

Detección de Anomalías en Problemas Espacio-Temporales

Una Aplicación al Mercado Inmobiliario de
Departamentos en Buenos Aires

Dr. Andrés Farall
Profesor en Ciencia de Datos – FCEyN - UBA
afarall@hotmail.com

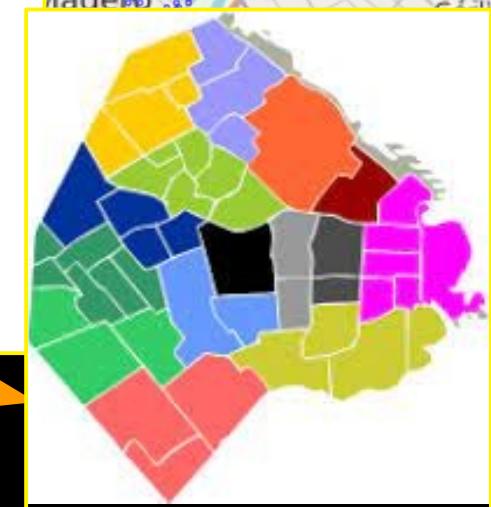
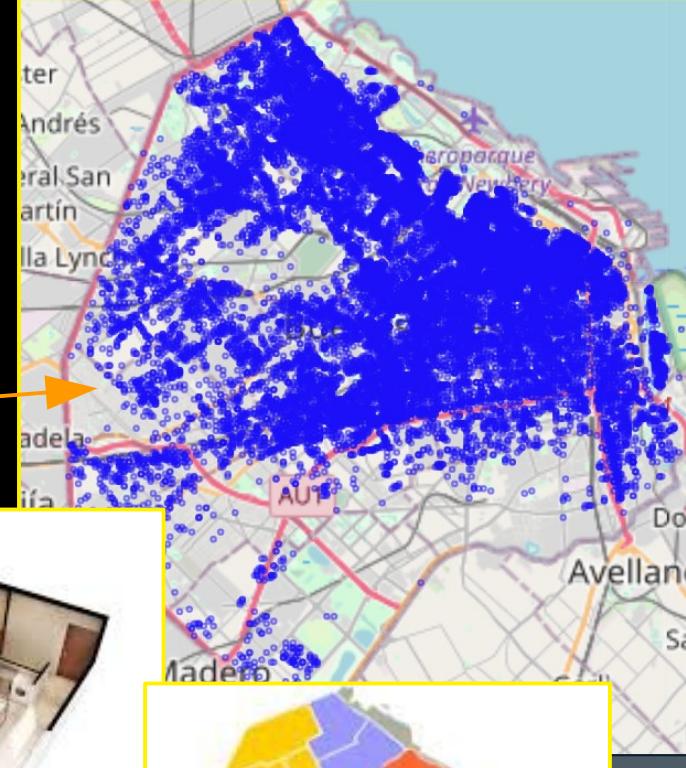
Temario

- Explorando Los Datos
- Definición de la Métodología
- Modelado del Precio
- Cálculo de Residuos Relativos
- Distribución Espacial de Residuos
- Hay Estructura en los Residuos ?
- Detección de Bumps
- Deteción de Anomalías en los Residuos
- Ubicación Espacial de las Anomalías
- Anomalías Detectadas

Los Datos

28.470 Departamentos con:

- Precio
- Georeferenciación
- Superficie Cubierta
- Cantidad de Baños
- Cantidad de Cuartos
- Cantidad de Ambientes
- Barrio del Anuncio
- Distancia al Metro
- Tiempo (Jul 2020-Jun 2021)

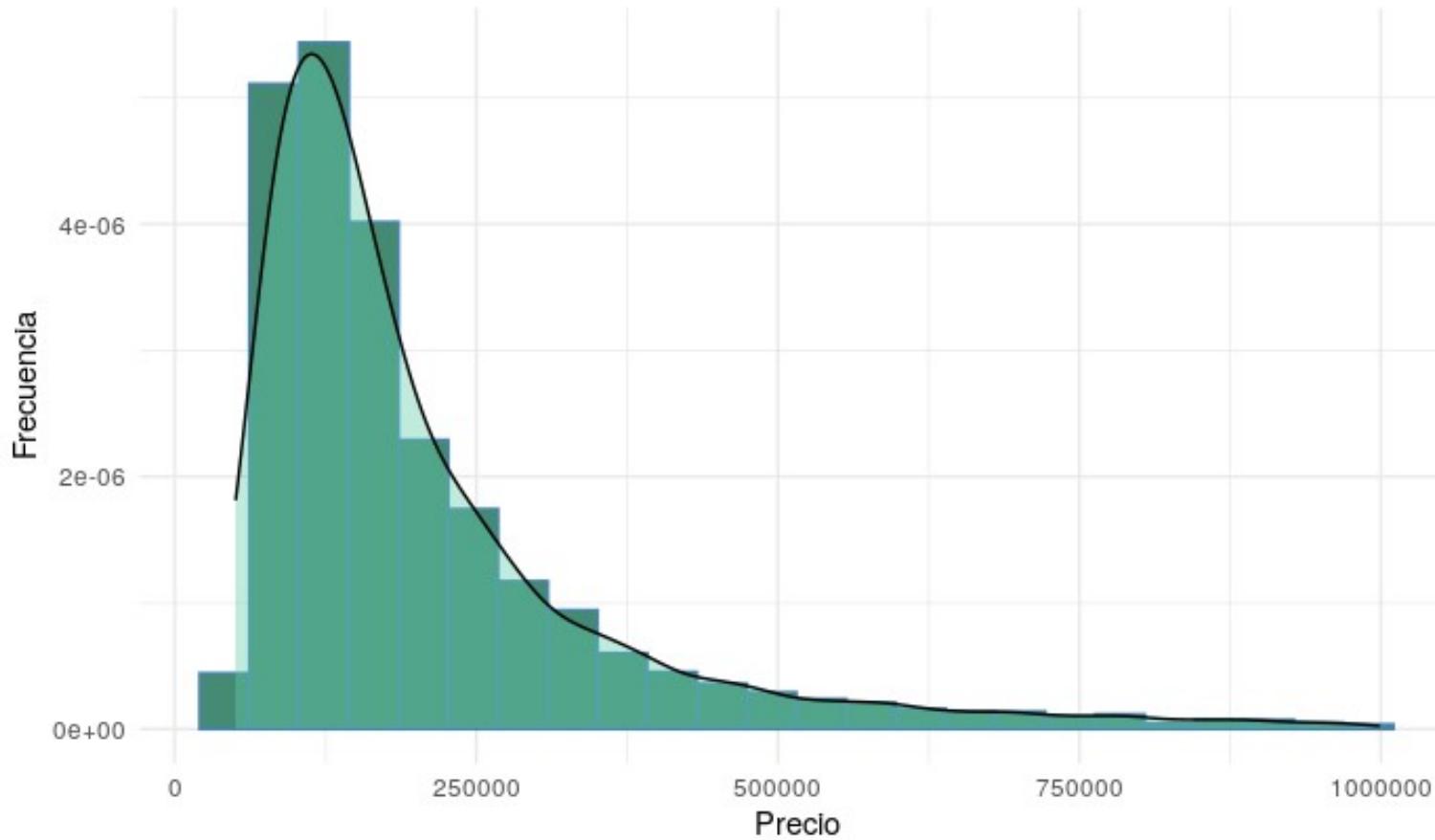


La Metodología

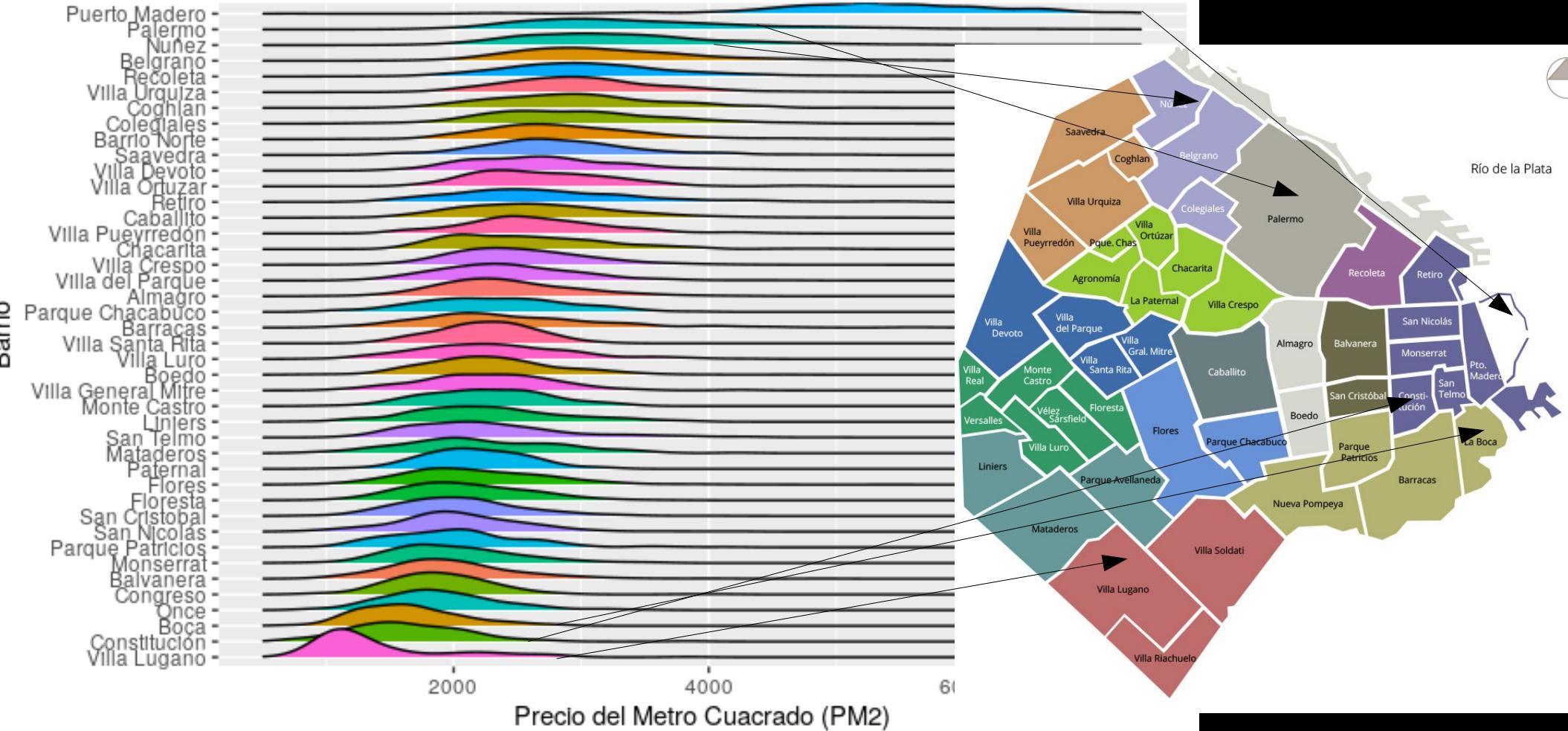
- Ajustamos un modelo supervisado para predecir el **precio** de los departamentos
 - Usando componentes **espacio-temporales** “suaves”
 - Usando las características de los departamentos
- Calculamos los **Resodos Relativos** (RR) del ajuste
- Verifiamos **Independencia** entre Puntos (Lon,Lat) y Marcas (RR)
- Detectamos Bumps con **PRIM**
- Aplicamos **OD** a los (RR+LON+LAT)
 - Usando Local Outlier Factor (**LOF**)
 - Usando Isolation Forest (**liForest**)
- Analizamos las **Anomalías**

El Precio

Distribución de Precios de los Departamentos

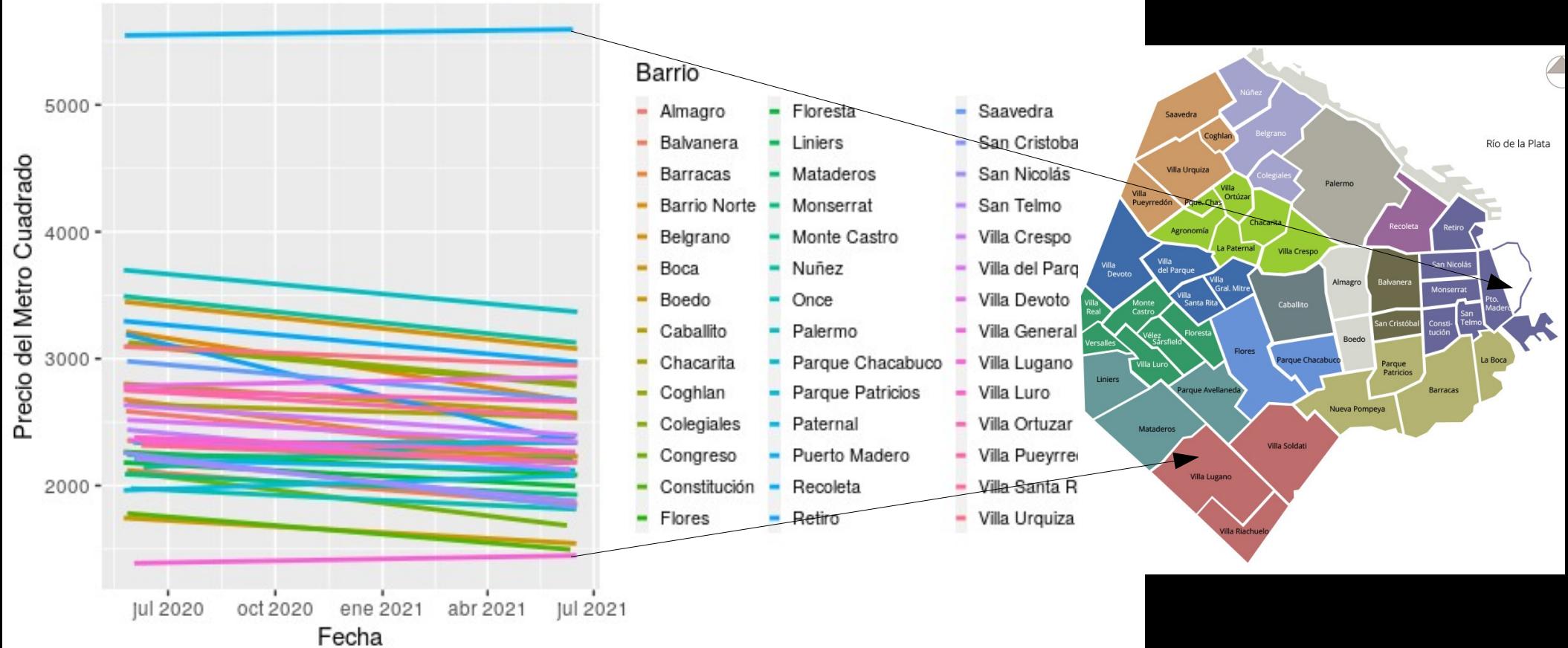


Precio del Metro Cuadrado X Barrio



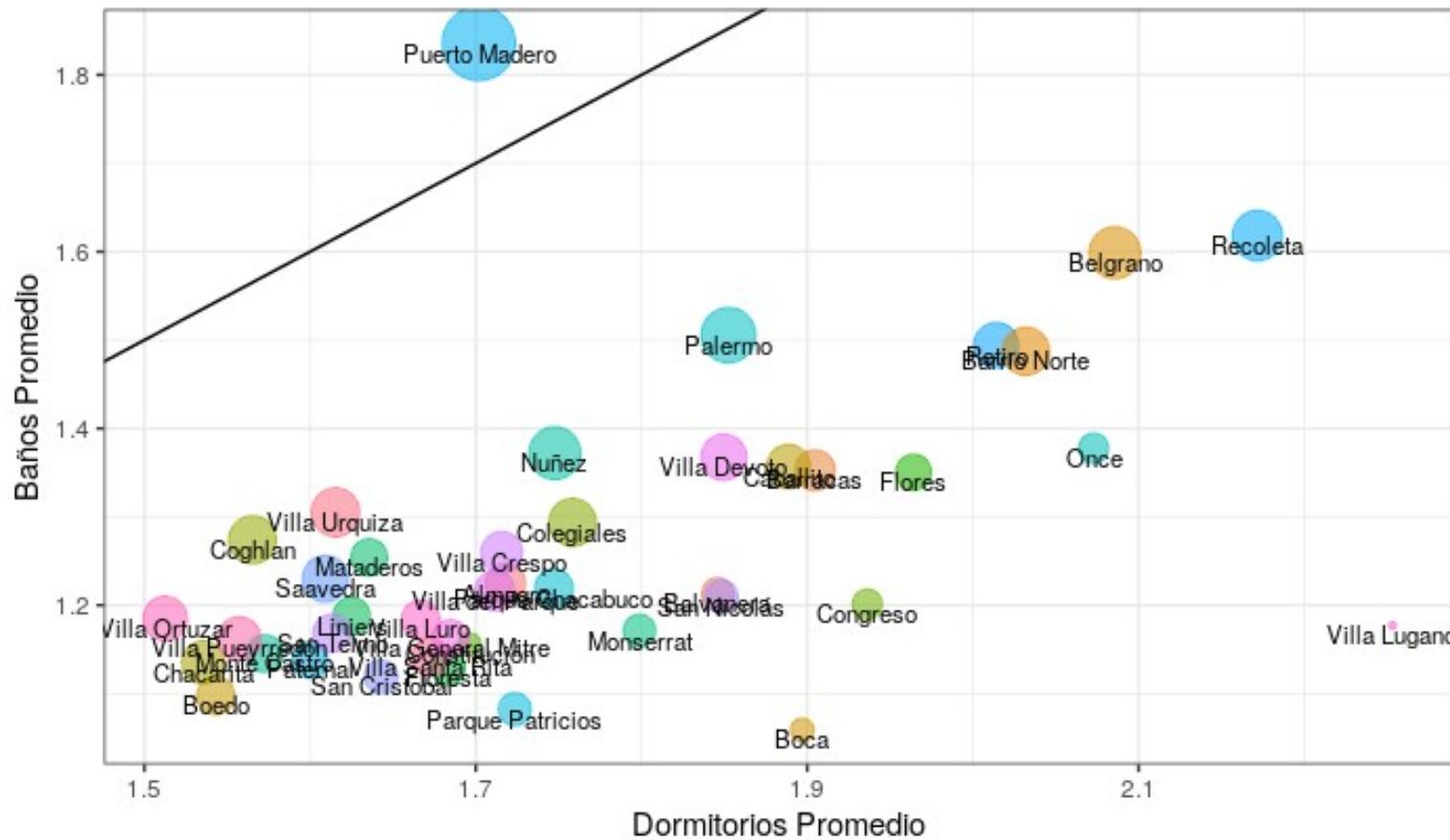
Evolución Temporal del Pm2

Evolución Temporal del Pm2 por Barrio



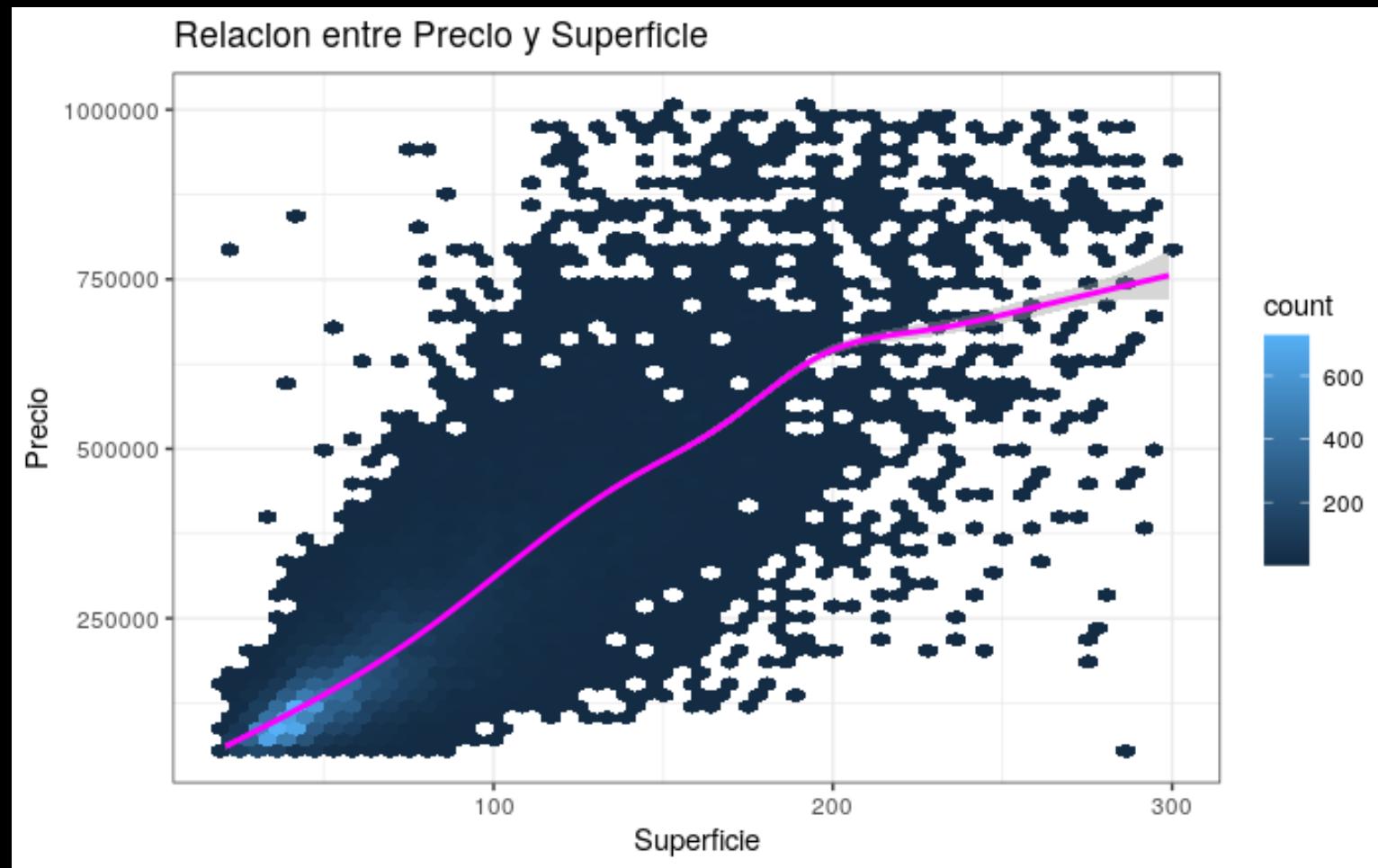
Relación Baños-Dormitorios

Relacion entre Baños y Dormitorios por Barrio

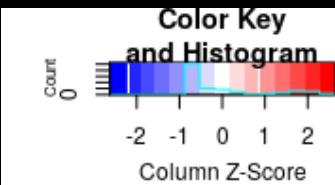
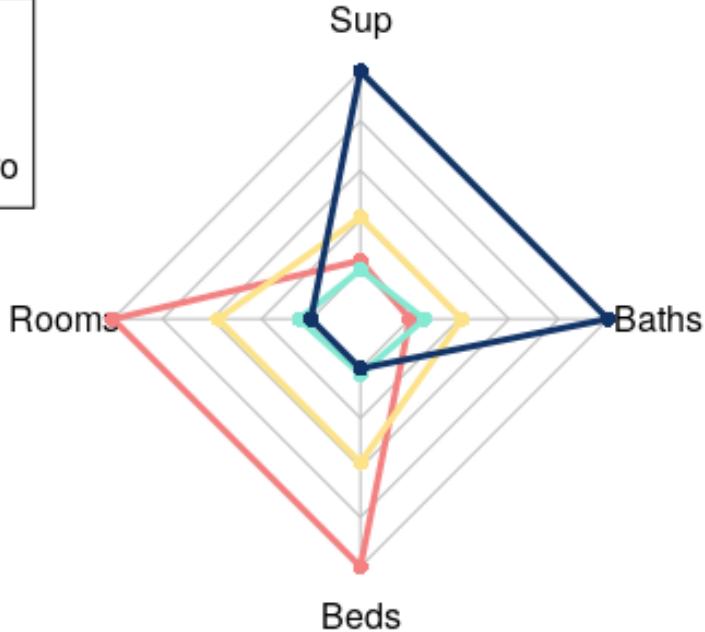


Tamaño del
Punto
proporcional al
PM2

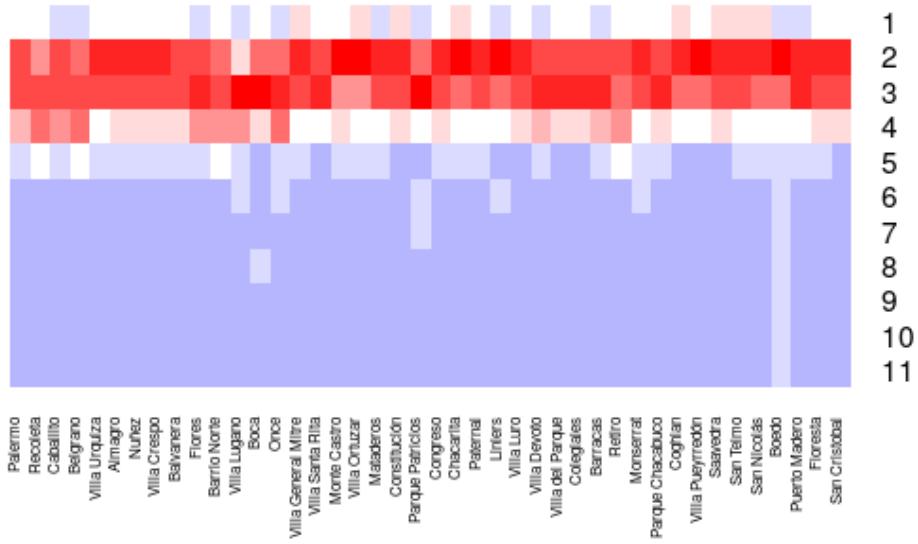
Precio X Superficie



Ambientes



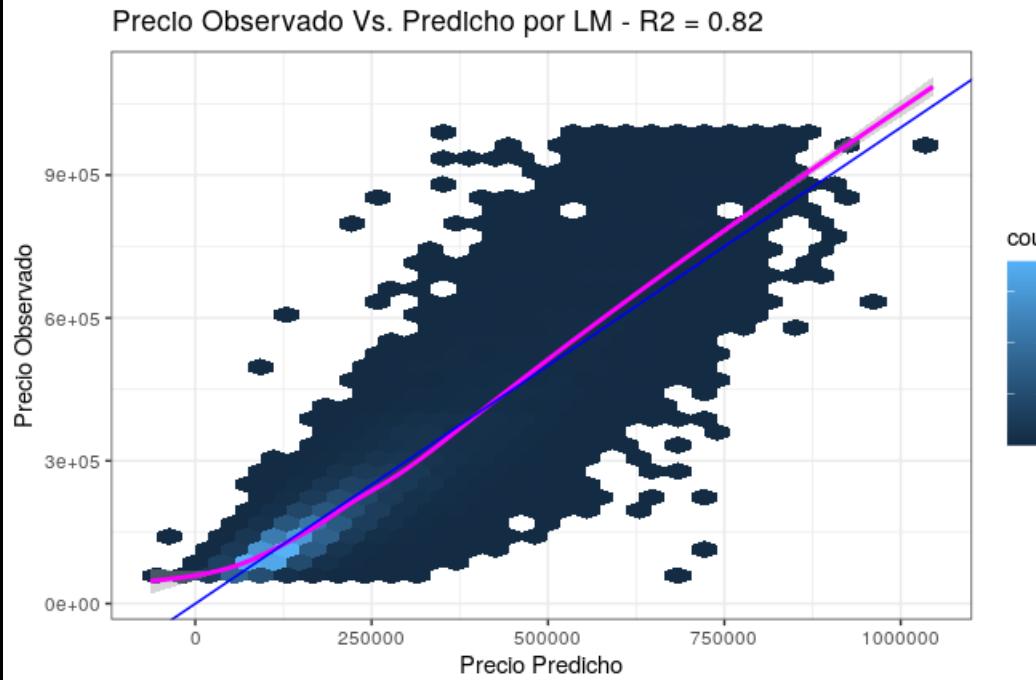
Cantidad de Ambientes por Barrio



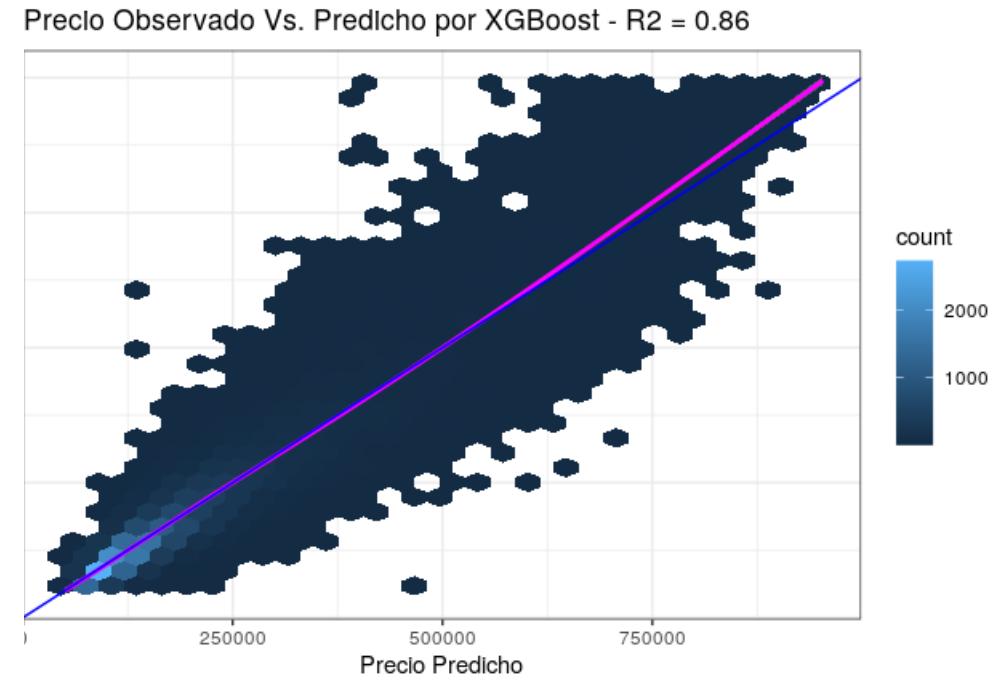
Modelado del Precio

Precio ~ Sup + Baths + Beds + S(Lat,Lon,Tiempo) + Subte + Barrio

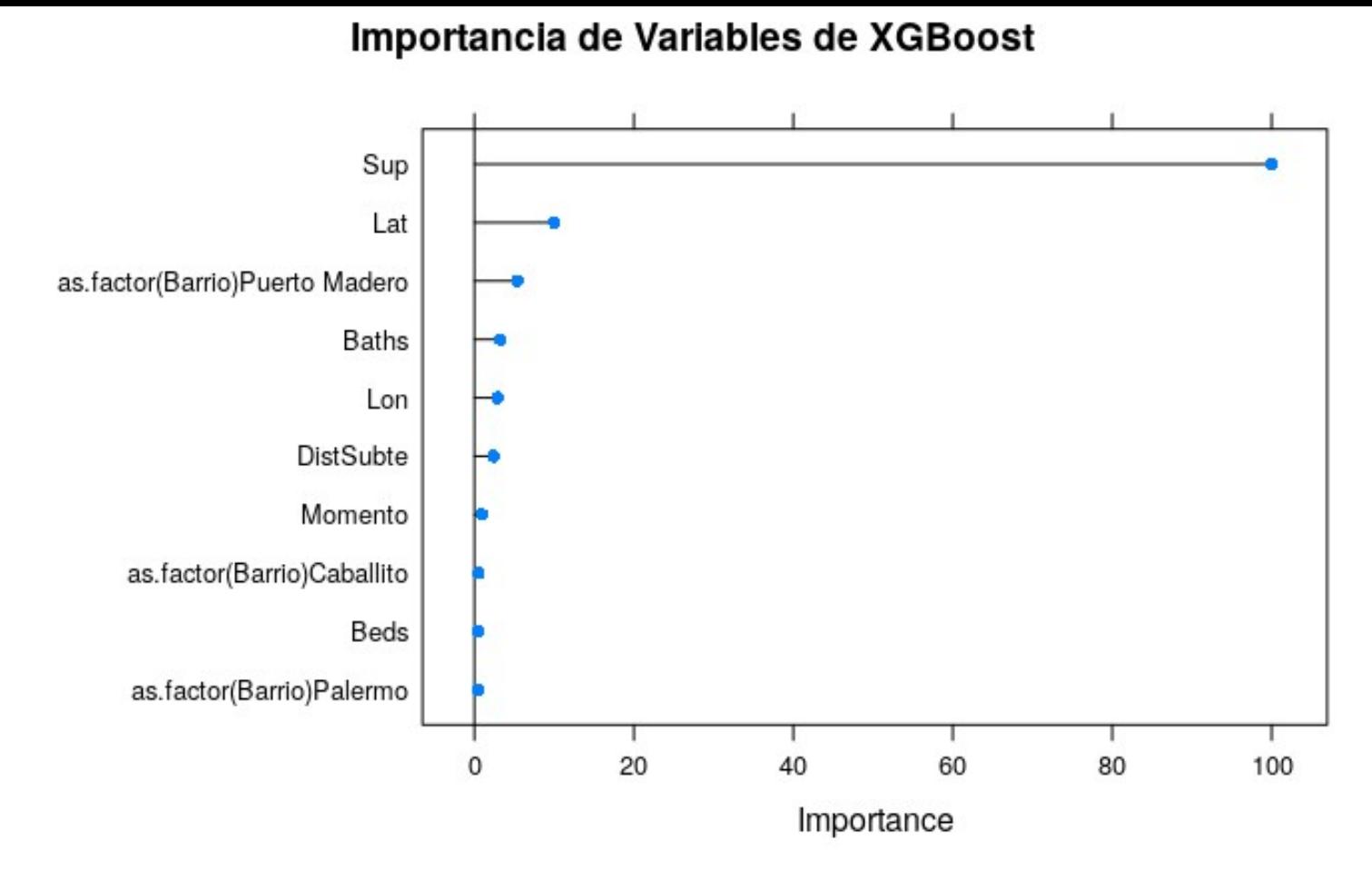
Modelo Lineal Múltiple



XGBoost



Importancia de Variables



Los Residuos

Residuos Relativos

$$RR = \frac{Obs - Pred}{Obs}$$

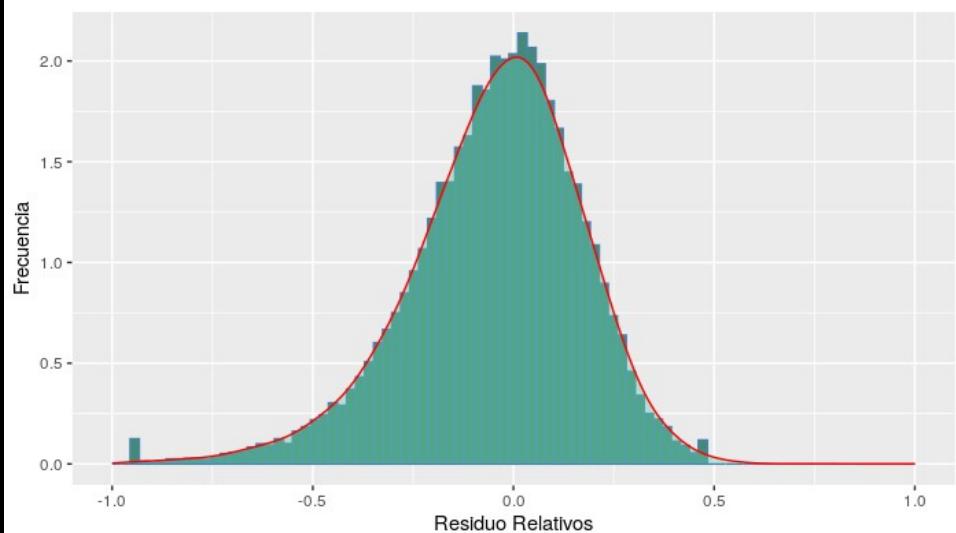
Residuos Relativos Topeados

$$RRT = \begin{cases} P_{0.25} & \text{si } RR < P_{0.25} \\ P_{99.75} & \text{si } RR > P_{99.75} \\ RR & \text{cc} \end{cases}$$

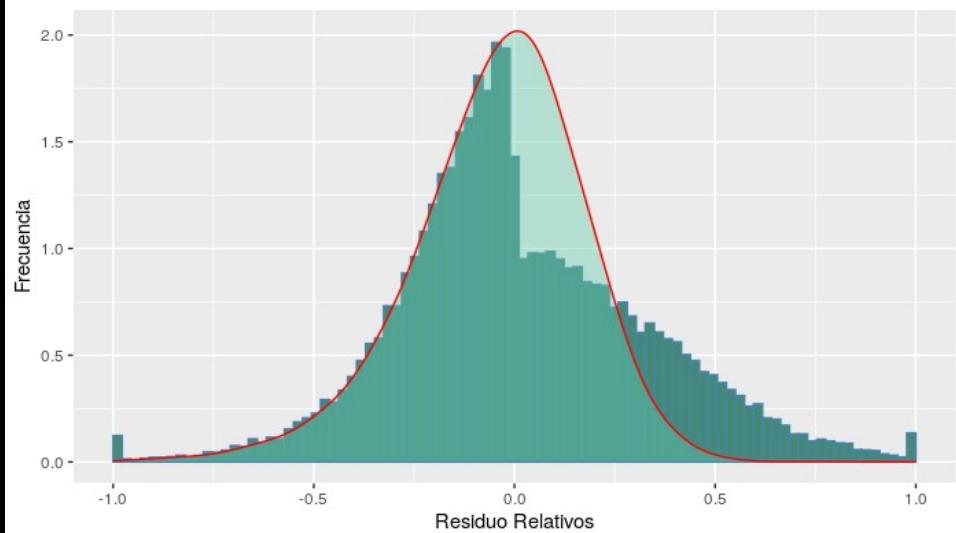
Residuos Relativos Topeados y Balanceados

$$RRTB = \begin{cases} \frac{RRT}{\text{Min}(RRT)} & RRT < 0 \\ \frac{RRT}{\text{Max}(RRT)} & RRT > 0 \end{cases}$$

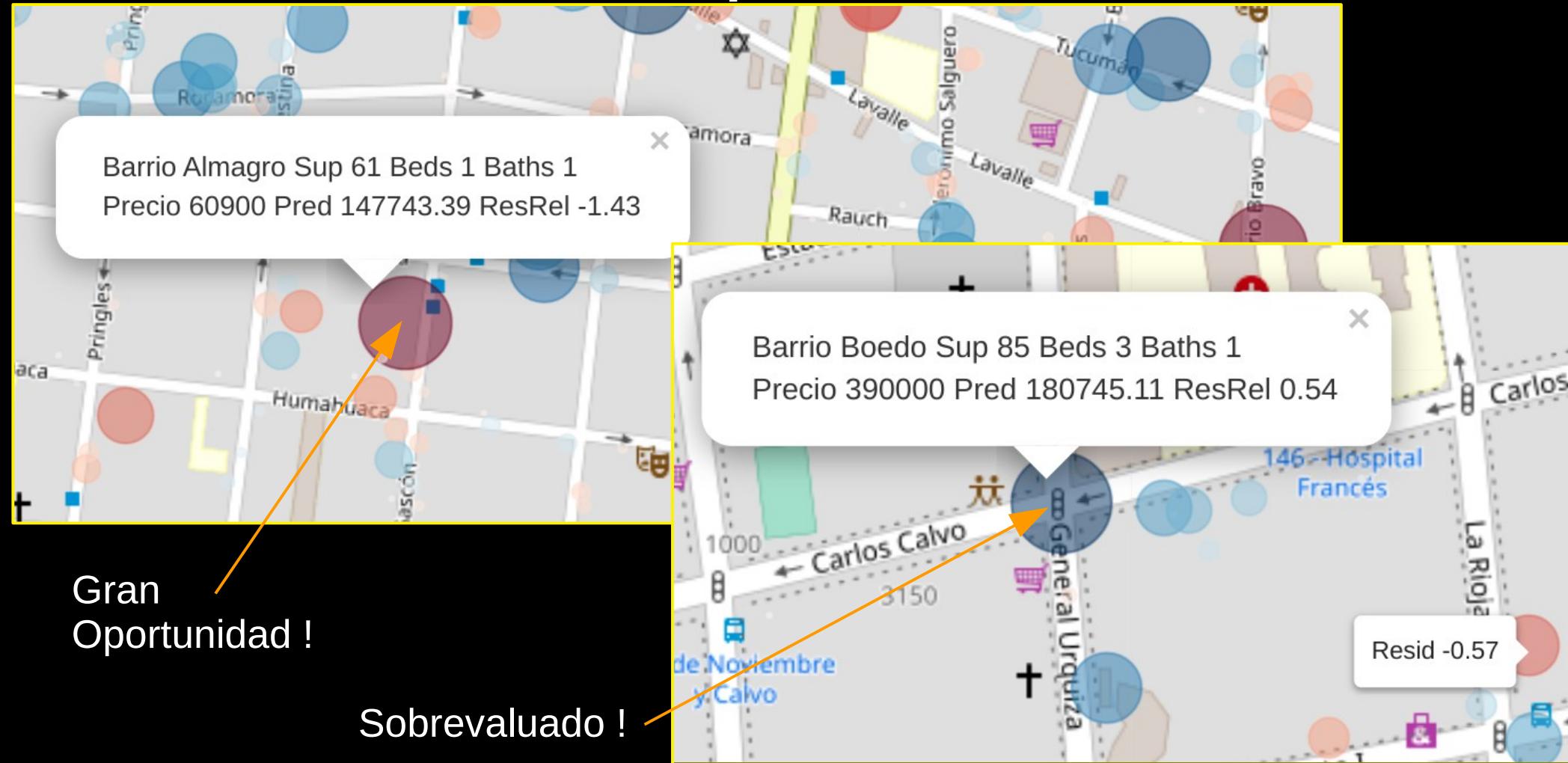
Distribución de los Residuos Relativos



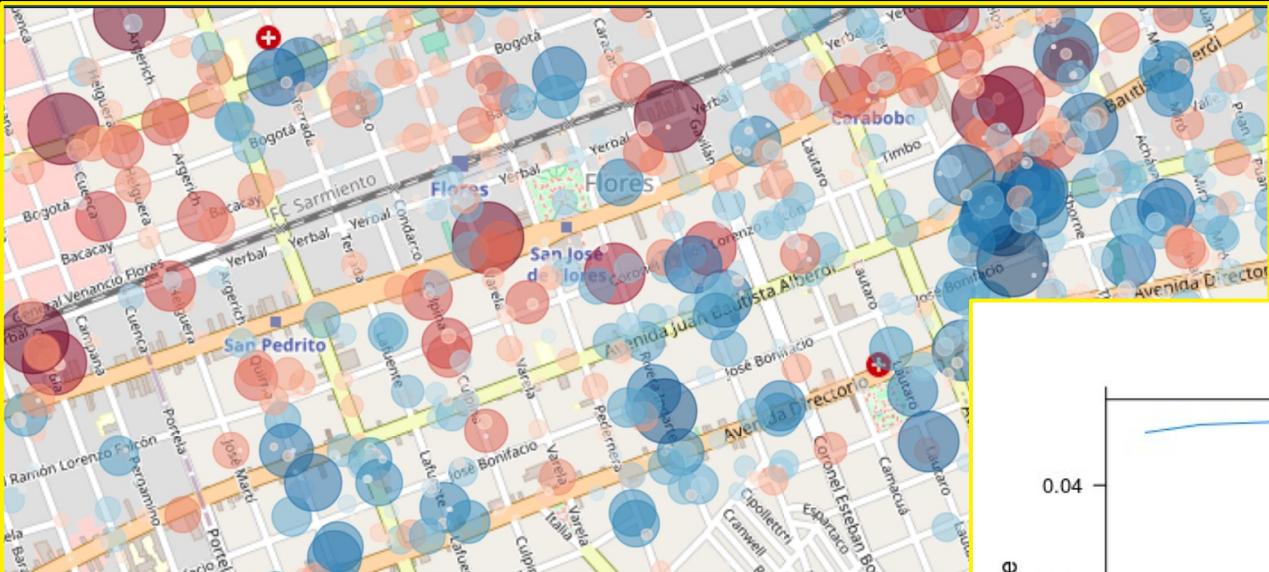
Distribución de los Residuos Balanceados



Distribución Espacial de Residuos

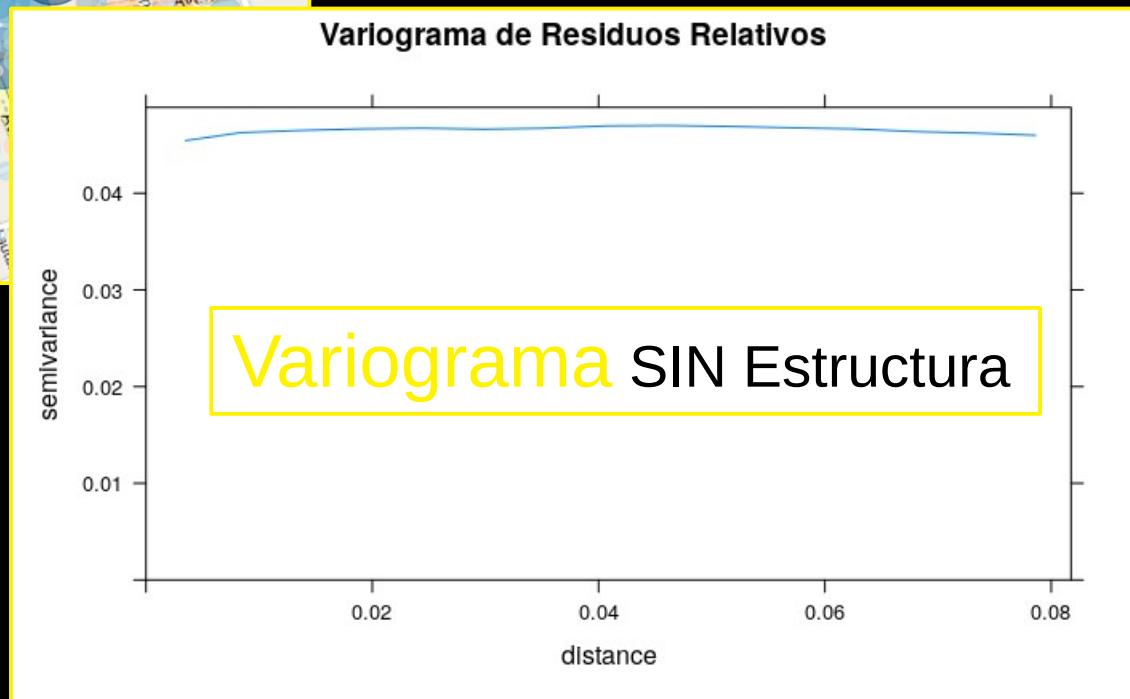


Distribución Espacial de Residuos

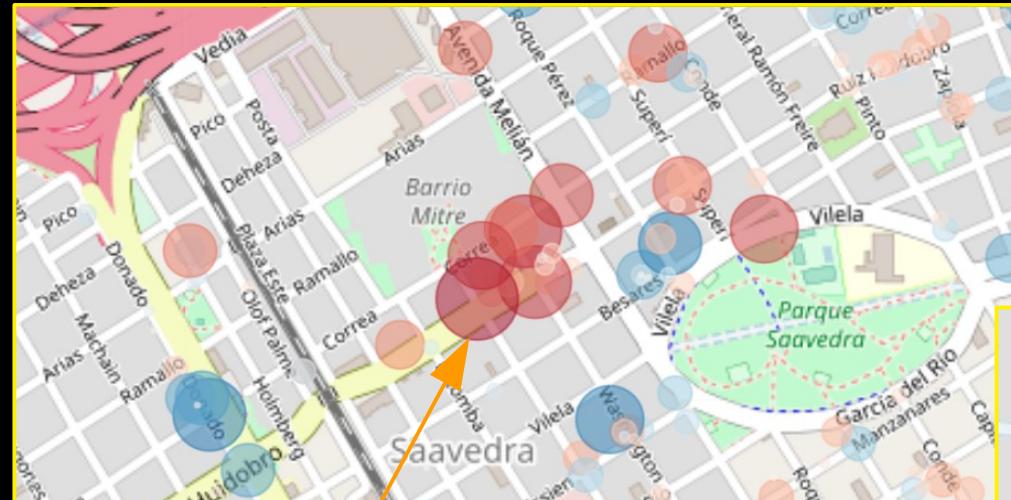


No existe Autocorrelación
Espacial: **Moran I** = -0.0016
Test: pvalor=0.8

Sin Estructura !!!



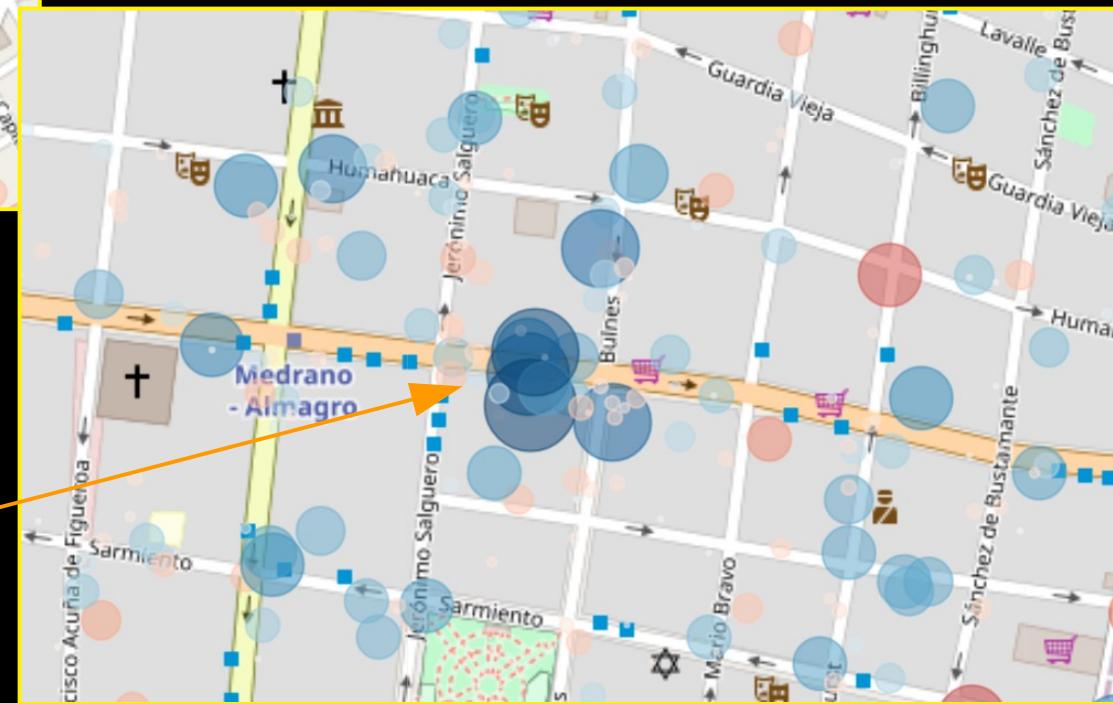
Distribución Espacial de Residuos



Cluster de
SUB-valuaciones

Cluster de
SOBRE-valuaciones

Pero HAY
CLUSTERS !!!



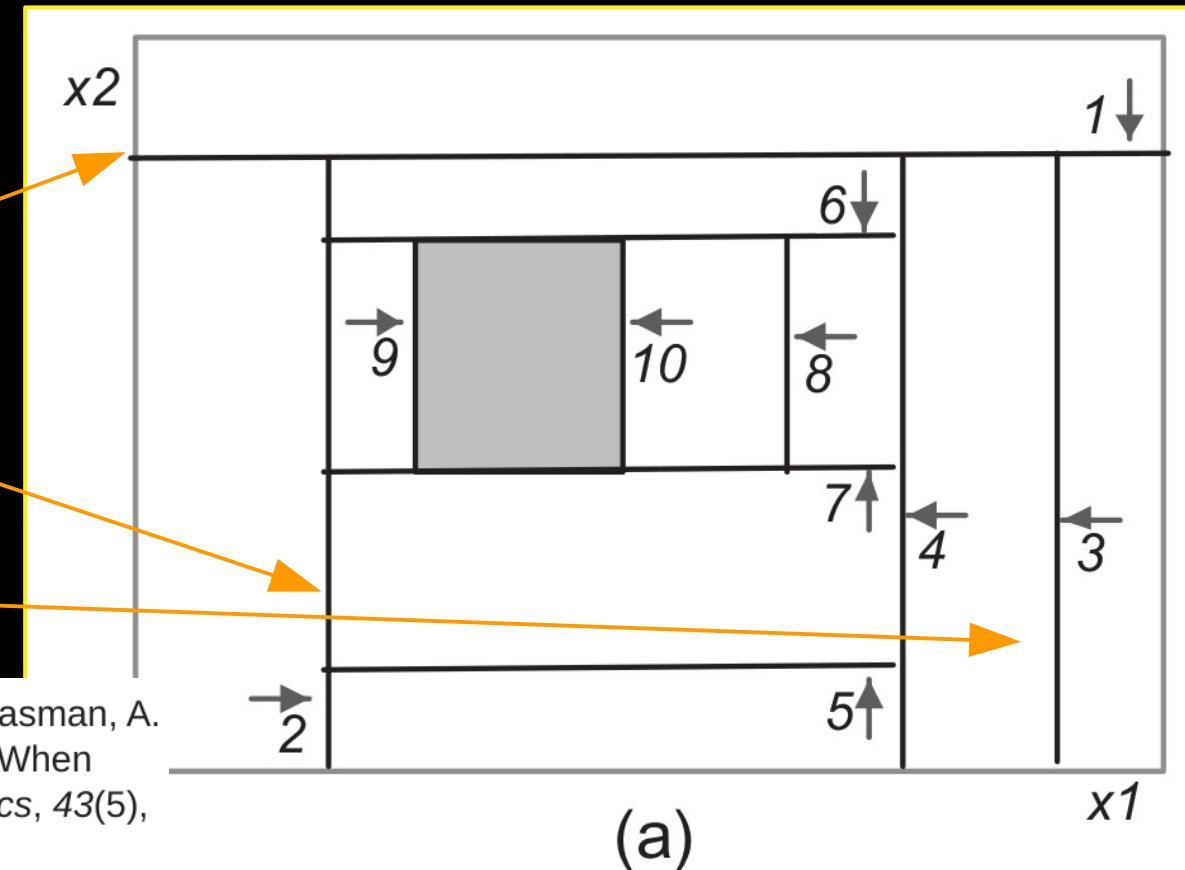
Patient Rule Induction Method (PRIM)

Friedman, J. H., & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and computing*, 9(2), 123-143.

Primer Recorte

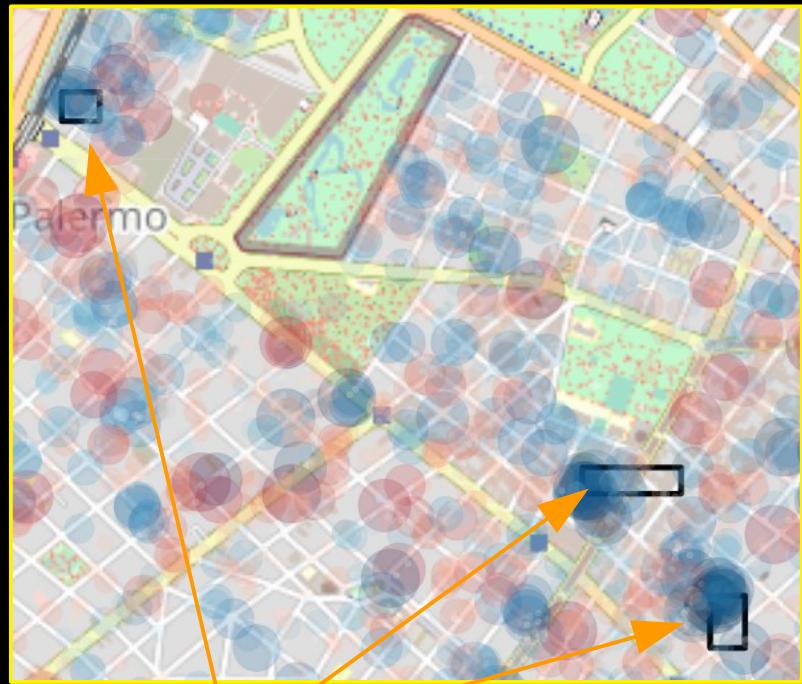
Segundo
Recorte

Tercer
Recorte



Abu-Hanna, A., Nannings, B., Dongelmans, D., & Hasman, A. (2010). PRIM versus CART in subgroup discovery: When patience is harmful. *Journal of Biomedical Informatics*, 43(5), 701-708.

Detección de Bumps con PRIM



Cajas de
Residuos Altos
(Sobre-
valuaciones)



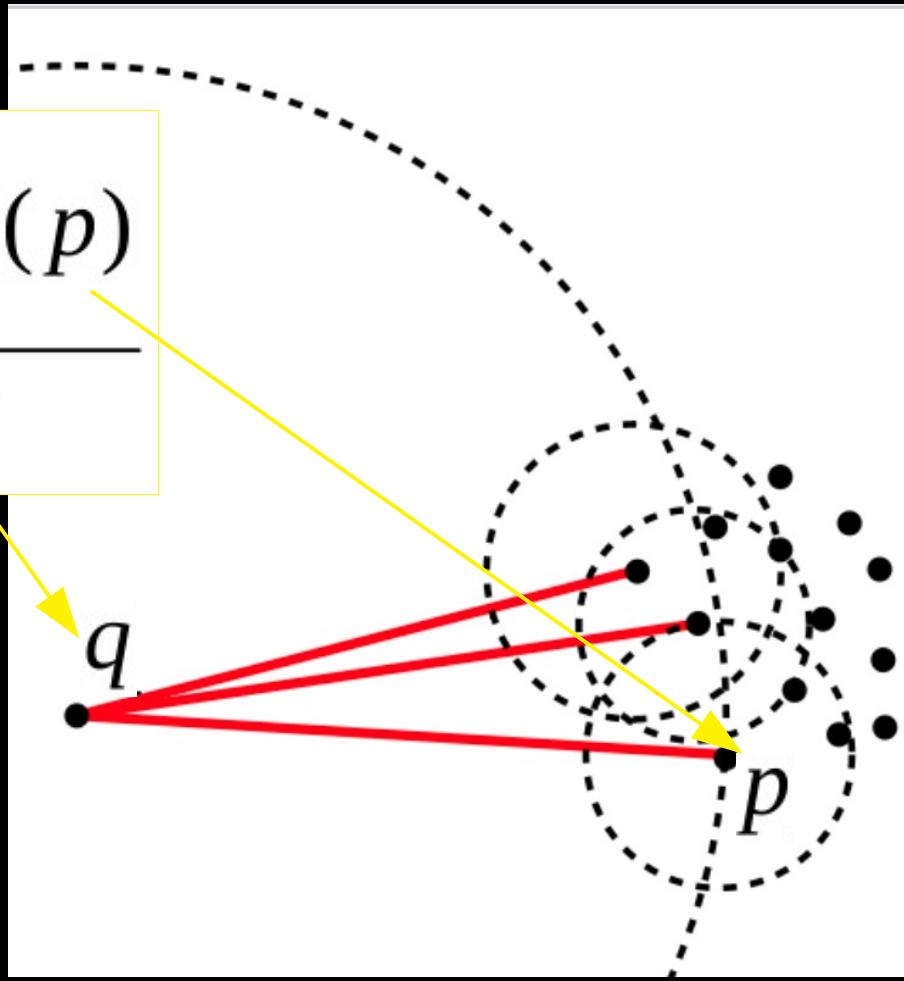
Cajas de Residuos Bajos
(Sub-valuaciones)

Detección de Anomalías

Local Outlier Factor

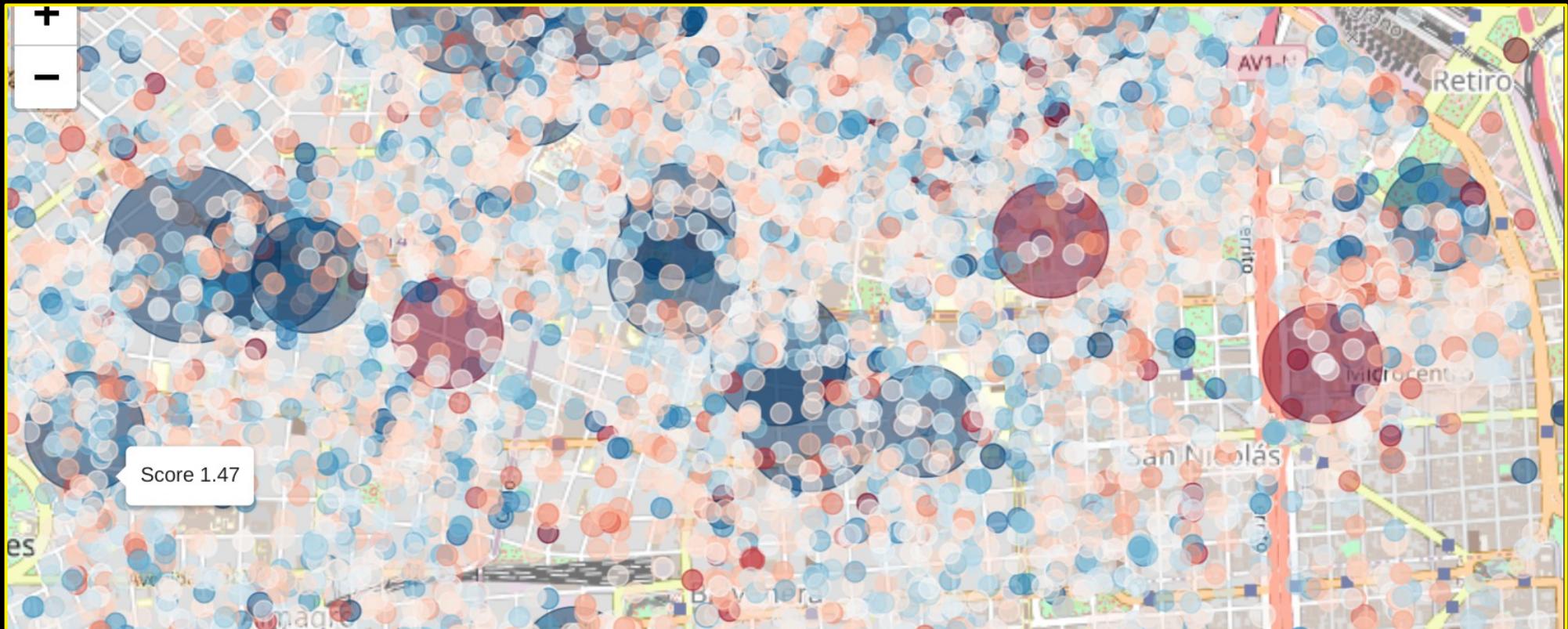
$$LOF(q) = \frac{\frac{1}{k} \sum_{p \in kNN(q)} lrd(p)}{lrd(q)}$$

local reachability
density (lrd)

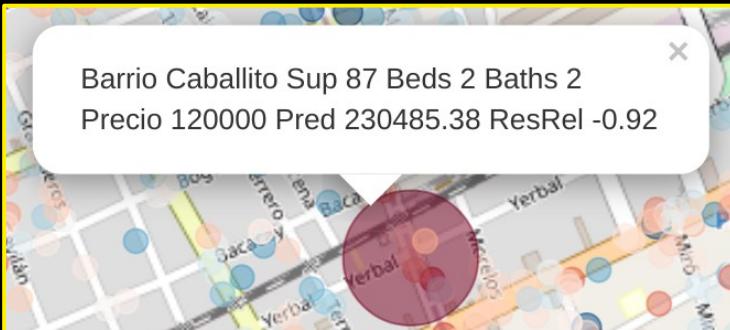
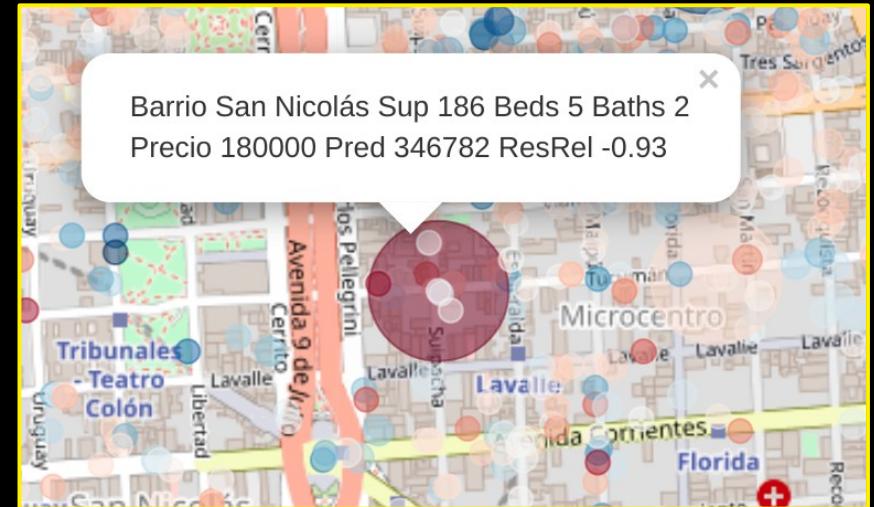
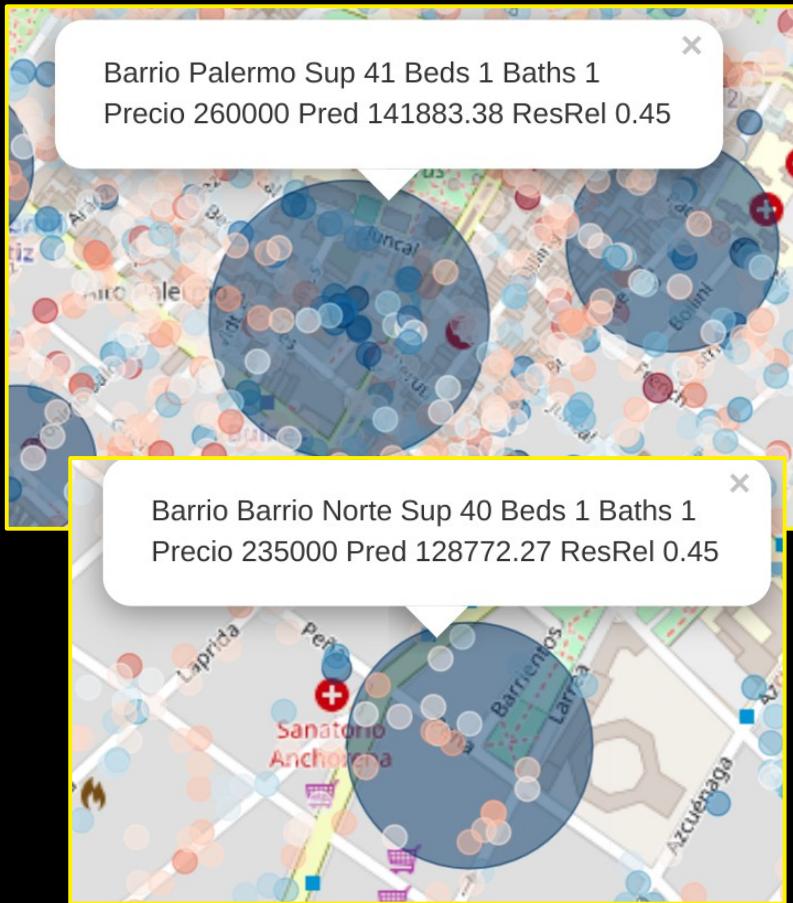


Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104).

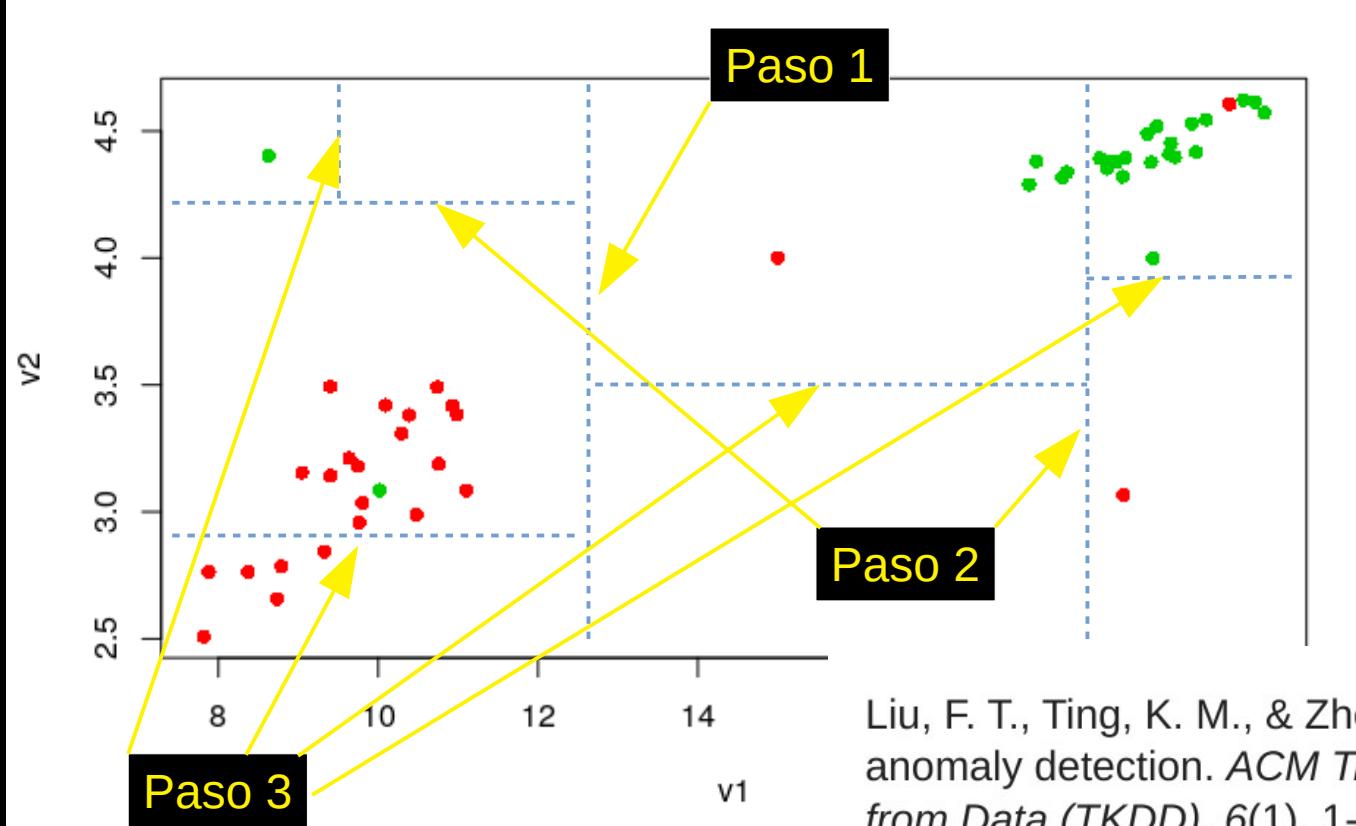
Anomalías Detectadas



Anomalías Detectadas



Detección de Anomalías Isolation Forests (iForests)

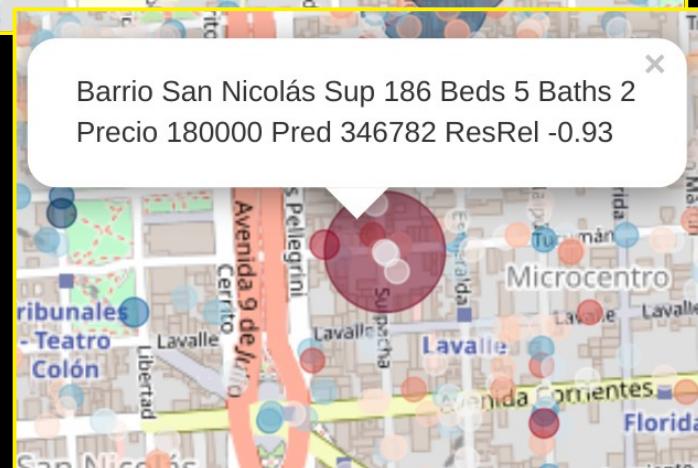
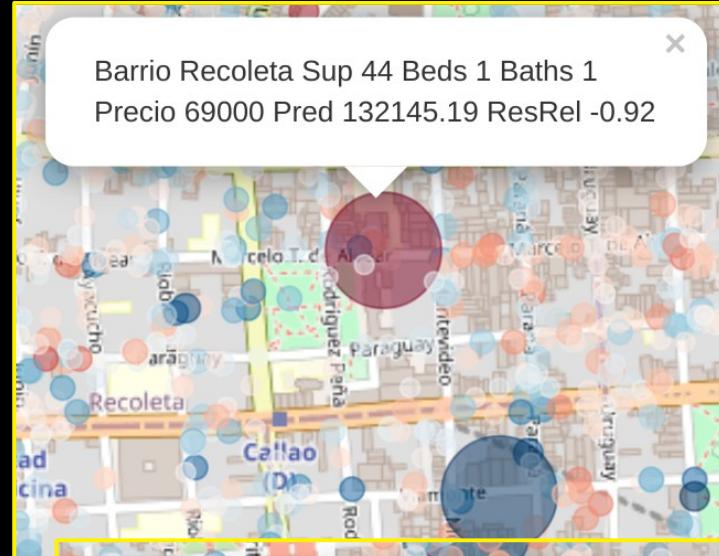
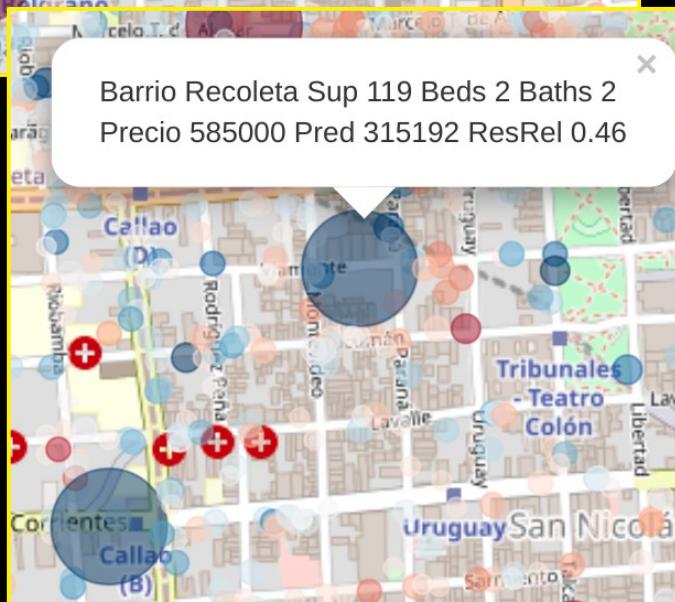
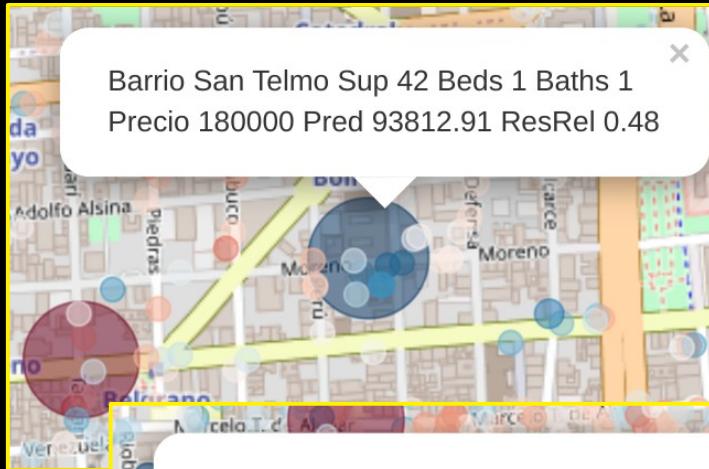


Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1), 1-39.

Anomalías Detectadas



Anomalías Detectadas



Gracias !

References

- Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., ... & Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1), 1-15.
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310.
- Halevy, A., Norvig, P., & Pereira, F. (2009).
The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8-12.
- Predicting vs. Explaining
And Why Data Science Needs More “Half-Bayesians”
(<https://towardsdatascience.com/predicting-vs-explaining-69b516f90796>)