# Predicción de Temperatura

Len, Nicolás [1] Len, Julian [2] Mascitti, Julio [3]

*Departamento de Computación - FCEyN*
*Universidad de Buenos Aires*
*Buenos Aires, Argentina*

**Abstract**

Estudio del comportamiento del clima mundial en términos del tiempo y de la emisión de gases. Se utilizó como herramienta el modelo de Cuadrados Mínimos Lineales para aproximar las observaciones y se logró realizar predicciones de la temperatura a futuro según distintos factores.

*Keywords:* CML, Clima, Temperatura, CO2

[1] Email:nicolaslen@gmail.com
[2] Email:julianlen@gmail.com
[3] Email:mascittija@gmail.com

# 1 Introducción

La capacidad de tener pronósticos del clima precisos es un tema que impacta fuertemente a toda la sociedad, por diversos motivos que ya conocemos, como por ejemplo el simple hecho de saber si necesitamos salir de casa con paraguas o no.

Este trabajo práctico tiene como objetivo principal, llevar a la práctica y conocer con un caso en particular, el poder que tiene la técnica de *cuadrados mínimos*. Para esto, se puso como meta intentar predecir ciertos datos del clima a partir de datos que se fueron obteniendo durante muchos años.

Cuadrados mínimos es una técnica en la que, dados un conjunto de pares ordenados (variable independiente, variable dependiente) y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de error cuadrático mínimo.

Es por eso que podemos intentar predecir la temperatura en ciertas ubicaciones geográficas, o en el planeta entero, haciendo uso de esta técnica, utilizando como variable dependiente la temperatura, y como variable independiente algún factor que consideremos que pueda tener alguna relación. Para esto, una vez decidido qué factor se considerará, volcaremos los datos obtenidos de la temperatura en función de ese patrón y pensaremos una familia de funciones que aproxime los datos.

Sabemos que el calentamiento global es un factor que hoy en día tiene mucha influencia en el clima mundial. Las variaciones de temperatura son más dispersas y los cambios de clima según la estación del año dejaron de ser tan ortodoxos como lo eran hace un tiempo. Es por eso que creemos que analizar algún factor que influya directamente en el calentamiento global, nos puede llegar a dar información interesante acerca de la variación de la temperatura durante el tiempo.

Creemos que la emisión de gases contaminantes que se genera durante todo el mundo influye directamente en la temperatura del planeta. Las potencias mundiales a nivel industrial, como Estados Unidos y China, son aquellos países que generan mayor emisión de gas, y es por eso que también creemos que en esos países la temperatura promedio de cada año crece durante el tiempo.

En síntesis, observaremos por un lado la variación de la temperatura del planeta y de algunos países (Argentina, China y Estados Unidos), en función de la emisión de gases en cada lugar respectivamente. En cada caso, planteando como hipótesis que la emisión de gases influye directamente en la temperatura de un país o del planeta. Y que lo seguirá haciendo en el futuro.

Por otro lado, observaremos la variación de la temperatura del planeta en función del tiempo, planteando como hipótesis que el promedio de temperatura fue creciendo de manera constante durante el tiempo. Y que lo seguirá haciendo en el futuro.

## 2 Desarrollo

Para cada experimento decidimos obtener predicciones a partir de distintos intervalos de datos, es decir, distintos entrenamientos, para hacer Cross Validation. De esta manera evitamos caer en Overfitting. Luego, elegimos la familia de funciones que mejor se adapte al modelo en general. Es decir, aquella que, independientemente del intervalo de entrenamiento, genere mejor predicción.

También, en algunos casos, acotamos la cantidad de datos obtenida según algún criterio que nos permita obtener únicamente el rango mas relevante (con mayor cantidad de información).

### *2.1 Variación de la temperatura del planeta en función del tiempo*

Como primer eje de estudio nos planteamos evaluar la variación de temperatura del planeta Tierra en función del tiempo, específicamente, en función de los años.

Al mismo tiempo decidimos dividir el experimento en dos partes. Tratar de predecir temperaturas con pocos (15 años de entrenamiento, 3 de predicción) y muchos datos (50 años de entrenamiento, 20 de predicción), teniendo como hipótesis, que a mayor cantidad de datos, la complejidad del modelo para realizar una predicción certera es mayor.

#### *2.1.1 Los datos*

Tomamos como set de datos las temperaturas anuales obtenidas entre los años 1880 y 2012.

#### *2.1.2 Los experimentos*

Una vez elegidos los datos con los cuales trabajar, graficamos la temperatura del planeta en función del tiempo para poder darnos alguna idea cómo evoluciona la temperatura. Se plantearon distintos entrenamientos divididos en dos tipos. Por un lado, entrenamientos donde se utilizaron los datos que se tenían dentro de distintos intervalos disjuntos de 50 años, y se realizaron predicciones sobre los siguientes 20 años. Por otro lado, entrenamientos donde se

utilizaron los datos que se tenían dentro de distintos intervalos disjuntos de 15 años, y se realizaron predicciones sobre los siguientes 3 años. En ambos casos se realizaron predicciones con distintas familias de funciones.

### *2.2 Variación de la temperatura del planeta en función de la emisión de gas*

Como segundo eje de estudio nos planteamos evaluar la variación de temperatura del planeta Tierra en función de la emisión de gas promedio de algunos países. Elegimos 4 países dentro de los cuales están China, Japón y Estados Unidos, potencias mundiales a nivel industrial y por lo tanto principales emisores de gases contaminantes. También elegimos Argentina para completar la elección de 4 países distintos.

#### *2.2.1 Los datos*

Tomamos como set de datos las temperaturas anuales y las emisiones de gases, obtenidos entre los años 1960 y 2012. Elegimos esos años dado que previo al 1960 no existe mucha información. Para determinar las emisiones de gases, se calculó un promedio de la emisión de $CO_2$ medidas en kt [4] de Argentina, China, Japón y Estados Unidos para cada año dentro del intervalo mencionado. Con respecto a la emisión de gases, se comenzó analizando la totalidad de los datos, observando así que para emisiones muy altas se tenía poca información acerca de la temperatura del planeta. Por lo tanto, acotamos nuestros datos con un rango de emisión de gases (menor a 2500000 kt) en el cual se encontraba la mayor cantidad de datos.

#### *2.2.2 Los experimentos*

Una vez elegidos los datos con los cuales trabajar, graficamos la temperatura del planeta en función de la emisión de gas para poder darnos alguna idea de cómo evoluciona la temperatura promedio. Se acotaron los datos en base a la emisión de gases (a partir de 700000 kt) dado que no se contaban con una relevante cantidad de datos fuera de este intervalo. Luego hicimos entrenamientos cada 450000 kt y predicciones de los próximos 350000 kt.

[4] Toneladas métricas per cápita. Centro de Análisis de Información sobre Dióxido de Carbono, División de Ciencias Ambientales del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos).[1]

### *2.3 Variación de la temperatura de Argentina, China y Estados Unidos en función de la emisión de gas en cada país respectivamente*

Como tercer eje de estudio, nos planteamos evaluar la variación de temperatura de los países, en función de su emisión de gases. Para esto nos quedamos con los datos obtenidos de Argentina, China y Estados Unidos, con la idea de encontrar una familia de funciones que minimice el error de predicción de los tres países simultáneamente.

#### *2.3.1 Los datos*

Al igual que en el experimento anterior, utilizamos los datos de la emisión de gas de cada uno de los países mencionados, desde el año 1960 hasta el año 2012. Y luego los cruzamos con los datos de temperatura provista por la cátedra, para poder realizar el experimento de cada país.

#### *2.3.2 Los experimentos*

Una vez elegidos los datos, graficamos la temperatura en función de la emisión de gas de cada país, definiendo distintos intervalos de entrenamiento y predicción para calcular el ECM, tomando en cuenta la intención de no caer en Overfitting, y luego calculamos el promedio del ECM de todos los países.

## 3 Resultados y conclusiones

A continuación detallamos los resultados y conclusiones de cada experimento realizado.

Vale aclarar que en cada experimento probamos distintas familias de funciones, y fuimos buscando la que mejor optimizaba el error cuadrático medio en los diferentes intervalos. Una vez que se obtuvieron las familias de funciones que menor ECM tenían, observamos en los gráficos cuáles se parecían más a los datos que figuraban en el gráfico.

### *3.1 Variación de la temperatura del planeta en función del tiempo*

En función de los datos, se pudo percibir el fenómeno de calentamiento global, con la evolución incremental de la temperatura a lo largo del tiempo.

Al mismo tiempo, pudimos verificar nuestra hipótesis. Es decir, para intervalos cortos de tiempo, utilizar modelos complejos no necesariamente aumenta la efectividad de la predicción, sino que hace todo lo contrario. Se puede ver en los gráficos que se pudimos obtener muy buenos resultados realizando aproxi-

maciones con constantes y rectas. Resultando como modelo de funciones que minimizan el ECM las rectas.

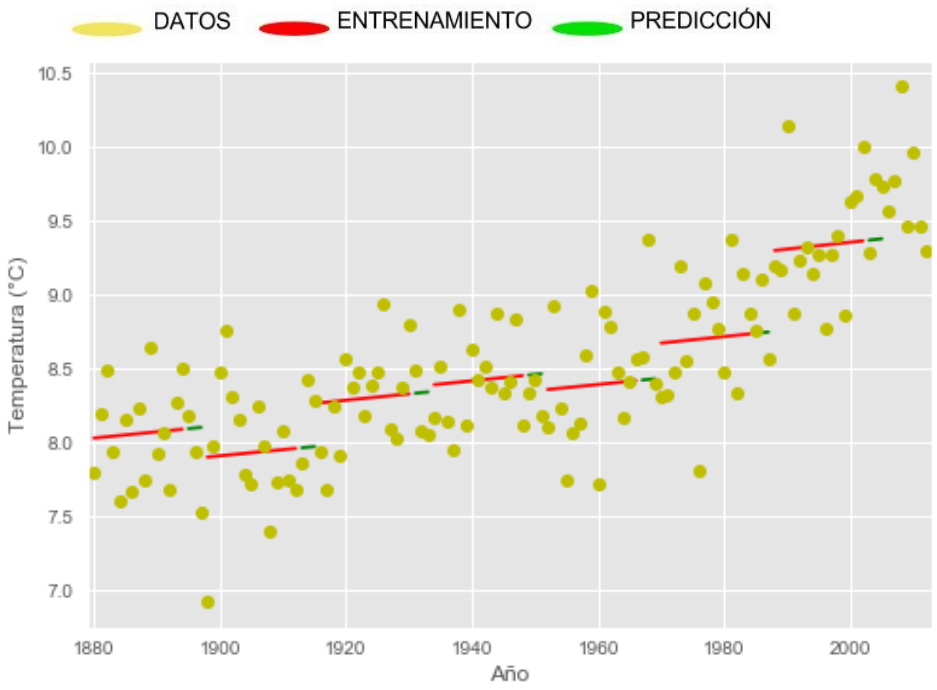


Fig. 1. Entrenamiento: 15 - Predicción: 3

| **SHORT** | |
|---|---|
| **Año** | **Error de Predicción** |
| 1880 | 0.12 |
| 1898 | 0.10 |
| 1916 | 0.06 |
| 1934 | 0.03 |
| 1952 | 0.31 |
| 1970 | 0.05 |
| 1988 | 0.10 |
| Promedio: 0.112 | |

Para largos períodos de tiempo, pudimos ver como a medida que fuimos agregando funciones que intenten captar la evolución de los datos a lo largo del tiempo, la precisión de las predicciones iban aumentando progresivamente. Por ejemplo, agregar rectas nos sirvió para intentar imitar la tendencia alcista de los datos. Luego con polinomios de grado 2 pudimos imitar más fielmente el incremento paulatino, o por ejemplo utilizando funciones trigonométricas (como el seno) conseguimos ajustar un poco mejor la pequeña dispersión de los datos año a año. La familia de funciones que minimiza el ECM en este caso es una composición de constantes, rectas y seno, es decir: $a*1+b*t+c*sen(t^5)$.

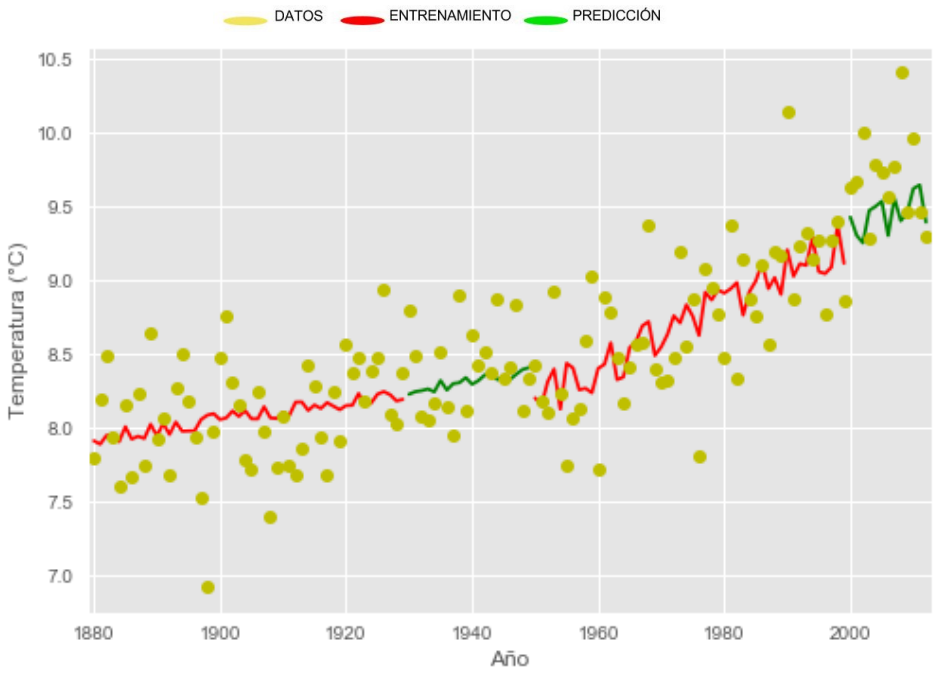


Fig. 2. Entrenamiento: 50 - Predicción: 20

| **LONG** | |
|---|---|
| **Año** | **Error de Predicción** |
| 1880 | 0.09 |
| 1898 | 0.17 |
| Promedio: 0.129 | |

### *3.2 Variación de la temperatura del planeta en función de la emisión de gas*

Para el siguiente experimento, empezamos por

Luego de probar con varias familias de funciones, se encontraron dos combinaciones de familias con un bajo error cuadrático medio. Estas fueron:

- $a * 1 + b * t + c * cos(t^5)$. Cuyo error fue 0.132.
- $a * 1 + b * (1/t)$ Cuyo error fue 0.121.

Si bien la segunda familia de funciones obtenía un menor error cuadrático medio, nos dimos cuenta que donde aparecían ciclos, ésta familia de funciones no se adecuaba a la información. Es por eso que llegamos a la conclusión que quedarnos con la primera familia de funciones era la mejor decisión.

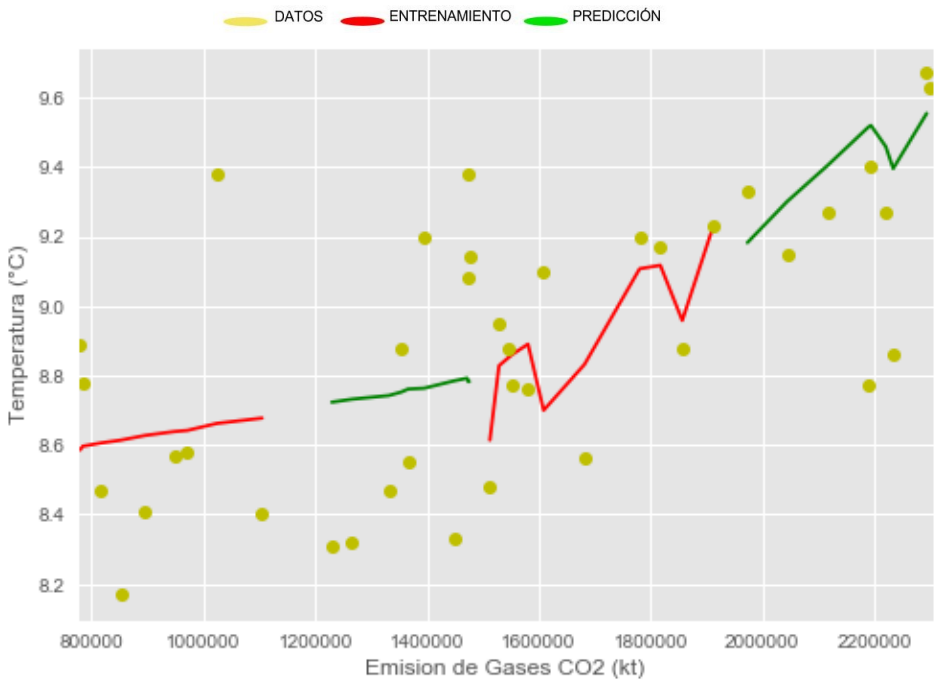


Fig. 3. Gases del mundo

| **Emisión** | **Error de predicción** |
|---|---|
| 700000 | 0.14 |
| 1500000 | 0.12 |
| Promedio: 0.132 | |

### *3.3 Variación de la temperatura de Argentina, China y Estados Unidos en función de la emisión de gas en cada país respectivamente*

Para este experimento fuimos probando distintas funciones incrementalmente. Primero, una única función, luego la combinación de dos funciones y por último tres funciones distintas.

Un resultado importante fue, que la combinación de dos funciones que por separado tenían buenos resultados, al ser combinadas, no garantizaban la reducción del ECM. Por ejemplo, conseguimos el mejor ECM con $a * 1 + b * 1/t^2$, cuando $b/t^2$ tenía un importante ECM al evaluarla individualmente, en contraposición a funciones como $log(t)$ o $sin(t)$ que tenían un ECM bastante reducido.

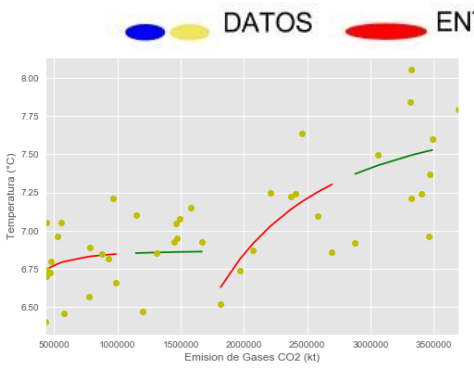


Fig. 4. China

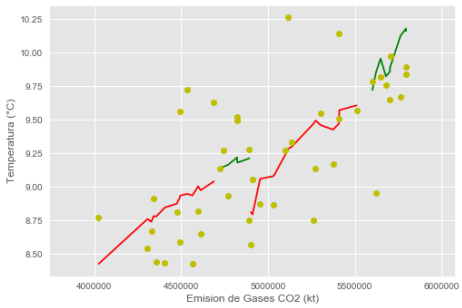


Fig. 5. Estados Unidos

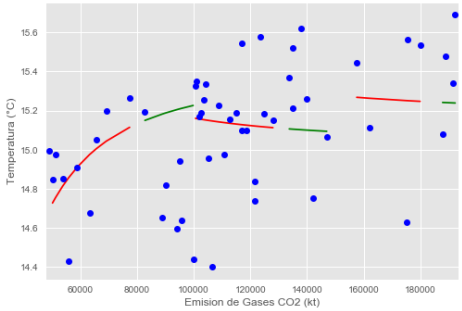


Fig. 6. Argentina

| China | | EEUU | | Argentina | |
|---|---|---|---|---|---|
| **Emisión** | 0.**Error** | **Emisión** | **Error** | **Emisión** | **Error** |
| 50000 | 0.04 | 4000000 | 0.08 | 50000 | 0.26 |
| 1800000 | 0.13 | 4900000 | 0.09 | 100000 | 0.10 |
| | | | | 150000 | 0.07 |
| Promedio: 0.085 | | Promedio: 0.084 | | Promedio: 0.143 | |
| Promedio Total: 0.255 | | | | | |

Luego de realizar los experimentos, no pudimos encontrar un sustento teórico al por qué de las familias que representaban los datos. Tuvimos buenos resultados con $a*1+b*log(t)+c*sin(t)$, $a*t+b*log(t)$, $a*1+b*log(t)$ y varias combinaciones más, pero la que minimizó el ECM fue $a*1+b*1/t^2$.

En estos experimentos fue complicado encontrar alguna tendencia en la distribución de los datos. Ya que los países tienen una emisión de gases bastante diferente, y sin ninguna correlación entre ellos. Al mismo tiempo, sus temperaturas también difieren bastante dado las ubicaciones geográficas.

## 4 Referencias

[1] http://datos.bancomundial.org/indicador/EN.ATM.CO2E.PC