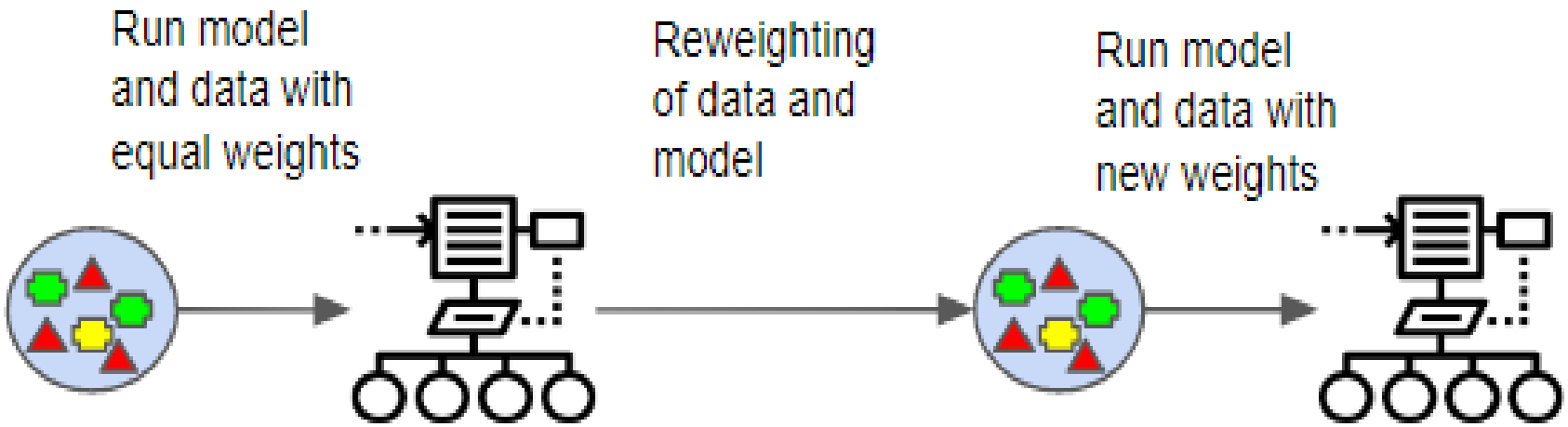
Ensamble Learning - Bagging

- Construcción de estimadores independientes -Boostrap-
- Combinación las predicciones mediante función agregación.
- Ejemplos: Random Forest, ExtraTrees, etc.



¿Cómo lidiar con valores perdidos?

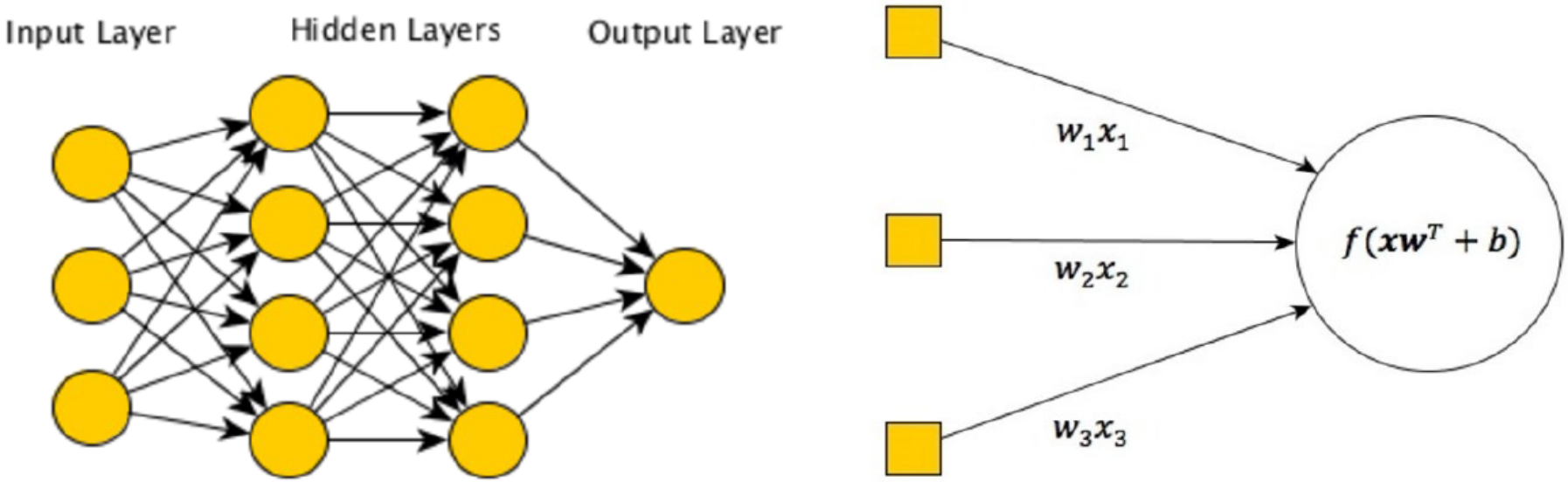
Ensamble Learning - Boosting



- Construcción secuencial de los estimadores
- Mayor peso en aquellos casos en los que se observa una peor performance.
- Ejemplos: AdaBoost y Gradient Tree Boosting, XGBoost.

¿Cómo lidiar con valores perdidos?

Ensamble Learning - Multi Layer Perceptron



Fuente: https://technology.condenast.com/story/a-neural-network-primer

- Cada neurona aplica una transformación lineal $x_i w_i^T + b$ seguida de una función de activación
- Al apilar capas de neuronas se aplican sucesivas de transformaciones lineales que permiten la construcción de modelos altamente no lineales

¿Cómo lidiar con valores perdidos?

Ensamble Learning - Bagging-LASSO

SaberEs

REVISTA DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

INICIO ACERCA DE INICIAR SESIÓN REGISTRARSE BUSCAR ACTUAL ARCHIVOS AVISOS

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

SciELO Scientific Electronic Library Online

latindex CATALOGO

Núcleo Básico de Revistas Científicas Argentinas NB

REDIB Red Iberoamericana de Innovación y Conocimiento Científico

Dialnet

EconBib

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

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.

Palabras clave

Regularización; LASSO; No respuesta

Experimentos con EPH

Bagging-LASSO

- Se aplica el algoritmo bagging a la imputación de ingresos laborales en la EPH del II trimestre de 2015
- En cada remuestra se estima la siguiente regresión LASSO

$$log_{10}(y_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + e_i \tag{1}$$

- Buscando minimizar la siguiente función de costo:

$$CF = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

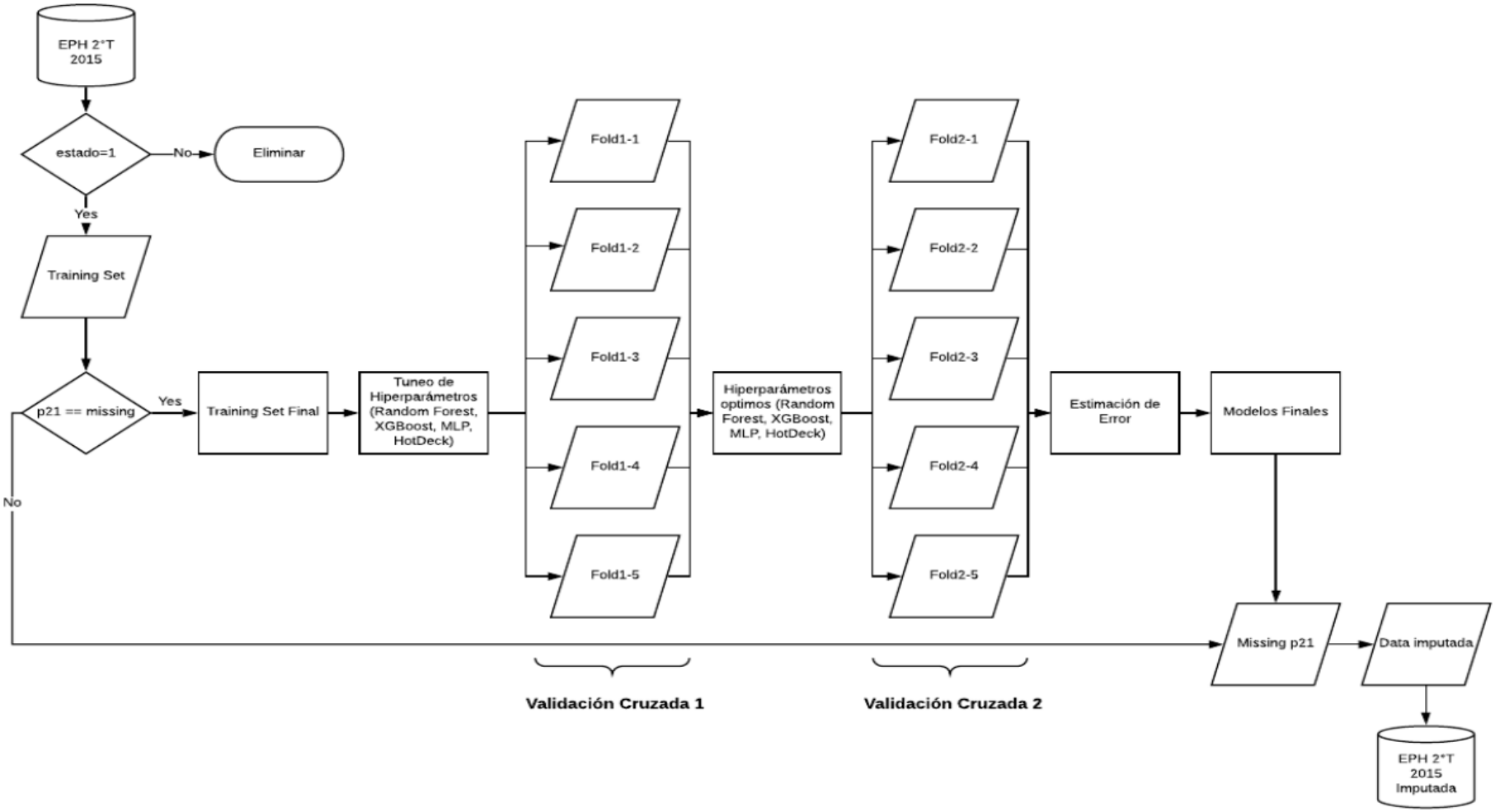
Experimento con EPH

Pipeline

- Dataset: EPH 2do. trimestre de 2015
- Población: Ocupados en la semana de referencia
- Variables predictoras sociodemográficas, laborales y otros ingresos
- Repo: https://github.com/gefero/ML_imputation

Experimento con EPH

Pipeline



Experimento con EPH

Estrategia de validación 1

- Estimación de métricas de error
- Supuesto: Proceso de generación de datos perdidos MCAR o MAR

Tabla 3. Métricas de performance predictiva de los diferentes algoritmos entrenadas

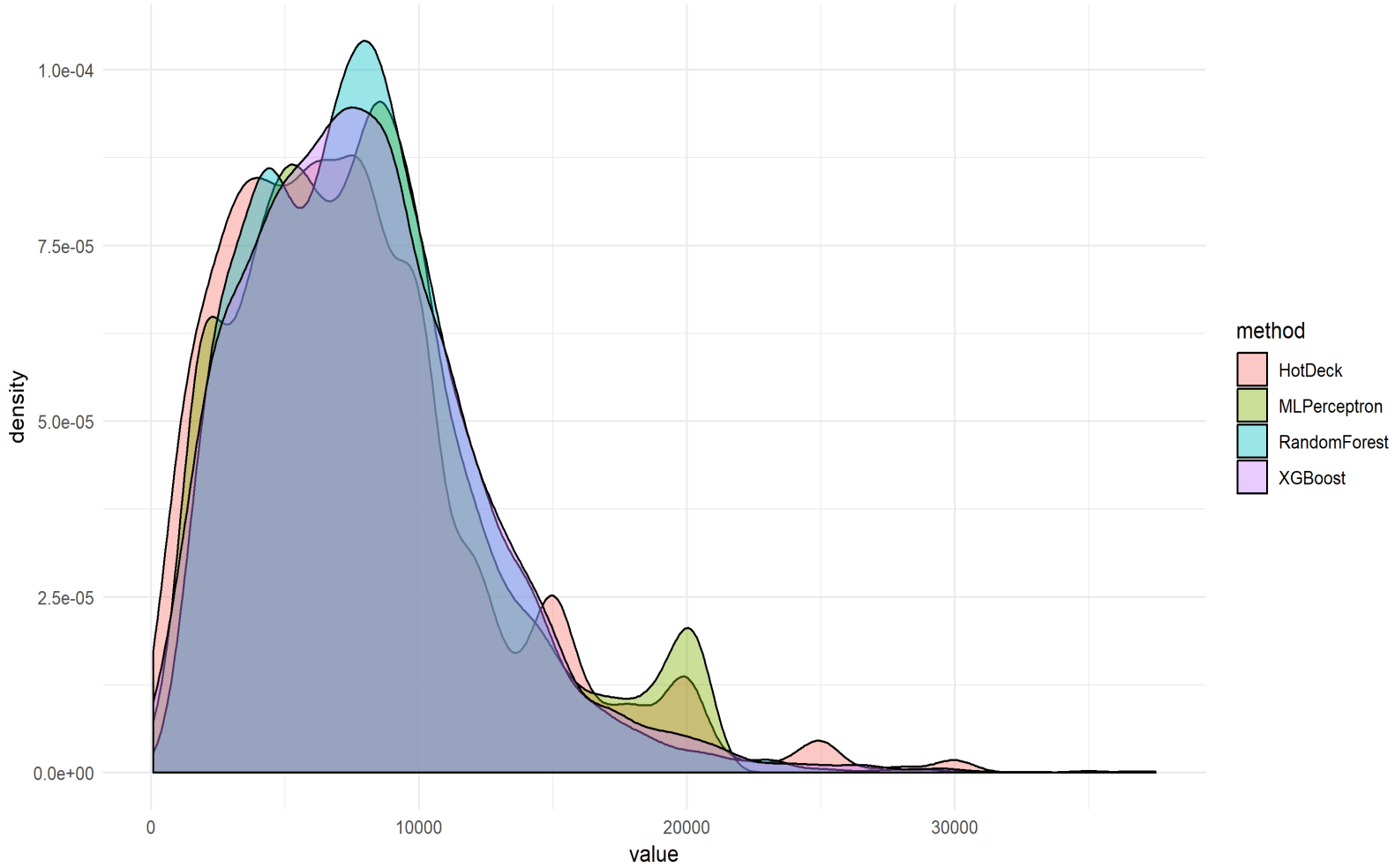
Algoritmo	RMSE	MAE
Hot Deck	$5930.6	$3740.6
Random Forest	$2800.6	$1561.9
XGBoost	$3260.8	$2016.8
MLP	$3974.2	$2293.1

Fuente: elaboración propia en base a microdatos de la EPH - 2do. trimestre de 2015

Experimento con EPH

Estrategia de validación 2

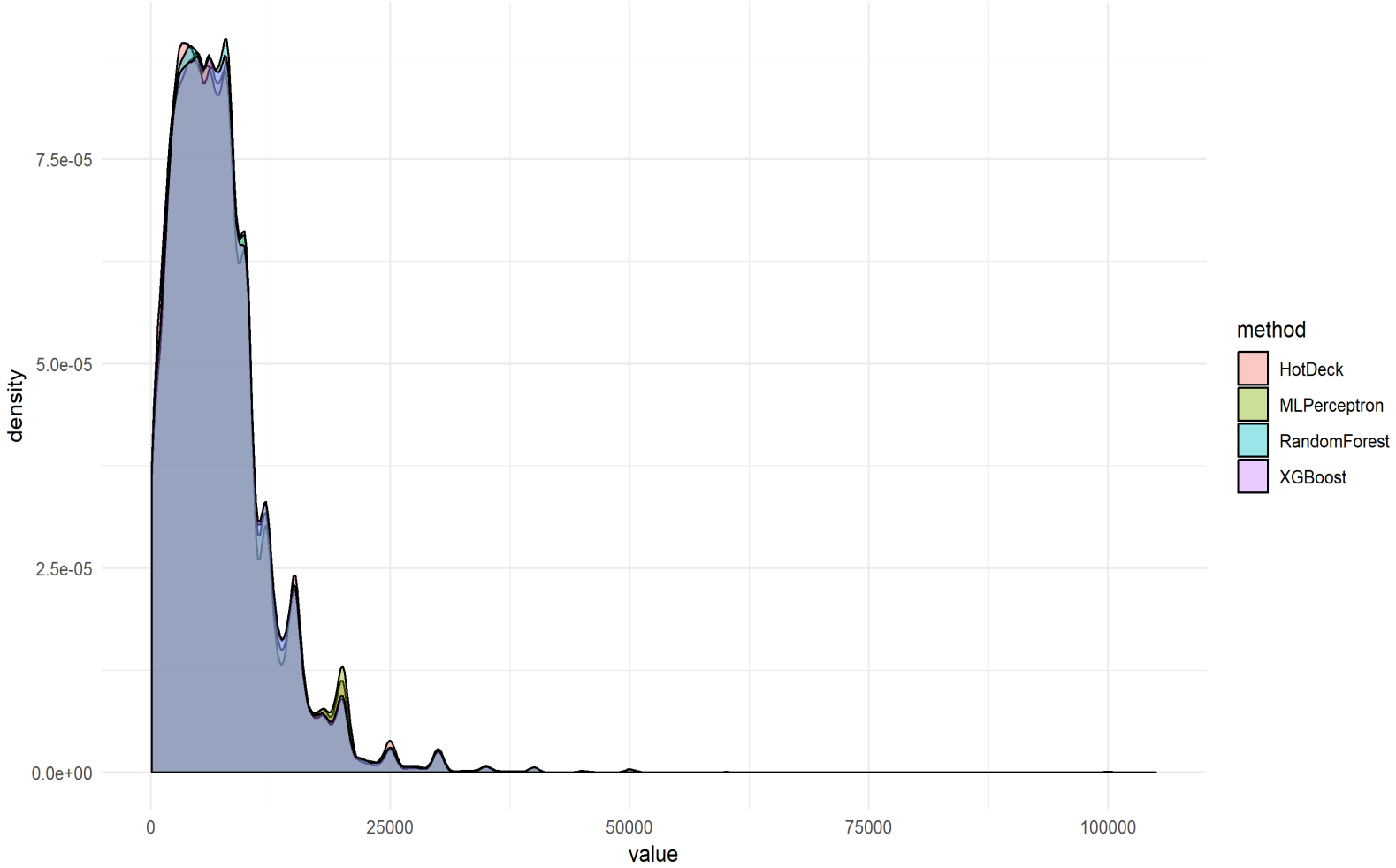
Comparación de distribuciones sobre datos perdidos reales (es decir, imputados por INDEC)



Experimento con EPH

Estrategia de validación 2

Comparación de distribución de datos completos (imputados + respuesta)



Resumen

- Machine Learning como alternativa para la imputación
- Reducción considerable en el *RMSE* entre casos perdidos comparado a Hot Deck -entre 30 % y 50 %-
- Problemas a futuro
  - Extensión del alcance del ejercicio
  - Mejoras en tuneo de hiperparámetros (algoritmos de búsqueda más inteligentes, diferentes funciones de activación, etc.)
  - Propiedades de los estimadores y estimaciones de medidas basadas en ingresos al utilizar estas técnicas
  - Performance relativa a HotDeck en procesos de generación de datos no aleatorios

La Yapa...

rindec/eph

README.md

eph

build passing | codecov 97% | CRAN 0.1.0

Caja de Herramientas para el procesamiento de la Encuesta Permanente de Hogares

Descripción

La librería `eph` tiene por objecto facilitar el trabajo de aquellos usuarios y usuarias de la Encuesta Permanente de Hogares - INDEC que deseen procesar datos de la misma mediante R.

Sus principales funciones son:

- `get_microdata()` : Descarga las bases de microdatos directamente de la página de INDEC
- `organize_labels()` : Etiqueta las bases siguiendo el último diseño de registro
- `calculate_tabulates()` : Crea tabulados uni o bivariados con ponderacion, totales parciales y porcentajes.
- `calculate_poverty()` : Replica el cálculo de pobreza e indigencia del INDEC, pero para las bases trimestrales^[el calculo oficial se realiza sobre bases semestrales no publicadas]
- `get_poverty_lines()` : Descarga de canasta basica alimentaria y canasta basica total
- `orgnize_panels()` : Arma un pool de datos para trabajar con panel en la EPH continua

https://github.com/rindec/eph

La Yapa...

rindec/eph

```
####Levanto Bases####
Individual_t117 <- read.table("Fuentes/usu_individual_t117.txt",
                              sep=";",
                              dec=",",
                              header = TRUE, fill = TRUE)

Individual_t216 <- read.table("Fuentes/usu_individual_t216.txt",
                              sep=";",
                              dec=",",
                              header = TRUE, fill = TRUE) %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)

Individual_t316 <- read.table("Fuentes/usu_individual_t316.txt",
                              sep=";",
                              dec=",",
                              header = TRUE, fill = TRUE) %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)

Individual_t416 <- read.table("Fuentes/usu_individual_t416.txt",
                              sep=";",
                              dec=",",
                              header = TRUE, fill = TRUE) %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)
```



```
library(eph)

####Levanto Bases####
Individual_t117 <- get_microdata(year = 2017,
                                 trimester = 1,
                                 type = "individual")

Individual_t216 <- get_microdata(year = 2016,
                                 trimester = 2,
                                 type = "individual") %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)

Individual_t316 <- get_microdata(year = 2016,
                                 trimester = 3,
                                 type = "individual") %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)

Individual_t416 <- get_microdata(year = 2016,
                                 trimester = 4,
                                 type = "individual") %>%
  select(ANO4,TRIMESTRE,P21,PONDIIO,IPCF, PONDIH)
```

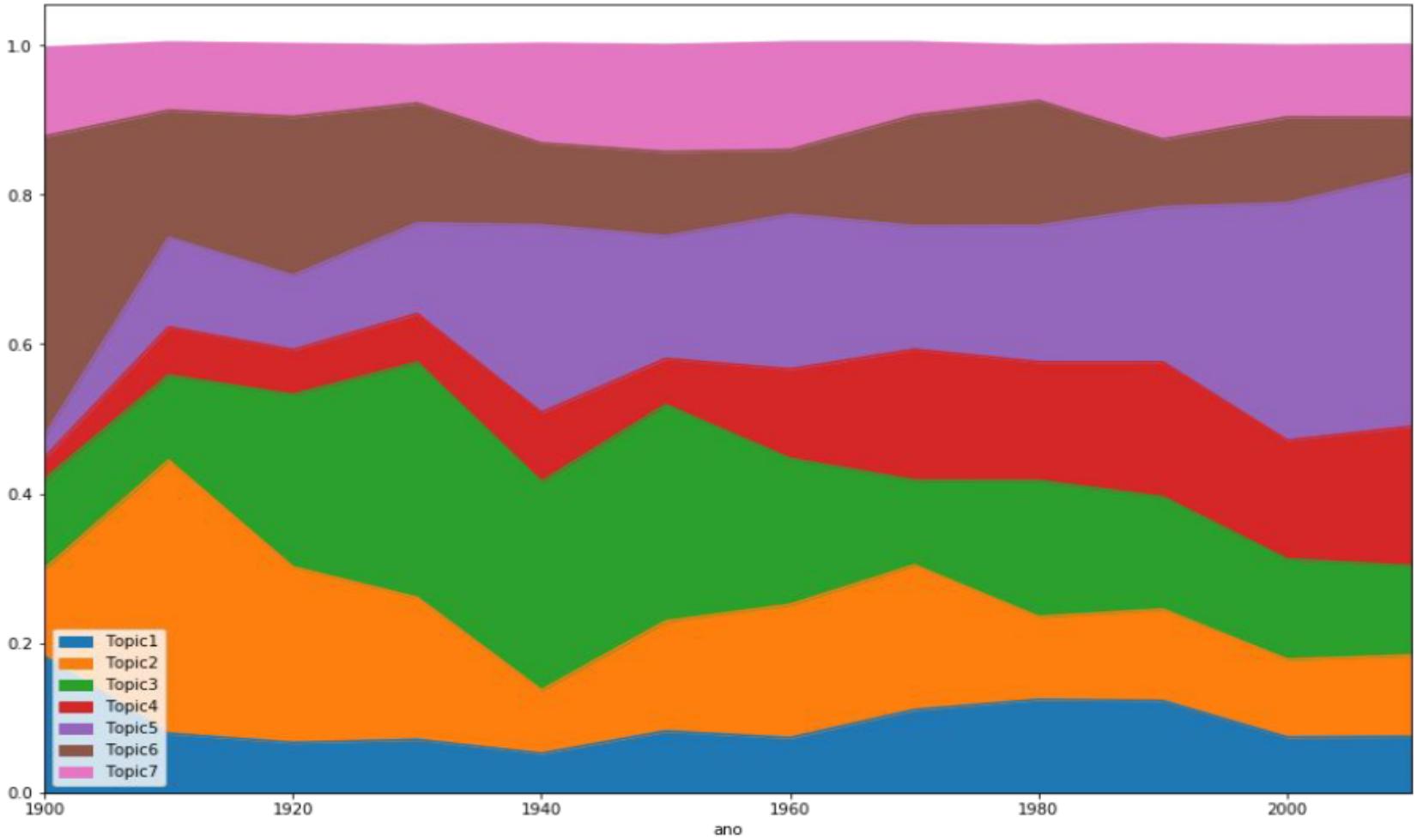
Composición de tópicos de algunos tangos

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
	Amor signo -	Imág. naturales	Amor signo +	Miscelaneo	Ciudad	Tango	Personif.
Arrabal amargo	0.02	0.02	0.02	0.02	**0.85**	0.02	0.02
Barrio reo	0.03	0.03	0.03	**0.53**	0.03	**0.34**	0.03
Cafetin de Buenos Aires	0.02	0.02	**0.49**	**0.38**	0.02	0.02	0.02
Garua	0.03	0.03	0.03	0.03	**0.85**	0.03	0.03
Lejana Tierra mía	0.03	0.03	0.03	0.03	**0.84**	0.03	0.03

La Yapa...

NLP - letras de tango

Evolución temporal de los tópicos



¿Preguntas?

@Crst_C

german.rosati@gmail.com

https://gefero.github.io/