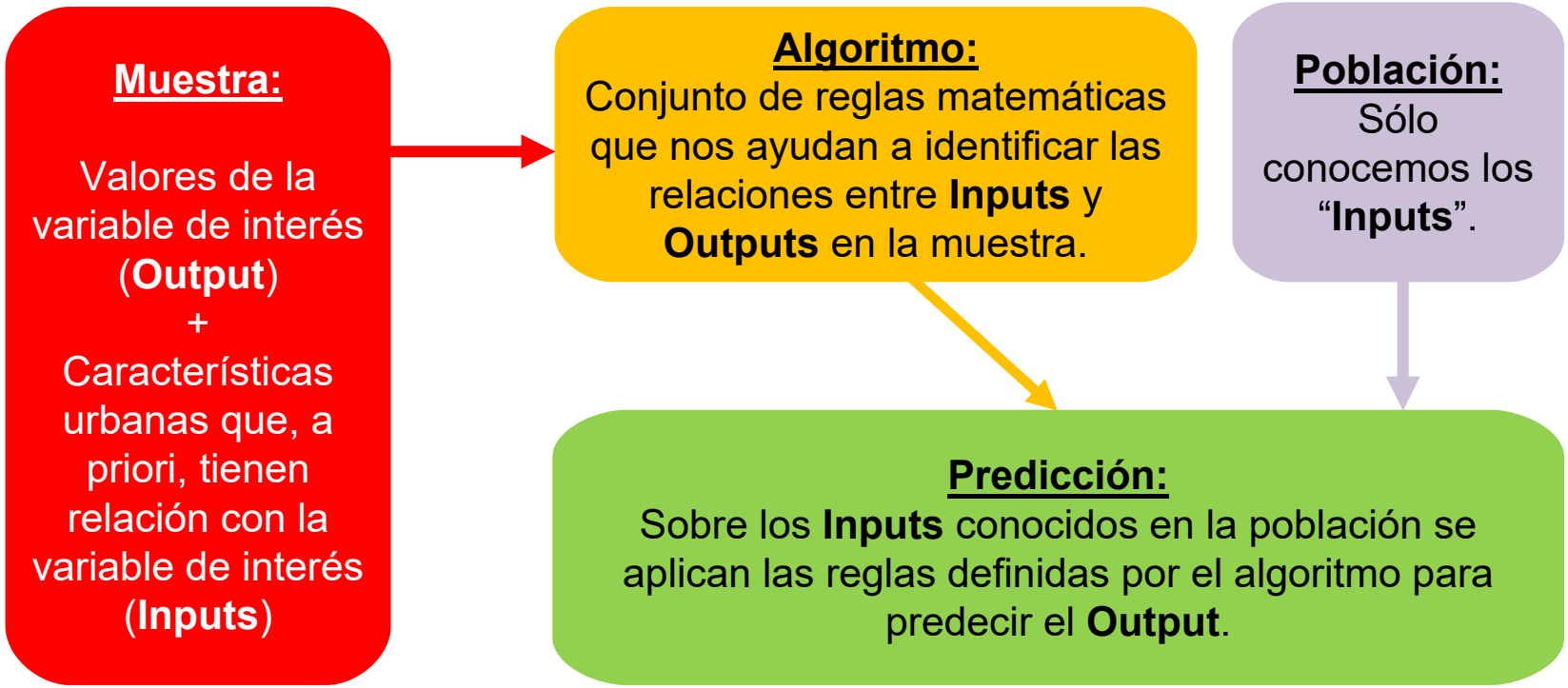
Aplicación de **Inteligencia Artificial** en avalúos masivos.

Mgter. Juan Pablo Carranza
Jefe de Modelización y Métodos Estadísticos
Proyecto de Estudio Territorial Inmobiliario
Gobierno de la Provincia de Córdoba, Argentina.

XI Simposio CPCI
Cancun, Mx, 5 al 7 de setiembre de 2018

¿En qué consiste el aprendizaje estadístico?



Abordaje de la estadística clásica.

Se impone una forma funcional al problema de estudio:

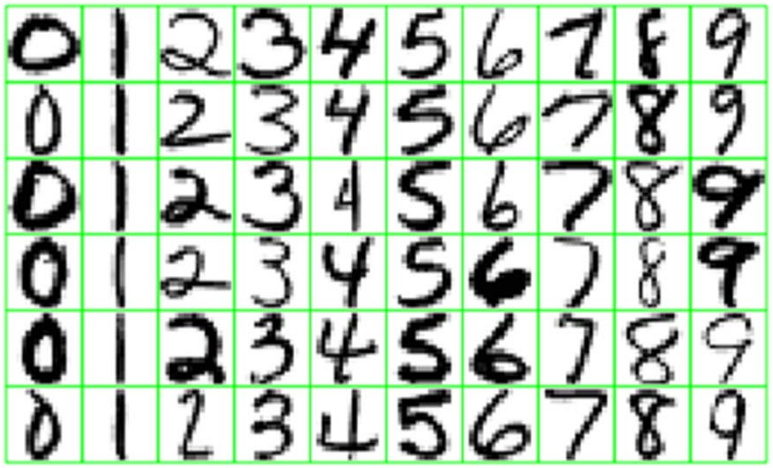
$$y = b_0 + b_1 X_1 + \ldots + b_n X_n + x$$

Abordaje algorítmico, aprendizaje estadístico.

Se respeta la estructura de la información, la no-linealidad, las propiedades emergentes propias de fenómenos caóticos.

La [IA] cada vez más a nuestro alrededor… algunas aplicaciones:

Ejemplo clásico!
Digitalizando el mundo físico.

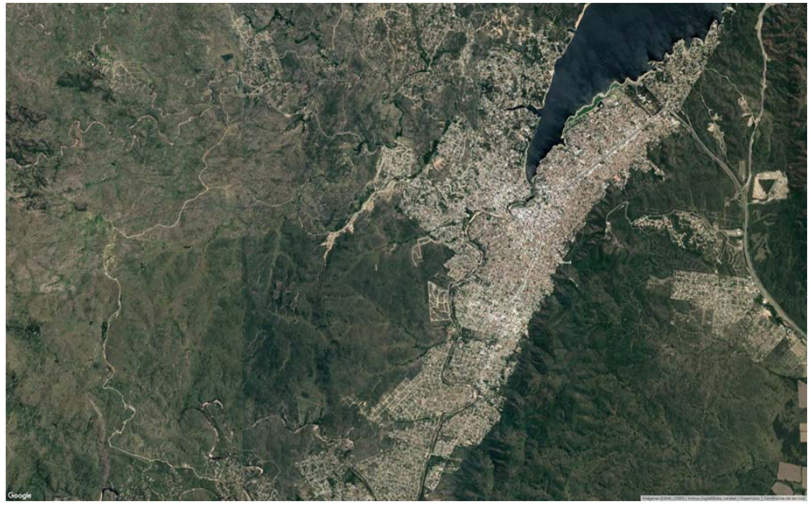


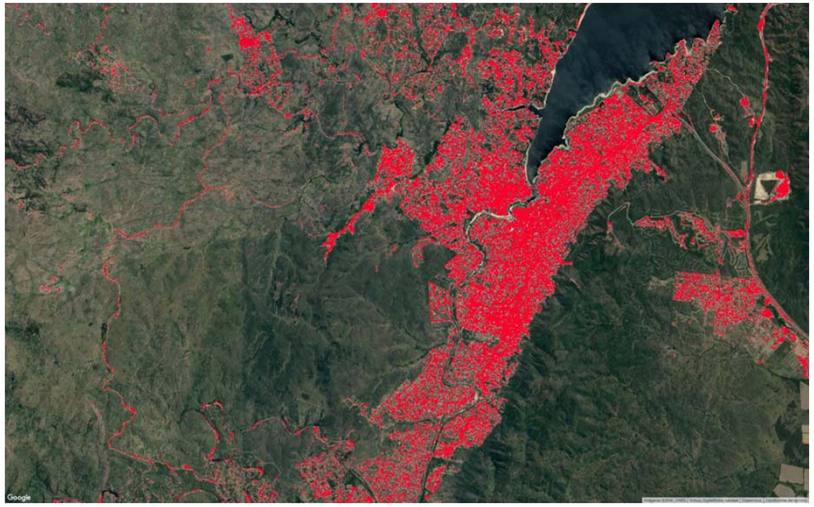
Tomado de Hastie, Tibshirani, Friedman: “The Elements of Statistical Learning”… Recomiendo!!!

En la actualidad?
Foco en la interacción con el ser humano.

Tomado de Iizuka (et. al.): “Globally and Locally Consistent Image Completion”.

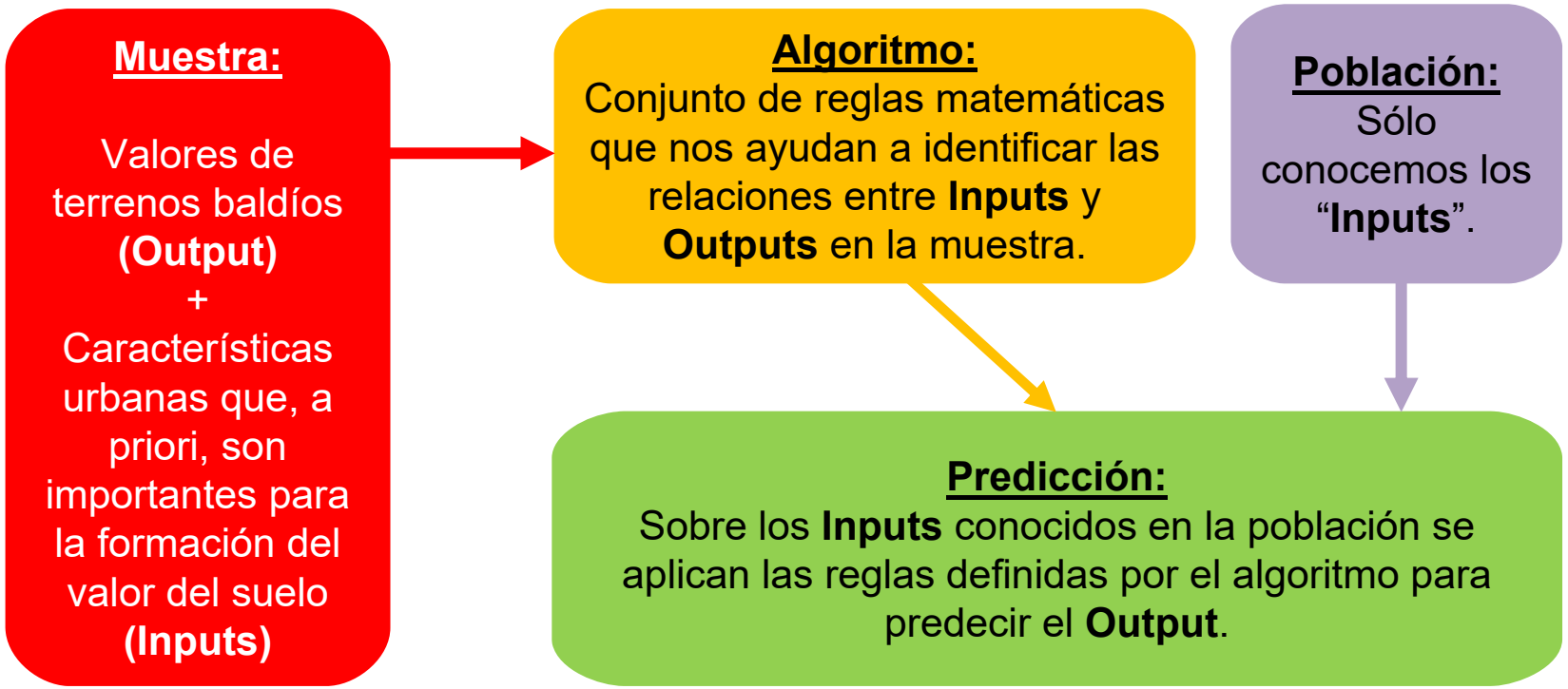
La [IA] cada vez más a nuestro alrededor… acercándonos al estudio territorial:





En base a muestras, el algoritmo puede clasificar las áreas construidas en una ciudad

¿Cómo se aplica al estudio del valor del suelo?



¿Cuál será nuestro "output"?
No todas las muestras de mercado son iguales!

Recurrimos a la **econometría espacial** para homogeneizar valores:

$$\log(y) = b_0 + b_1 X + b_2 W y + b_3 W u + h$$

Diferenciando la expresión con respecto a X:

$$b_1 \; @ \; +gy/y) \; / \; gX$$

¿Cuáles serán nuestros “inputs”?

Catastrales de entorno:
densidad construida en el entorno.
disponibilidad de baldíos en el entorno.
tamaño promedio del lote en el entorno.
Dinámica inmobiliaria en el entorno.

Distancias:
al centro.
a vías principales.
a vías secundarias.
a zonas de bajo perfil inmobiliario
a zonas de alto perfil inmobiliario.
al río.
a vías de FFCC.
a la ruta.
a zonas de depreciación.
etc...

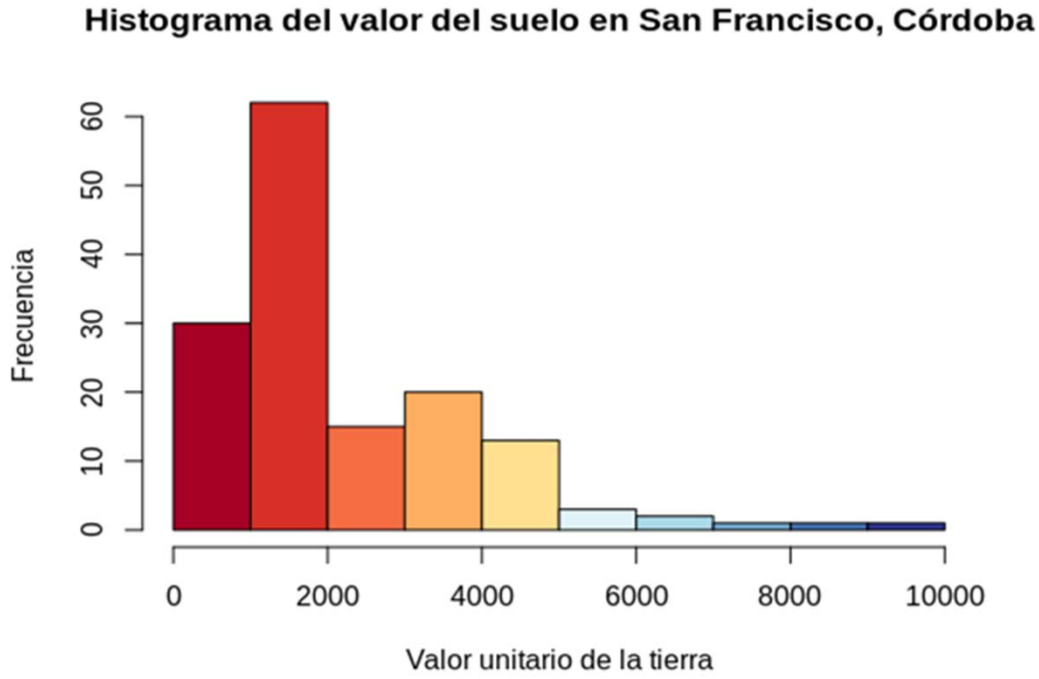
Satelitales de entorno:
área construida
área no construida
dimensión fractal
índices de fragmentación

Terminando de conformar nuestra base de datos… ¿Cómo sabemos que tenemos datos de buena calidad?

Estadística clásica: Eliminar outliers.

¡NO!

El problema no sigue una distribución normal e imponer esa condición limita el análisis.

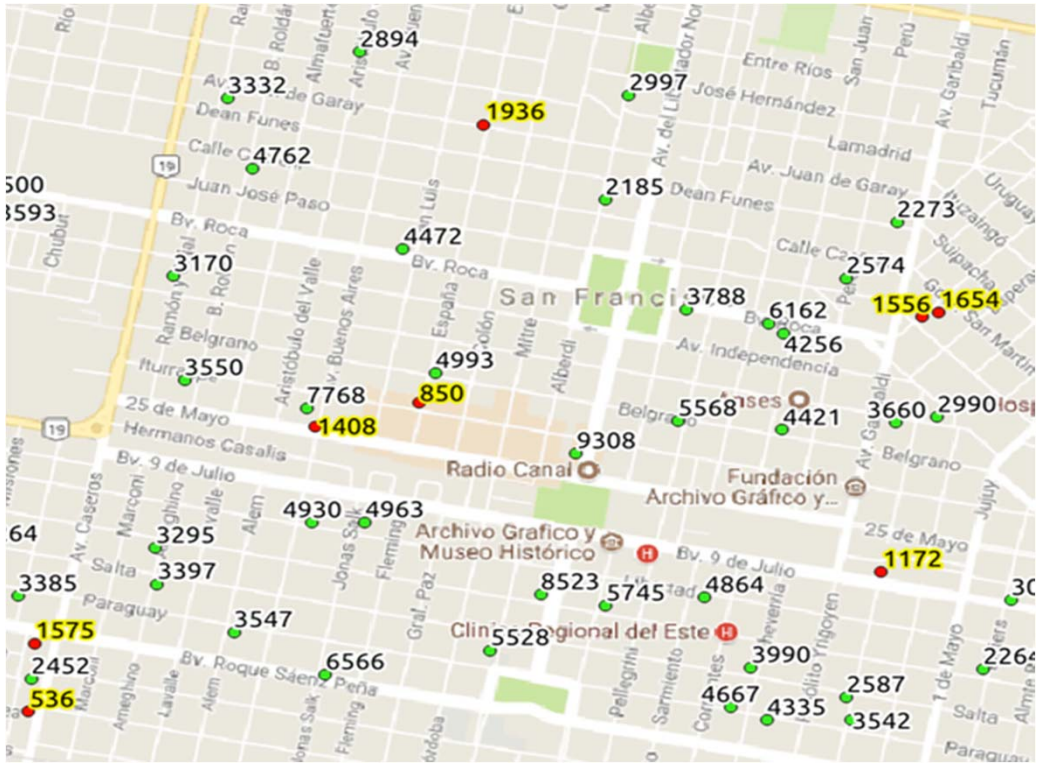


Terminando de conformar nuestra base de datos… ¿Cómo sabemos que tenemos datos de buena calidad?

Sí eliminamos Outliers Espaciales (o inliers) mediante el **índice de Moran local**

$$I = \frac{\frac{N}{S_0}\sum_i \sum_j W_{ij} Z_i Z_j}{\sum_i Z_i^2}$$

Outlier espacial: dato atípico en su entorno.



Pero… ¿qué es un algoritmo?
Por ejemplo: Un árbol de regresión

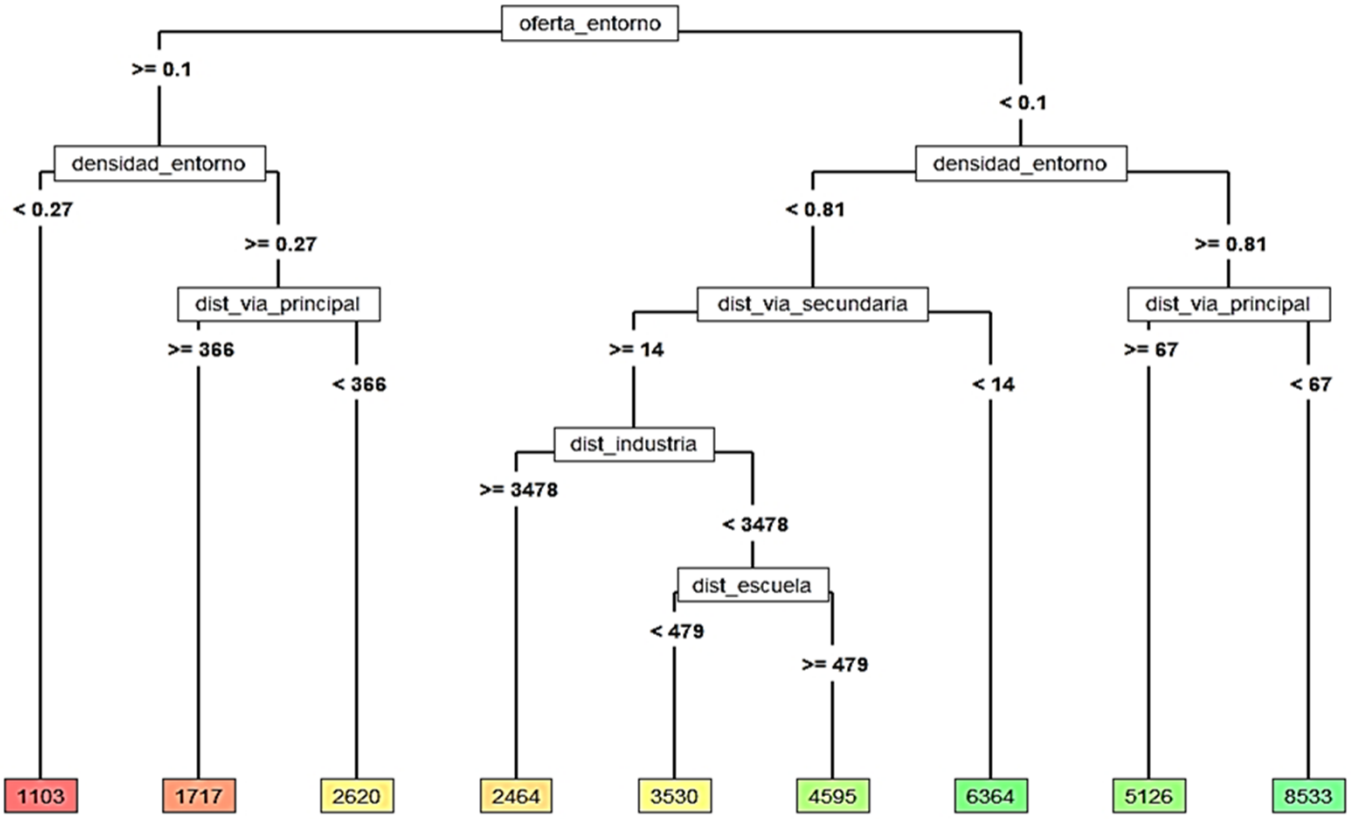
$$R_1(j,s) = \{X | X_j \leq s\} \text{ and } R_2(j,s) = \{X | X_j > s\}$$

$$\min_{j,\,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right]$$

$$\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j,s))$$

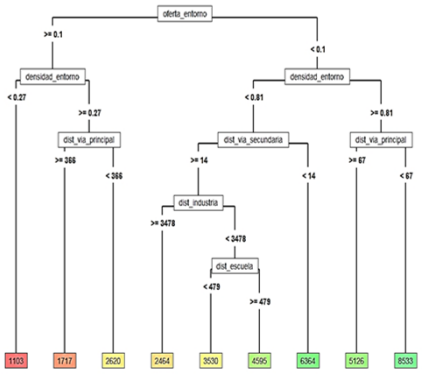
Esto es un árbol de regresión!
Su misión es separar grupos de terrenos baldíos en grupos lo más **homogéneos** posibles y lo más **heterogéneos** con el resto.

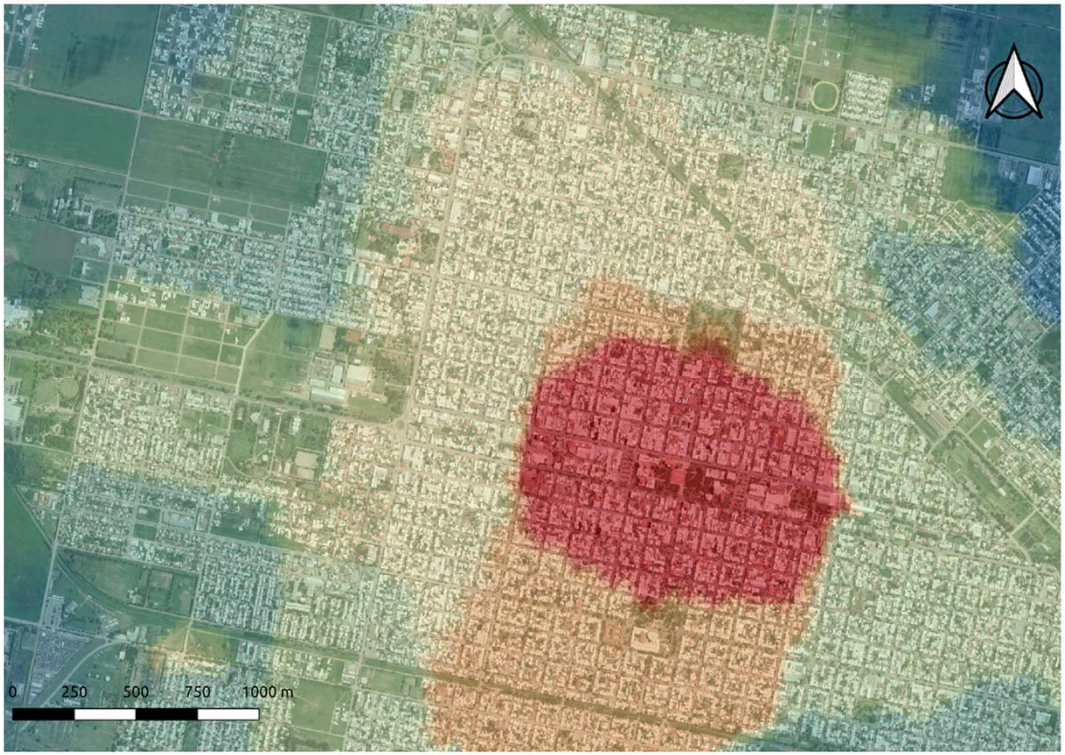
Visualmente es más fácil de comprender...



Not such an intelligent intelligence! Damn it!
Problema: Overfitting.

La predicción del árbol de regresión devuelve unas pocas zonas con valor homogéneo





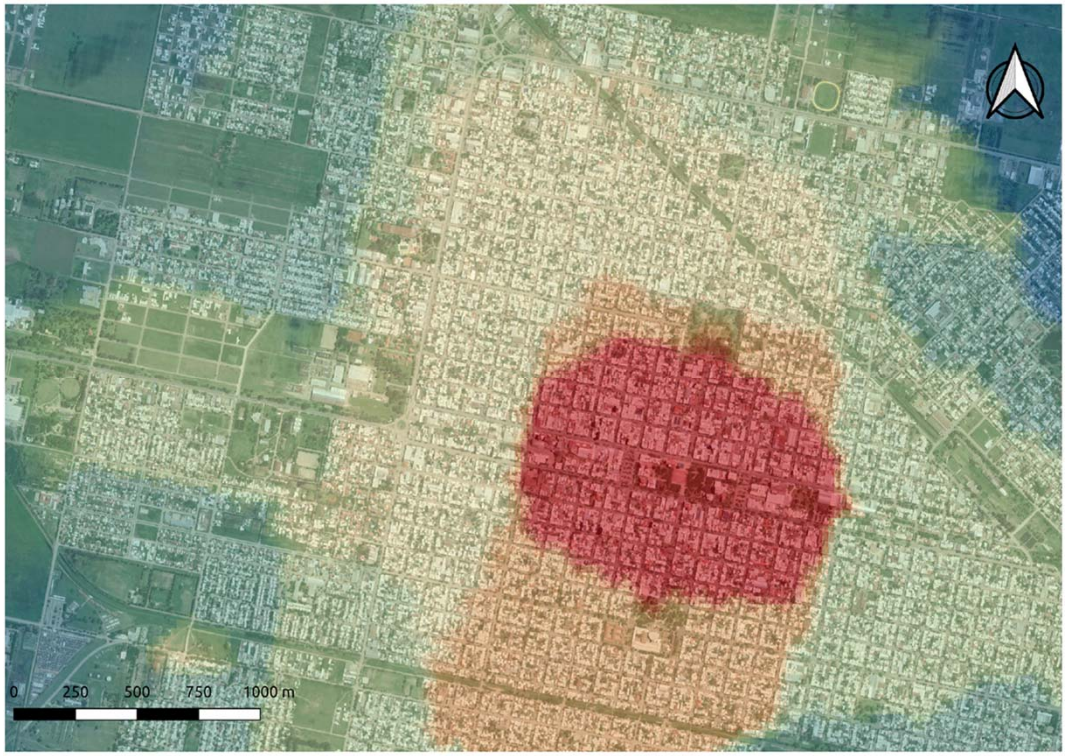
¿Cómo solucionamos el Overfitting del árbol de regresión?
Inventamos muchos árboles para intentar simular las situaciones no observadas en la muestra.

Pero… ¿Cómo hacemos si ya hemos usado todos los datos y todos los inputs?
 # Usamos menos datos (con reposición) en cada árbol (bootstrap)
 # Usamos menos inputs en cada nodo.

¿Y cómo unificamos las predicciones de todos estos árboles?
Hay muchas formas. La más sencilla: promediamos las predicciones de cada uno de los árboles para cada uno de los puntos a predecir.

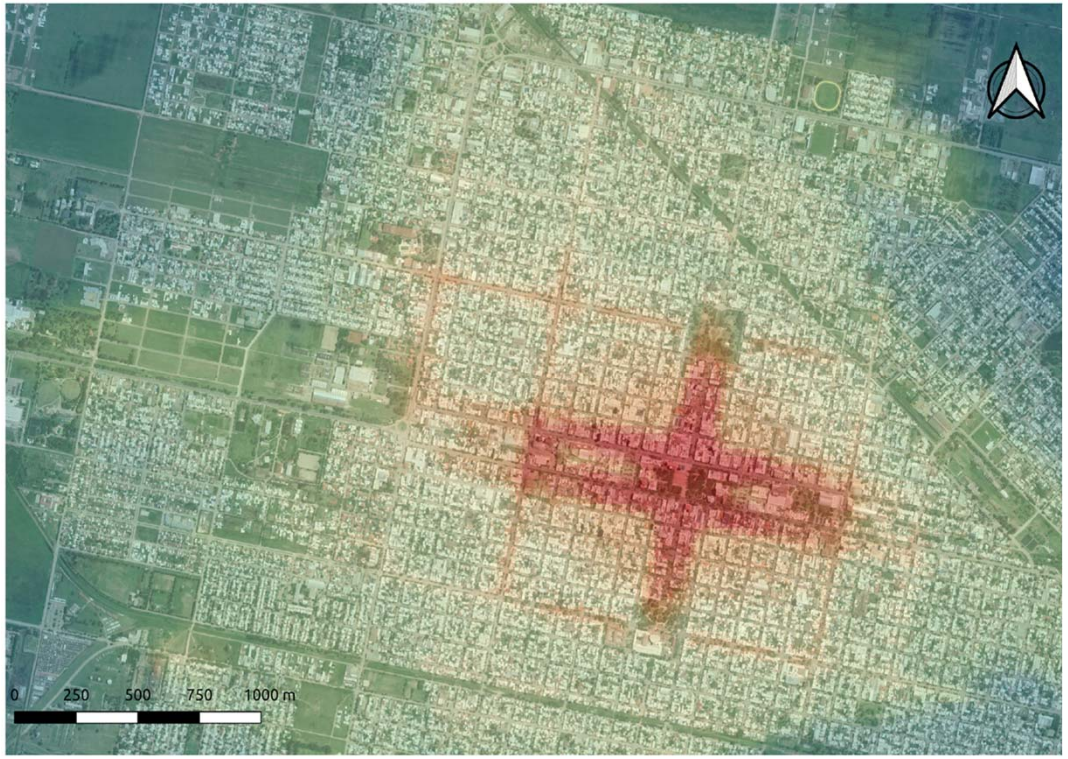
Haciendo esto aplicamos una técnica llamada: RANDOM FOREST

Aplicando **Random Forest** pasamos de esta situación….



… A una estimación mucho más consistente.

Random Forest generaliza mejor.



Ventajas de Random Forest

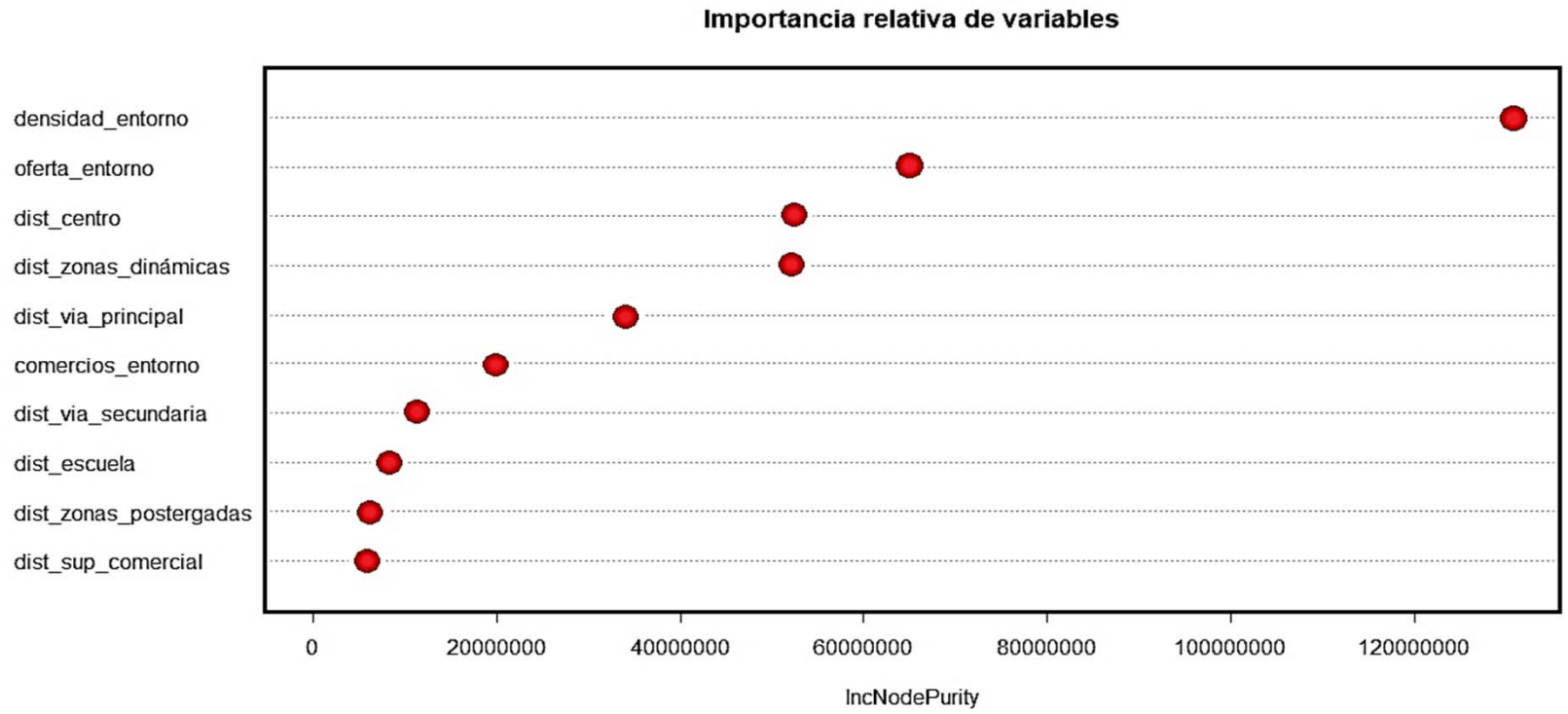
- ❑ Reduce la varianza en la estimación (es decir, es menos propenso al overfitting).
- ❑ Logra una calidad predictiva mucho más elevada. Generaliza mejor!

Desventajas de Random Forest

- ❑ Ya no hay sólo un árbol que nos permita comprender esquemáticamente cómo se conforma el valor del suelo en función de los inputs utilizados.

Su utilización dependerá de las características del problema de investigación.

Sin embargo, sí podemos conocer con mucha precisión cuál es la importancia relativa de cada input en la predicción:



¿Cómo medimos el error de predicción de nuestras estimaciones?

Validación Cruzada	Grupos									
Acción:	1	2	3	4	5	6	7	8	9	10
Sale el grupo 1	**1**	2	3	4	5	6	7	8	9	10

Se estima el modelo con los datos de los grupos 2 a 10

Se predice el grupo 1 y se mide el error de predicción

Sale el grupo 2	1	**2**	3	4	5	6	7	8	9	10
Sale el grupo 3	1	2	**3**	4	5	6	7	8	9	10
Sale el grupo 4	1	2	3	**4**	5	6	7	8	9	10
Sale el grupo 5	1	2	3	4	**5**	6	7	8	9	10
Sale el grupo 6	1	2	3	4	5	**6**	7	8	9	10
Sale el grupo 7	1	2	3	4	5	6	**7**	8	9	10
Sale el grupo 8	1	2	3	4	5	6	7	**8**	9	10
Sale el grupo 9	1	2	3	4	5	6	7	8	**9**	10
Sale el grupo 10	1	2	3	4	5	6	7	8	9	**10**

Se calculan diferentes medidas de error…

- Error relativo promedio en valor absoluto:

$$ERP = \frac{\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i}\right)}{n}$$

- Coeficiente de variación (IAAO):

$$CV = \frac{\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i} - \frac{\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i}\right)}{n}\right)}{\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i}\right)}$$

Evaluación comparativa de diferentes modelos aplicados (caso Ciudad de San Francisco):

Console | Terminal

~/

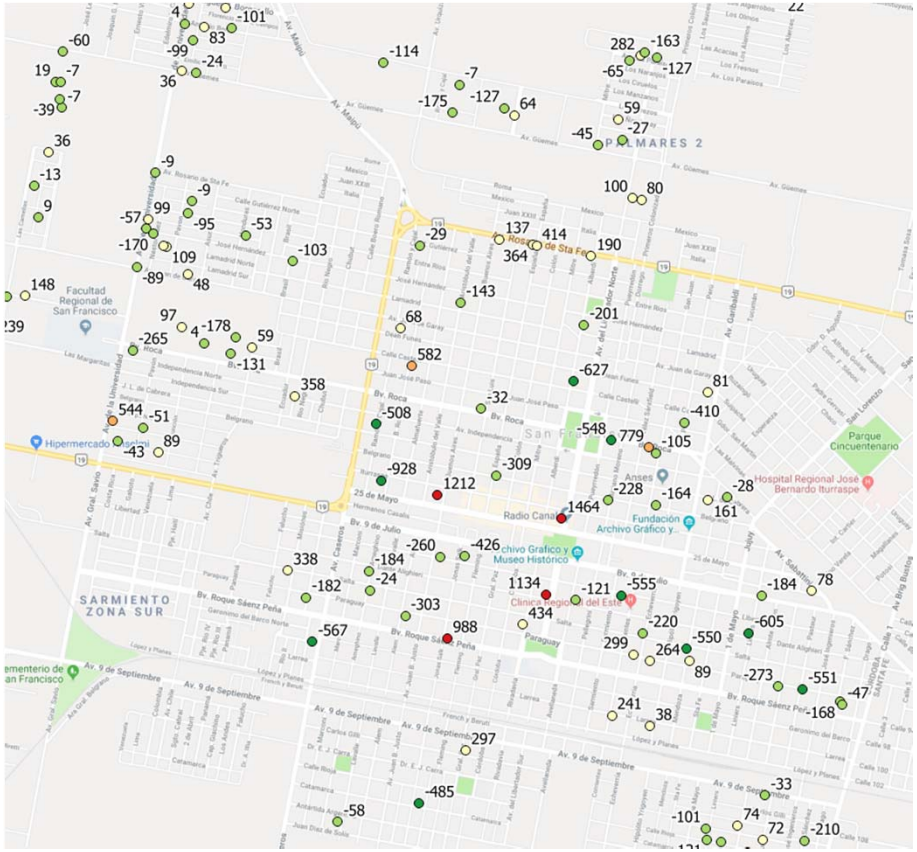
>

Modelo:	ERP	PROMEDIO	MEDIANA	CV	CD
Random Forest + Kriging	0.1960	1.0597	1.0328	0.1841	0.1878
Gradient Boosting + Kriging	0.1993	1.0771	1.0452	0.1819	0.1860
Random Forest	0.2009	1.0716	1.0479	0.1839	0.1866
Ensamble Machine Learning + Kriging	0.2021	1.0592	1.0034	0.1950	0.2014
Gradient Boosting	0.2049	1.0676	1.0505	0.1873	0.1897
Ensamble Machine Learning	0.2071	1.0520	1.0108	0.2001	0.2046
Kriging Ordinario	0.2133	1.0559	1.0401	0.1999	0.2027
Redes Neuronales + Kriging	0.2224	1.0623	1.0240	0.2107	0.2162
Regression Kriging	0.2376	1.0555	1.0170	0.2271	0.2332
Partial Least Squares + Kriging	0.2379	1.0505	1.0027	0.2300	0.2373
Redes Neuronales	0.2387	1.0605	1.0042	0.2276	0.2377
CART	0.2583	1.0600	1.0111	0.2457	0.2551
Supported Vector Regression + Kriging	0.3472	1.1972	1.1391	0.2811	0.2906
Supported Vector Regression	0.4054	1.2507	1.1753	0.3044	0.3193

Incorporación de la Auto-Correlación espacial:

La estimación realizada mediante Random Forest tiene un error intrínseco.
Para reducir su impacto, se realiza un Kriging de los errores y esta nueva estimación se suma a la estimación original.

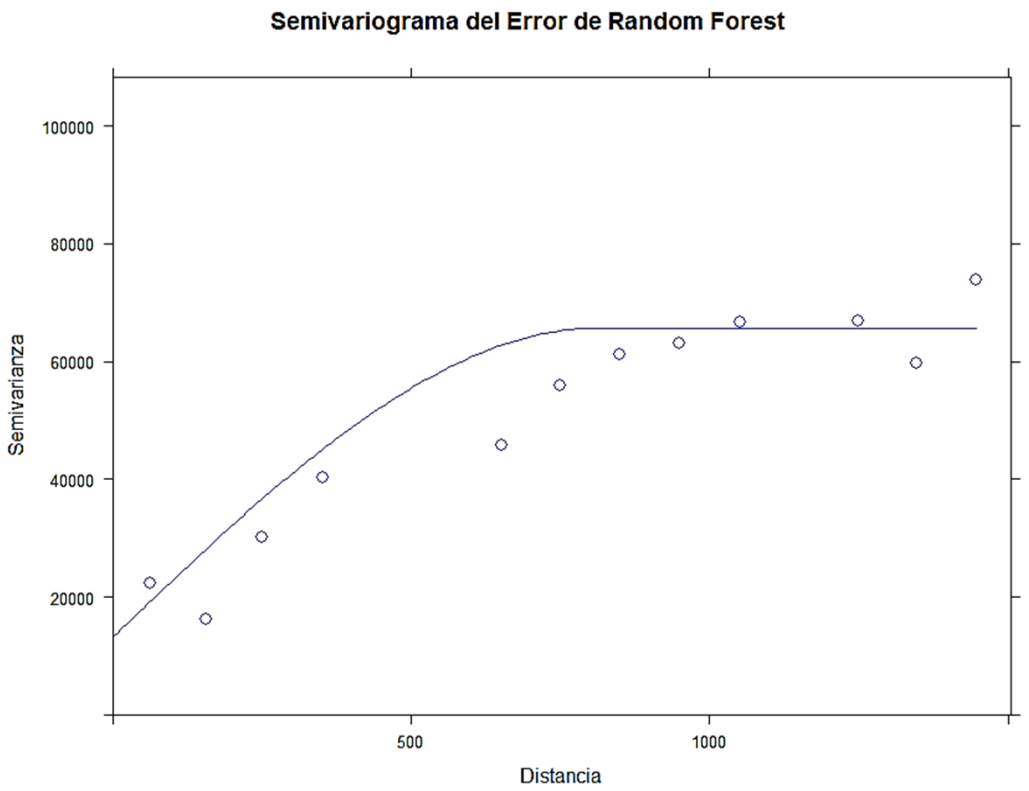
Estimación final = Estimación Random Forest + Kriging Ordinario del error.



Interpolación del ajuste del error:

A los fines de ajustar el error de la predicción inicial se utiliza la dependencia espacial de las observaciones, capturadas mediante el semivariograma.

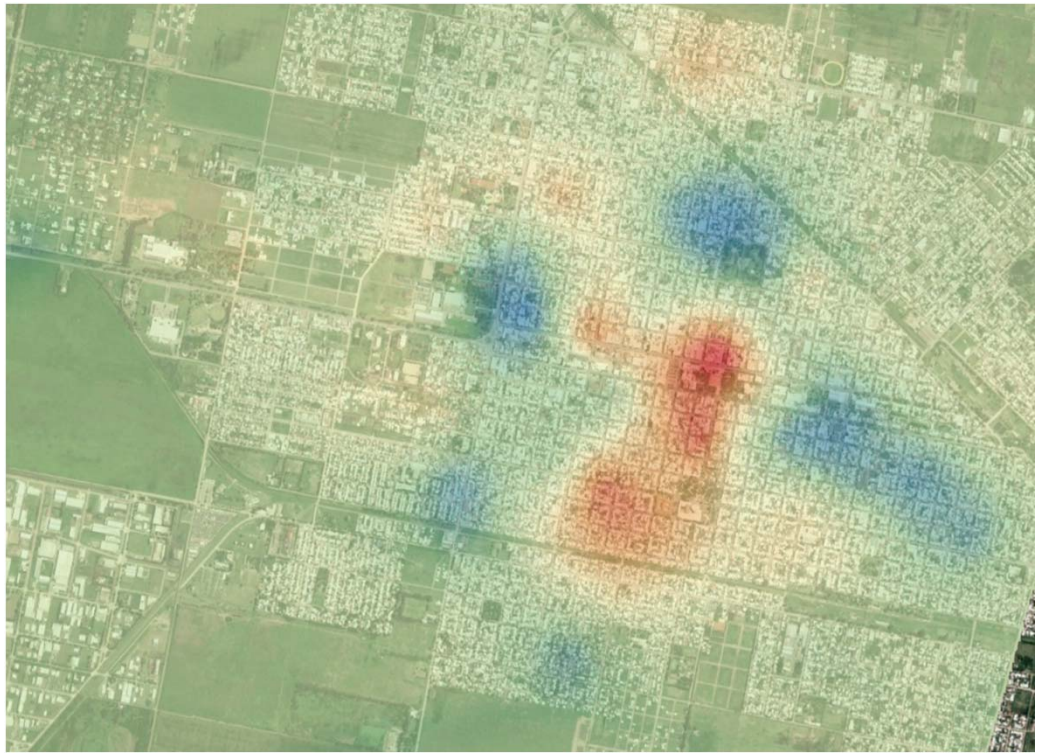
Ley de Tobler: Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes.



Interpolación del ajuste del residuo:

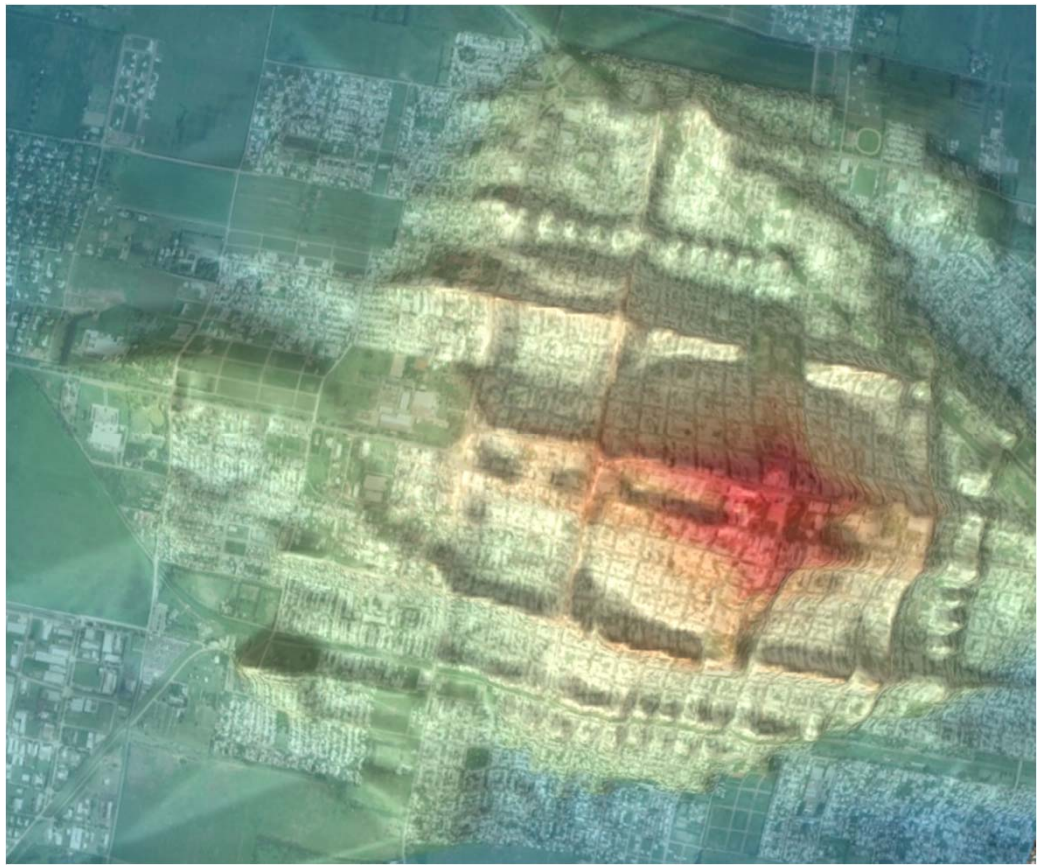
En rojo: Ajusta la predicción hacia arriba.

En azul: Ajusta la predicción hacia abajo.



Estimación final Random Forest + Kriging ordinario:

A la predicción original se le suma el Kriging del Error.



Evaluación comparativa de diferentes modelos aplicados (caso Ciudad de San Francisco):

Console | Terminal

```
~/
>
```

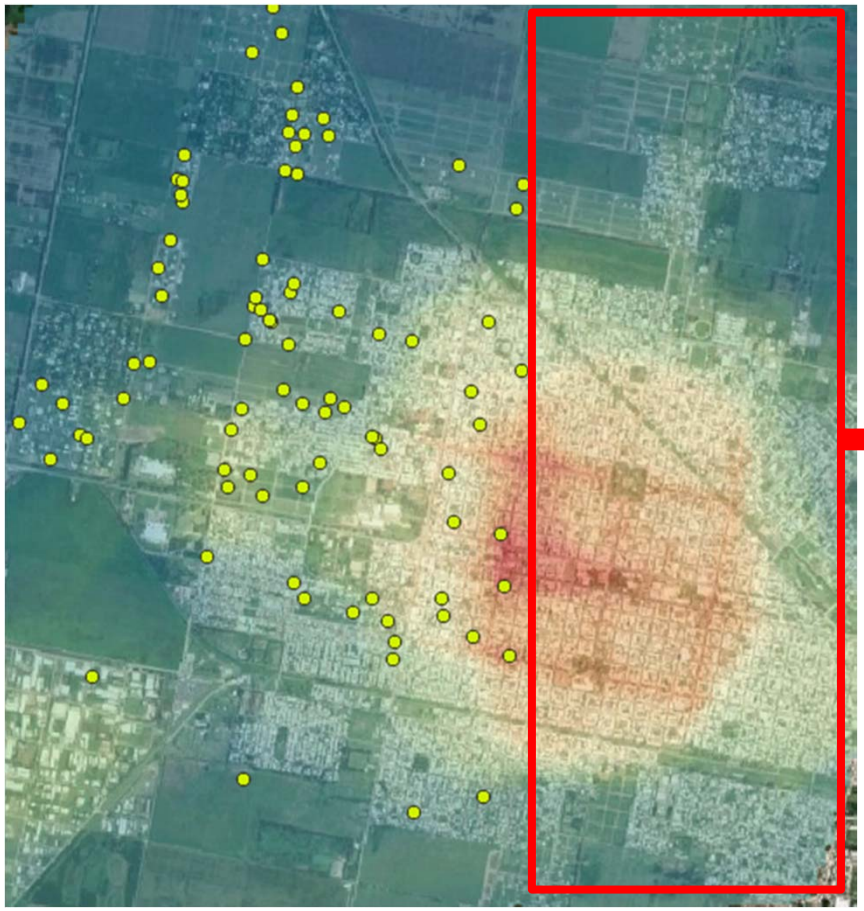
Modelo:	ERP	PROMEDIO	MEDIANA	CV	CD
Random Forest + Kriging	0.1960	1.0597	1.0328	0.1841	0.1878
Gradient Boosting + Kriging	0.1993	1.0771	1.0452	0.1819	0.1860
Random Forest	0.2009	1.0716	1.0479	0.1839	0.1866
Ensamble Machine Learning + Kriging	0.2021	1.0592	1.0034	0.1950	0.2014
Gradient Boosting	0.2049	1.0676	1.0505	0.1873	0.1897
Ensamble Machine Learning	0.2071	1.0520	1.0108	0.2001	0.2046
Kriging Ordinario	0.2133	1.0559	1.0401	0.1999	0.2027
Redes Neuronales + Kriging	0.2224	1.0623	1.0240	0.2107	0.2162
Regression Kriging	0.2376	1.0555	1.0170	0.2271	0.2332
Partial Least Squares + Kriging	0.2379	1.0505	1.0027	0.2300	0.2373
Redes Neuronales	0.2387	1.0605	1.0042	0.2276	0.2377
CART	0.2583	1.0600	1.0111	0.2457	0.2551
Supported Vector Regression + Kriging	0.3472	1.1972	1.1391	0.2811	0.2906
Supported Vector Regression	0.4054	1.2507	1.1753	0.3044	0.3193

Potencia del aprendizaje estadístico:

Capacidad de generalización!

(menos costoso y más preciso)

Ejemplo: Estimación realizada con datos sólo en la mitad oeste de la ciudad.

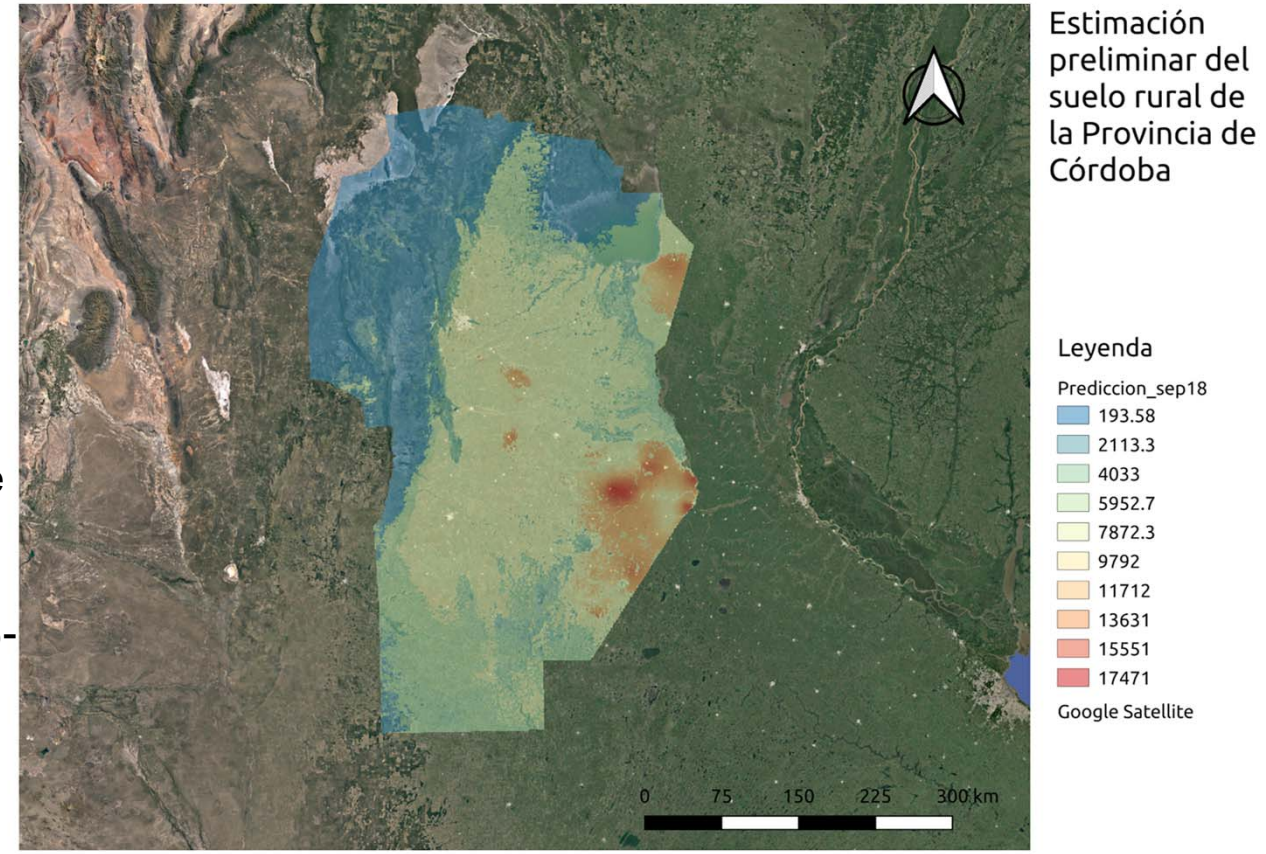


Avances en la estimación del suelo rural:

Métodos más adecuados:

Supported Vector Machine (error 18%)

Random Forest (error 20%-31%)



Variables utilizadas en la estimación….

TIPO	VARIABLE	TIPO	VARIABLE
Suelo	Capacidad de Uso de la Tierra	**Infraestructura**	Distancia a asentamientos urbanos
	Indice de Productividad		Distancia a red vial pavimentada
	Coberturas/Usos de la tierra		Distancia a red de Energía Eléctrica
	Pendiente		Distancia a red de Gas Natural
	Altura		Distancia a centros de acopio
	Suceptibilidad a inundación y/o anegamiento		Distancia a Balanzas Públicas
	Suceptibilidad a erosión eólica		Distancia a puerto San Lorenzo
	Deficiencia de Húmedad		Acceso a riego
	NDVI		Acceso a riego complentario
Hídrología	Distancia a cursos de agua		Distancia a obras hídricas
	Disponibilidad de agua subterránea	**Estructura productiva**	Producción Tambera
	Profundidad del nivel freático		Producción Ganadera
Climáticos	Precipitación media anual		Actividad turística
	Régimen de Temperaturas		Explotación minera
	Radiación Solar	**Económicas**	Rendimientos zonales por localidad
	Bioclimaticas		Arrendamientos zonales por localidad
	Vulnerabilidad de sequía		

carranzajuanp@gmail.com