Predicción de Temperatura con Cuadrados Mínimos

Ezequiel Alvarez Joel Esteban Cámera Sebastián Sicardi

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Resumen

Cuadrados Mínimos Lineales (CML) es un método comúnmente usado para estimar valores desconocidos a partir de valores conocidos. En este trabajo proponemos modelos para predecir la temperatura del planeta, países y ciudades utilizando el mencionado método.

Keywords: Cuadrados Mínimos, Temperatura, Predicciones

1. Introducción

La temperatura del planeta no es estática. La enorme cantidad de factores a tener en cuenta para poder saber la temperatura en un futuro es de una gran complejidad. El pronóstico de la temperatura es de gran importancia para muchos conjuntos de la sociedad, desde el correcto funcionamiento de grandes plantaciones y granjas hasta para saber cuánto abrigo debo llevar para el día de hoy.

En el presente trabajo, desarrollaremos predicciones de temperatura utilizando el método de cuadrados mínimos lineales intentando anticipar la temperatura de ciudades, países y el planeta.

2. Cuadrados mínimos lineales

El objetivo consiste en ajustar los parámetros de una familia de funciones tal que esta encaje mejor en el conjunto de datos que se posee. [1]

Sea un conjunto simple de datos que consiste en n puntos (en $\mathbb{R}^2$) (x_i, y_i), con $i = 1, ..., n$, done x_i es la variable independiente y y_i es la variable dependiente cuyo valor es dado en la observación; y sea $\{f_i(x)\}_{j=1}^{m}$ una familia de funciones (llamamos familia de funciones a una combinación lineal de una base de funciones). Queremos encontrar una función $f(x)$ que sea combinación lineal de la familia de funciones de modo que $f(x_k) \approx y_k$, donde:

$$f(x) = \sum_{j=1}^{m} c_j f_j(x)$$

Por tanto, se trata de hallar los m coeficientes c_j que hagan que la función aproximante $f(x)$ dé la "mejor aproximación" para los puntos dados (x_k, y_k). El criterio de "mejor aproximación" puede variar, pero en general se basa en aquél que minimice una acumulación del error individual (en cada punto) sobre el conjunto total. El error de la función $f(x)$ en un punto (x_k, y_k) se define como:

$$e_k = y_k - f(x_k)$$

El método de cuadrados mínimos lineales alcanza su óptimo cuando el error cuadrático medio es mínimo.

$$\text{ECM} = \frac{1}{n} \sum_{i=1}^{m} (e_i)^2 = \frac{1}{n} \sum_{i=1}^{m} (y_i - f(x_i))^2$$

Si se tratara de hallar el conjunto de coeficientes $\{c_j\}$ tal que $f(x)$ pase exactamente por todos los pares (x_i, y_i), con $i = 1, ..., n$, entonces tendría que cumplirse que:

$$\sum_{j=1}^{m} c_j f_j(x_k) = y_k$$

[1] Tomado de https://en.wikipedia.org/wiki/Least_squares

Que en forma matricial se expresa como:

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & \dots & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \dots & f_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_n) & f_2(x_n) & \dots & f_m(x_n) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = A.c = b$$

Esto establece un sistema de n ecuaciones y m incógnitas, y como en general $n > m$, quedaría sobredeterminado y no tendría siempre una solución general. Por lo tanto, la aproximación tratará en realidad de hallar el vector c que mejor aproxime $A.c = b$.

3. Temperatura global

En esta sección intentaremos ver como se podría predecir la temperatura de un año usando los años anteriores. Para esto usamos los datos de la *National Centers for Environmental Information* [2] . Estos consisten de la temperatura promedio anual desde 1880 hasta 2012 y podemos observarlos en la figura 1, donde también exhibimos la media móvil a derecha (el promedio de los 5 años anteriores, incluyendo al mismo, para cada año) para notar más claramente la tendencia.

Para construir un modelo interesante, lo primero que realizamos fue el coeficiente de correlación de Pearson [3] entre la temperatura t, el año y la temperatura de los 5 años anteriores:

	año	t_{i-1}	t_{i-2}	t_{i-3}	t_{i-4}	t_{i-5}
correlación con t	0,7661	0,6595	0,6645	0,6027	0,6547	0,6531

Aunque el año muestra una fuerte relación con la temperatura, decidimos no utilizarlo en este análisis e intentar predecir una fecha futura solo con las temperaturas pasadas.

Además, como los datos son muy ruidosos, decidimos utilizar la media móvil como el dato para cada año, por lo que por ejemplo las primeras filas quedarían:

[2] https://www.ncdc.noaa.gov/
[3] https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

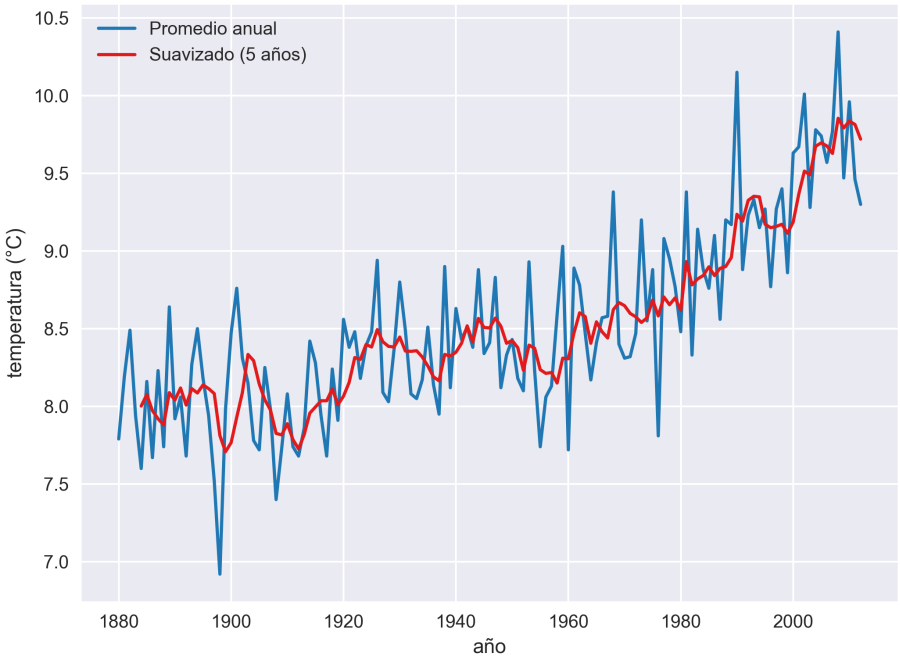

Figura 1. Temperatura superficial promedio por año. En rojo, la media móvil para los 5 años anteriores.

año	m	m_{i-1}	m_{i-2}	m_{i-3}	m_{i-4}	m_{i-5}
1890	8,040	8,088	7,880	7,920	7,972	8,076
1891	8,118	8,040	8,088	7,880	7,920	7,972
1892	8,008	8,118	8,040	8,088	7,880	7,920
1893	8,114	8,008	8,118	8,040	8,088	7,880
1894	8,086	8,114	8,008	8,118	8,040	8,088

Donde m es la media móvil para el año y $m_{i-1}, ..., m_{i-5}$ es la de los años anteriores. También tomamos como fecha inicial a 1890 tal que todos los años tengan esta información previa disponible.

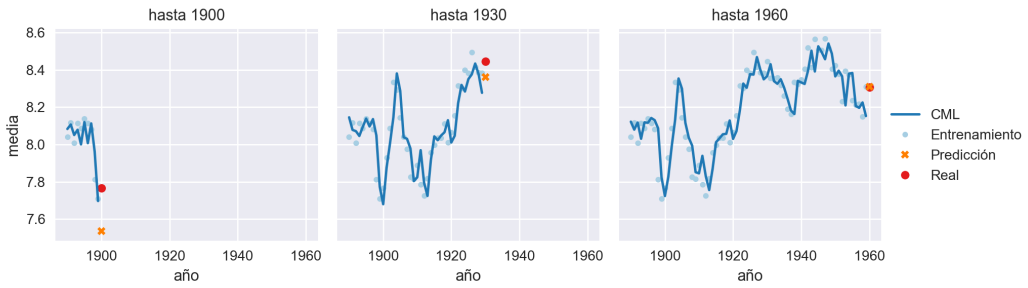

Figura 2. Algunos cortes de la validación cruzada. En naranja y rojo la temperatura predicha y real del año en cuestión respectivamente.

Proponemos entonces la siguiente familia de funciones:

$$f(X^{(i)}) = c_0 + \sum_{j=1}^{5} c_j X_j^{(i)}$$

Donde $X^{(i)} = (m_{i-1}, m_{i-2}, m_{i-3}, m_{i-4}, m_{i-5}), \forall i = 1, ..., n$ tiene las medias pasadas; $c_0, c_1, ..., c_5$ son los coeficientes de cuadrados mínimos; y la variable dependiente es:

$$Y = (m_1, m_2, ..., m_n)$$

Para evaluar la *performance* de la misma procedemos a hacer validación cruzada usando todas las fechas menores a un año y prediciendo contra este, $\forall a \in \{1900, 1901, ..., 2010\}$, como propone Rob J. Hyndman [4] (empezando en 1900 para tener al menos 20 años de entrenamiento). No olvidar que nuestra variable dependiente no es la temperatura del año siguiente, sino la media del mismo mas los 4 anteriores; en otras palabras estamos prediciendo como cambia la media rodante. En la figura 2 mostramos como se ajusta la función a los datos de entrenamiento y como predice para los años 1900, 1930 y 1960. Finalmente en esta tabla los resultados del experimento completo:

ECM_f	0,012628
error máximo	0,110602
error mínimo	8,52e−6

Como punto de comparación corrimos el mismo test con un polinomio grado 5 que toma el año como variable, y el error del mismo quedó:

[4] https://robjhyndman.com/hyndsight/tscv/

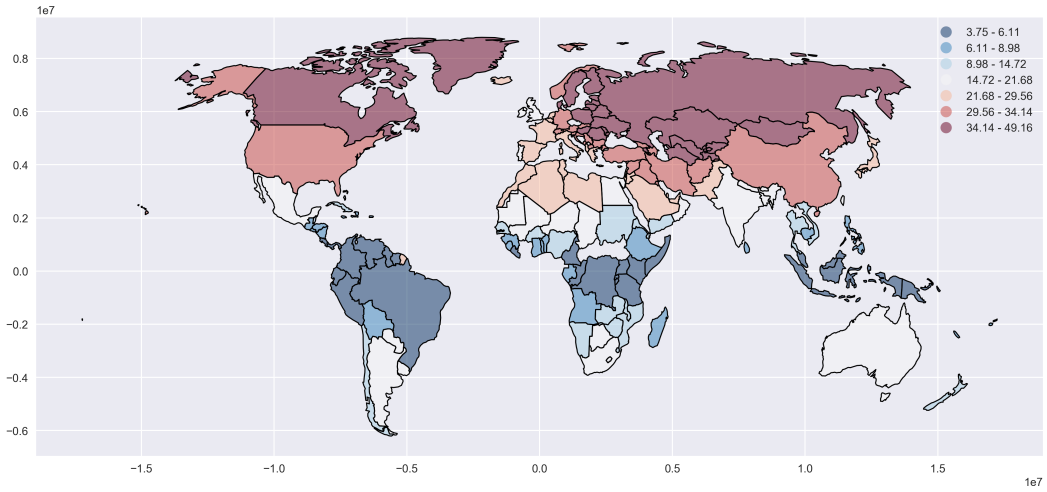

Figura 3. Coropleta de países por amplitud térmica. Un color mas azul indica una amplitud menor, y uno mas rojo una amplitud mayor.

ECM_{P}	0,038682
error máximo	0,210445
error mínimo	2,22e−6

La mejora de nuestro modelo no es grande, pero existe.

4. Predicción de la temperatura con países

Los datos utilizados en esta sección provienen de *Kaggle* [5] y poseen la temperatura promedio por mes de 242 países (con datos completos). El propósito es poder predecir el promedio mundial, calculado ahora con estos datos, de los próximos 10 años con solo un subconjunto de los países.

Para comenzar, observamos la amplitud térmica histórica por país. Para ello, utilizamos los datos geográficos de *Natural Earth* [6]. El resultado se puede contemplar en la figura 3.

A continuación, elegimos 4 países según su ubicación geográfica y amplitud térmica de tal forma de obtener datos heterogéneos:

[5] https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data
[6] http://www.naturalearthdata.com/

	mínima	máxima	amplitud
Canadá	-28,736	14,796	43,532
Reino Unido	-1,473	17,285	18,759
Australia	12,529	29,861	17,332
Brasil	21,797	27,151	5,353

La familia de funciones queda entonces dada por:

$$g(X^{(i)}) = c_0 + c_1 X^{(i)}_{\text{Canadá}} + c_2 X^{(i)}_{\text{UK}} + c_3 X^{(i)}_{\text{Australia}} + c_4 X^{(i)}_{\text{Brasil}}$$

Con $X_p^{(i)}$ la temperatura promedio del país p para el año i, y la variable dependiente el promedio anual entre todos los países.

El rango de fechas elegido abarca de 1860 a 2010 para que todos los países tengan la misma cantidad de datos. Como en la sección anterior, hacemos validación cruzada pero en este caso entrenando con el promedio mundial y nuestros países hasta una fecha, y prediciendo los siguientes 10 años. El experimento completo son 150 cortes y los resultados se ven en la siguiente tabla. Como ejemplo, en la figura 4 vemos el corte hasta 1920. A modo de ver como mejora la predicción cuantos más datos utilizamos, adjuntamos el ECM_{g} por corte en la figura 5.

ECM_{g} promedio	0,493460
error máximo	2,986647
error mínimo	0,041324

Comparar este modelo con el de la sección 3 resulta difícil ya que el primero predice un año en el futuro sin mas información que las temperaturas pasadas, mientras que el otro trata de predecir 10 años con la temperatura de algunos países. Dependiendo de los datos disponibles y que se busque predecir, se utilizará el modelo acorde.

5. Predicción de la temperatura por ciudades

Al igual que la sección anterior, los datos de temperatura promedio mensual, latitud y longitud utilizados en esta sección provienen de *Kaggle* donde se encuentran los mismos de aproximadamente 3500 ciudades, mientras que para los datos de alturas de ciudades se utilizaron los de *World City Locations Da-*

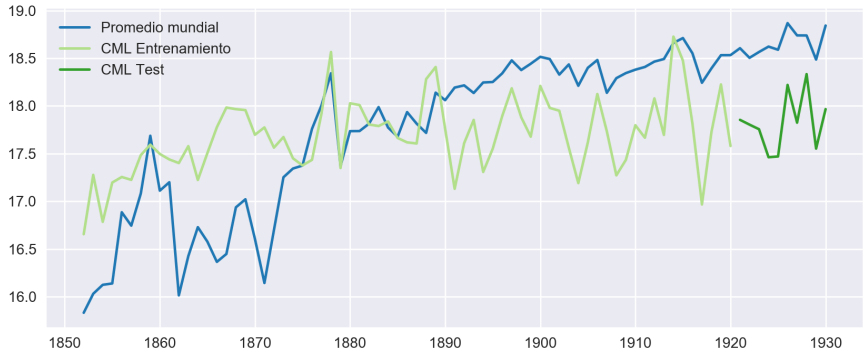

Figura 4. La función $g(X^{(i)})$ entrenada hasta 1920 y proyectada hasta 1930.

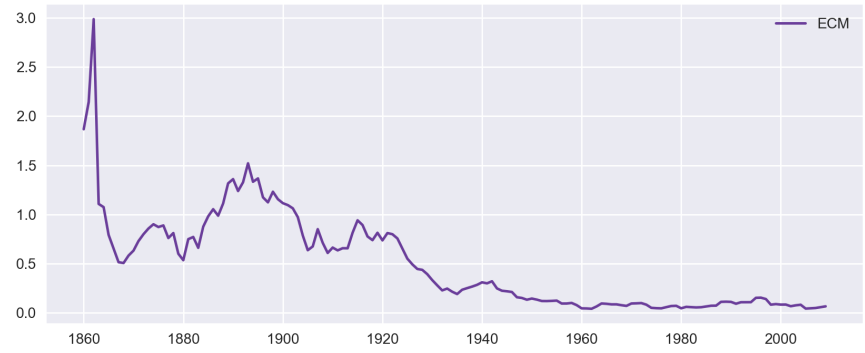

Figura 5. ECM_g de los 10 años siguientes al año del corte por año.

tabase[7]. El propósito de esta sección es poder predecir la temperatura de una ciudad utilizando datos de otras ciudades. Por ello, hemos decidido utilizar sólo ciudades de los Estados Unidos ya que la base posee una gran cantidad de ciudades con climas y geografías distintas entre sí. Nuestro experimento consiste en predecir la temperatura de una ciudad de la costa este utilizando ciudades de la costa oeste, y viceversa. Para esto definimos que una ciudad es de la costa oeste si está entre los meridianos -105 y -125, y de la costa este si está entre -65 y -85. Como ciudad del oeste elegimos El Paso y como ciudad del este a Cleveland[8].

El modelo propuesto toma las temperaturas promedio anuales de cada

[7] https://github.com/bahar/WorldCityLocations
[8] Ambas ciudades elegidas a dedo y sin tomar en cuenta nada particular de ellas.

ciudad de una costa como variables independientes, y la temperatura de la ciudad de la costa opuesta como variable dependiente. Este modelo se utiliza para predecir los 5 años posteriores a los de entrenamiento. Luego, procedemos a reducir la cantidad de ciudades por costa de acuerdo a estos criterios: Primero tomando ciudades a menos de 100 metros sobre el nivel del mar (donde obtenemos 15 ciudades por conjunto) y por último, tomando las 5 ciudades con mayor amplitud térmica (también del conjunto total). Los análisis fueron realizados de la misma forma que en las secciones anteriores, con el rango de años de 1900 a 2010. La validación cruzada empieza en 1920 y termina en 2005, de forma tal de tener al menos 20 años de entrenamiento y poder predecir 5 años en el último corte.

objetivo	subconjunto	ECM	error mínimo	error máximo
El Paso	costa este	0,436834	0,073482	2,321951
	altitud $<$ 100m	0,308179	0,034849	0,800785
	top amplitud	0,273082	0,033659	0,723342
Cleveland	costa oeste	0,821950	0,029300	2,576817
	altitud $<$ 100m	0,660223	0,041248	2,266057
	top amplitud	0,614763	0,019107	2,123448

Como el ECM no presentaba tendencia a disminuir al ir aumentando las fechas del corte, decidimos comparar los histogramas de las distintas variantes y eso te puede contemplar en la figura 6.

En el gráfico 6 se puede observar claramente como ambos modelos mejoran al reducir la cantidad de variables. En especial, el que utiliza la costa este para predecir la ciudad de El Paso usando las 5 ciudades de mayor amplitud térmica.

6. Conclusión

Al trabajar con datos provenientes de un fenómeno ultra complejo como lo es la temperatura, buscar un modelo matemático robusto que los prediga resulta muy complicado, especialmente al reducirnos al método de cuadrados mínimos lineales. Es nuestra opinión que un modelo eficaz tiene que alejarse de las funciones convencionales y enfocarse en los datos.

Por otro lado, el manejo de gran cantidad de datos es susceptible a errores

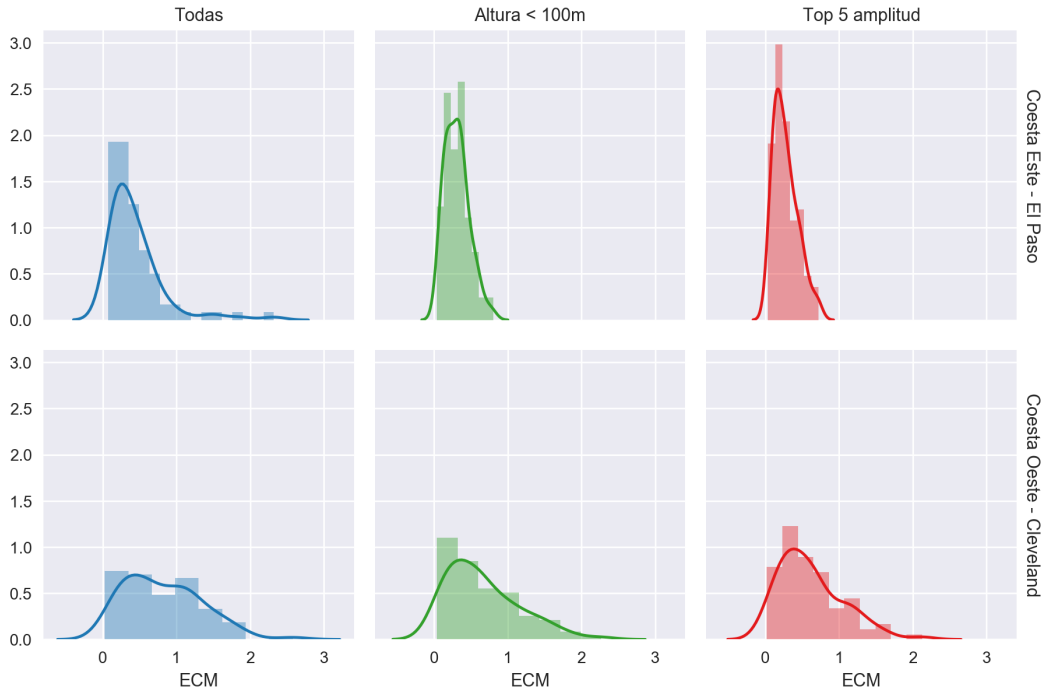

Figura 6. Histogramas del ECM por corte para las distintas variantes.

y tedioso. Si no fuese por herramientas como Jupyter y Pandas nos hubiese resultado muy difícil realizar este análisis.