

---

## Actividad 6 - Pronostico de Series de tiempo

---



Gustavo Alberto Medina Ferrer<sup>1</sup>

A219223438<sup>1</sup>

Numero(s):(81) 8309 6131<sup>1</sup>

gustavomedinaferrer@gmail.com<sup>1</sup>

---

Universidad de Sonora

Licenciatura en Fisica

Introduccion a la Fisica Moderna

21/02/2021

### Resumen

En este documento se hara un reporte de la actividad 6 de la materia de Fisica Computacional, impartida por el profesor C. Lizarraga en la Universidad de Sonora. La actividad se centra en analizar series de tiempo (Temperatura Maxima y minima de la estacion 26068) con el modelo ARIMA.

Para hacer el pronostico de las series de tiempo, primero se deberan de limpiar, esta es la primera parte de la Actividad, donde se veran dos diferentes formas de transformar una serie en una serie estacionaria. La segunda parte se centra en el analisis usando el modelo ARIMA.

## 1. INSTRUCCIONES

### 1.1. PARTE 1:

Esta primera parte se centra en "limpiar" la serie de tiempo (la cual en este caso es desde 1979 hasta 2009). Esto se hara determinando si la serie es estacionaria, es decir, si es una serie que se comporta de maneras similares a lo largo de las estaciones de los anos. Tambien se hara eliminando la *Tendencia* y la *Estacionalidad*.

La primera parte se basa en la *prueba de Dickey-Fuller*, la cual se usara para comprobar nuestra hipotesis no-nula, la cual es que la serie es *no-estacionaria*. Esto es porque este tipo de pruebas estadisticas se suelen hacer para desacreditar una hipotesis; dado que nosotros queremos analizar si es estacionaria, deberemos de tener una hipotesis opuesta a nuestras expectativas.

La prueba se basara en el valor de  $p$ , el cual es un indicador de que tan posible es que los valores analizados se den, si la hipotesis es cierta. En otras palabras, cual es la probabilidad de que la serie se comporte de la manera que se comporte si en realidad no fuera estacionaria. Dicho esto, se puede observar que buscamos un valor de  $p$  muy bajo, preferiblemente menor igual a 0.05, ya que esto nos diria que es casi imposible que la serie se comporte asi si no es estacionaria, confirmandonos que es estacionaria.

Despues de esto, se eliminara la Tendencia y la Estacionalidad, lo mas que se pueda para poder obtener un promedio movible y una desviacion estandar lineales. Esto se hara empleando dos metodos diferentes, quedandonos con el metodo que parezca darnos mejores resultados.

### 1.2. PARTE 2:

En esta segunda parte se hara el pronostico de la serie de tiempo usando los metodos de modelacion estadistica ARIMA (Auto Regressive Integrated Moving Average). Este metodo es similar a una ecuacion de regresion lineal, donde la prediccion depende de tres parametros  $p$ ,  $d$  y  $q$ , los cuales se encuentran con la *Funcion de autocorrelacion* ( $p$ ) y la *Funcion de autocorrelacion parcial* ( $q$ ).

Estas funciones se grafican y donde se cruzan con la linea de  $+1.96 \sigma$ , la desviacion estandar, es el valor de los valores  $p$  y  $q$ .

Se estudiaran cada caso: Autoregresion (AR) y promedios moviles (MA), despues integrandolos en el metodo de ARIMA, al igual que se utilizara el criterio de *Akaike* como indicador de cual es el mejor modelo, a partir del cual se construira el pronostico de la serie de tiempo.

## 2. ACTIVIDAD

### 2.1. PARTE 1:

En esta primera parte de la actividad fui donde mejor entendi los conceptos.

Primero que nada, agarre un periodo de tiempo bastante grande para hacer mis analisis. El periodo fue de 1979 a 2009 y, sorprendentemente, usando la funcion de `isnull().sum()`, vi que no faltaba ningun dato en TMax y TMin. Solo habian 8 datos faltantes en EVAP. Al principio crei que realmente era asi, sin embargo, descubri que hay un vacio de alrededor de un mes a finales del 2008, y de hecho no se tiene valores del 2009. No se si hice algo mal y realmente solo era hasta el 2008, pero no creo que haya sido asi.

Sin embargo, cargue la serie de codigo que el profesor nos puso en el portal y analice la estacionalidad de mi serie. Para mis sorpresa de nuevo, el valor de  $p$  era mucho menor que 0.05. Esto se puede deber a que escogi un periodo de tiempo muy grande, mientras que el profesor utilizo un periodo de tiempo relativamente corto. Sin embargo, decidí seguir haciendo los tratamientos que el profesor hizo, ya que aun cuando el valor de  $p$  era muy bajo, aun se veia muy inestable en cuanto a la desviacion estandar y el promedio.

Lo siguiente que se hizo fue reducir su escala a una logaritmica, lo que nos dio un promedio aun bastante inestable, pero al hacer el test de estacionaridad, se puede observar que el promedio es bastante lineal, pero la desviacion aun era bastante inestable.

Se prosiguió a usar promedios moviles exponenciales. Ahora, aqui estuve un tiempo ya que quise entender un poco mas a que se referia con exponenciales, y a lo que encontre parece ser que le da una mayor importancia a los datos mas recientes, o sea que es un tipo de promedio ponderado. Este ultimo metodo disminuyo un poco mas la inestabilidad de la desviacion estandar, sin embargo aun se podia observar la estacionaridad de la serie.

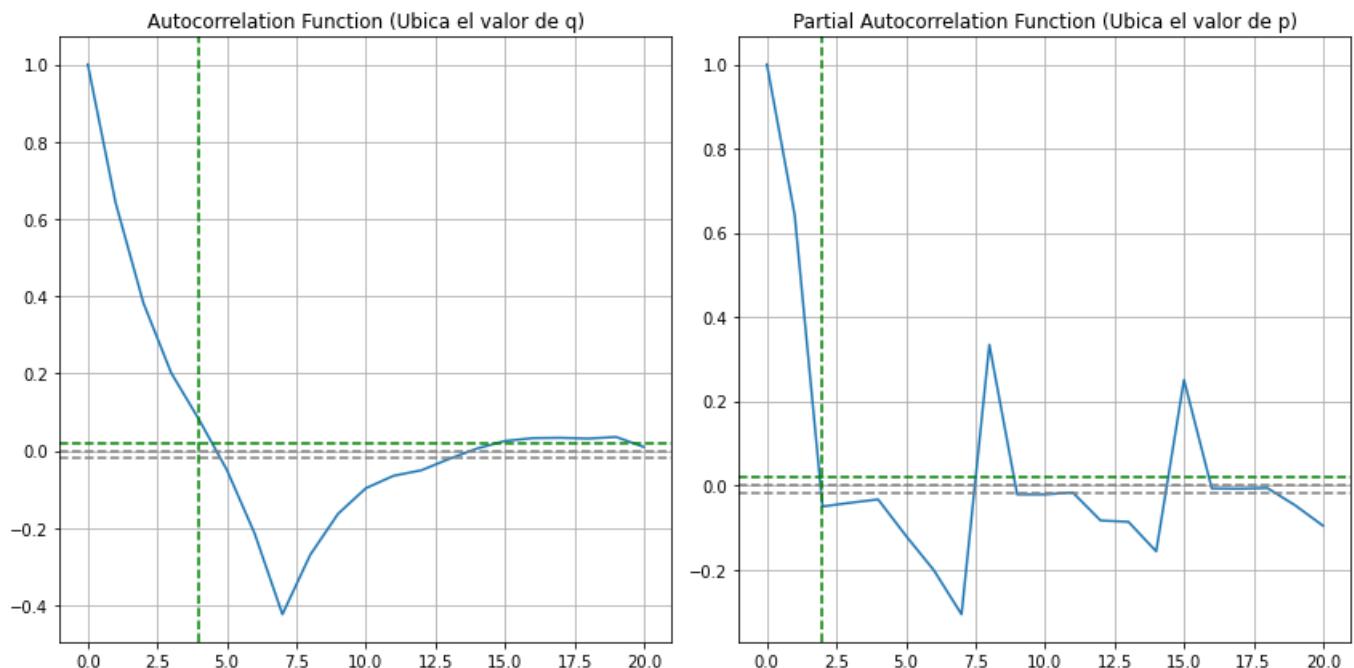
Despues de esto comence a jugar con el 'halflife' del comando, para ver exactamente en que afectaba, y conclui que entre mas corto lo hacia, mas lineal quedaban el promedio y la desviacion.

Para eliminar la estacionaridad de los datos, se uso la funcion `.shift()` de pandas y haciendo una diferencia del nuevo df que se crea con la funcion (que desfaza los datos por el numero de datos que le pongas) y el original. Con esto ultimo se pudo observar un promedio practicamente lineal, mientras que la desviacion aun tenia sus variaciones notables.

Para terminar esta primera parte se intento llegar al mismo resultado usando el metodo de descomposicion que se uso en la actividad pasada, sin embargo, esto termino siendo menos efectivo, ya que tanto el promedio como la desviacion eran menos lineales.

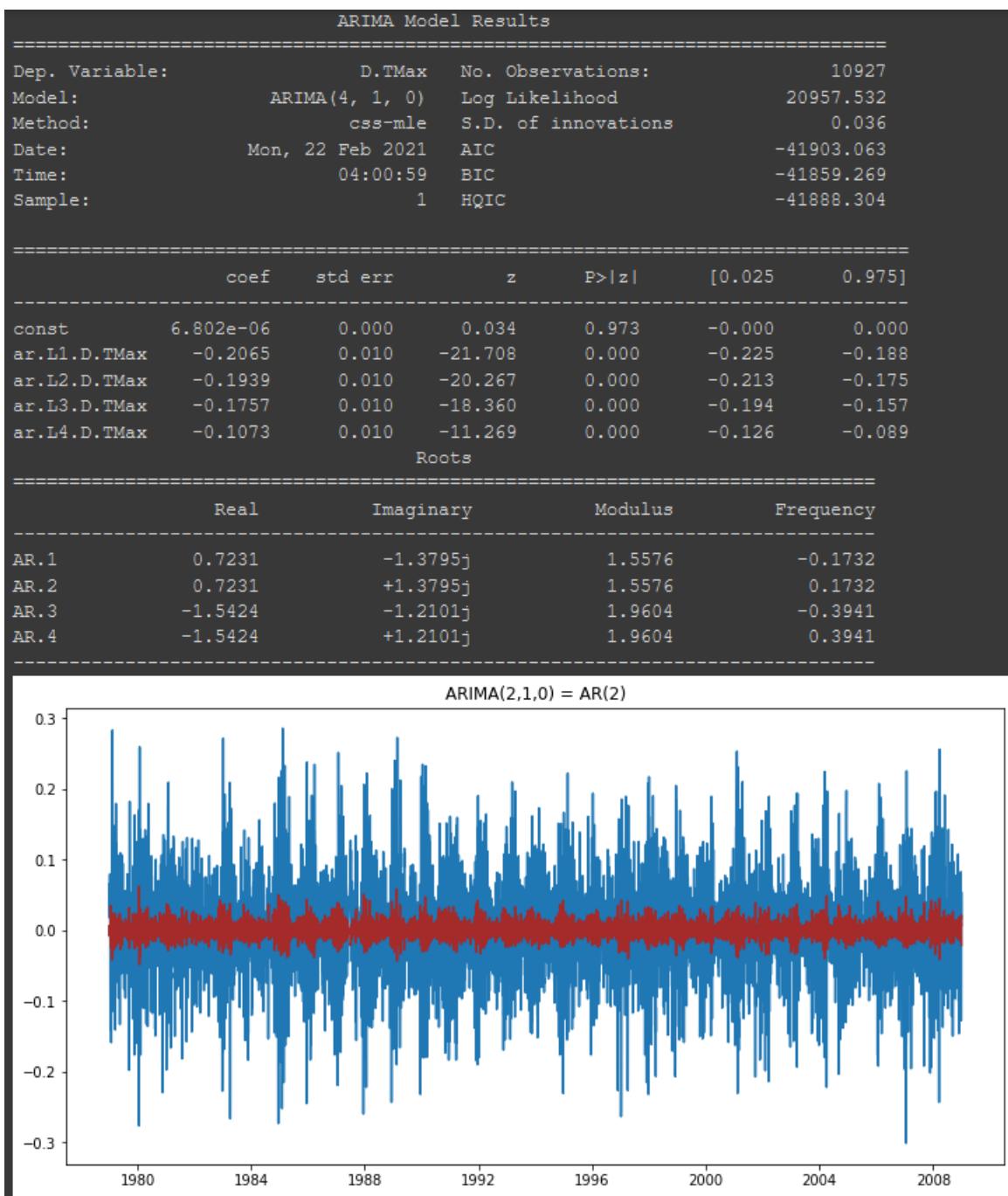
## 2.2. PARTE 2:

PAra comenzar la segunda parte, donde se utilizaran los metodos de ARIMA, se aproximan los valores de  $p$  y  $q$ , usando el metodo que se describio en las instrucciones. Las graficas que se generaron son las siguientes:

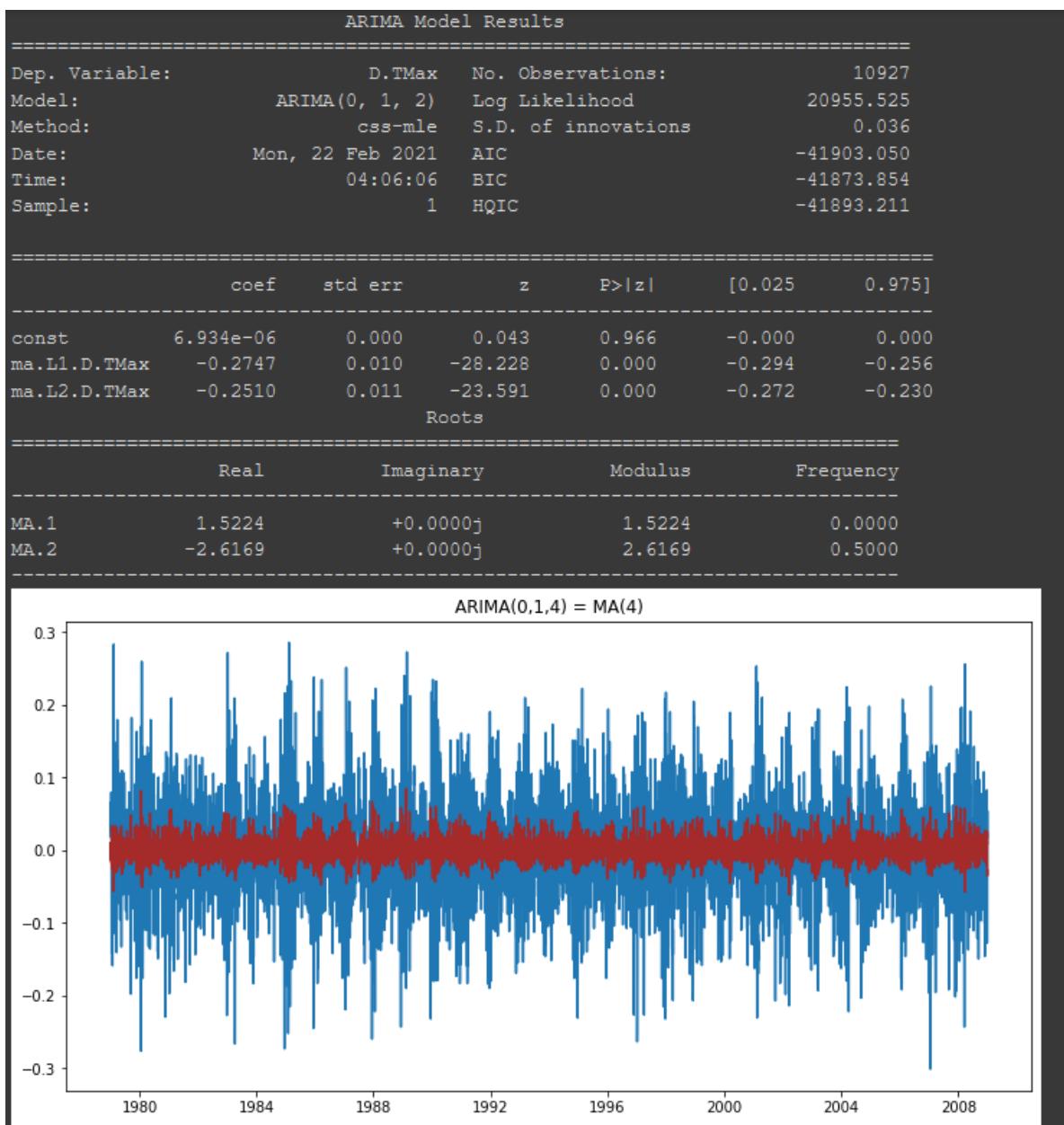


Aqui se puede observar que el valor de  $q$  es alrededor de 4, entonces  $q = 4$ , mientras que el valor de  $p$  es alrededor de 2, por lo que  $p = 2$ . Con estos dos valores, y usando el valor de  $d = 1$ , podemos hacer nuestros tres modelos, los cuales resultaron de la siguiente manera:

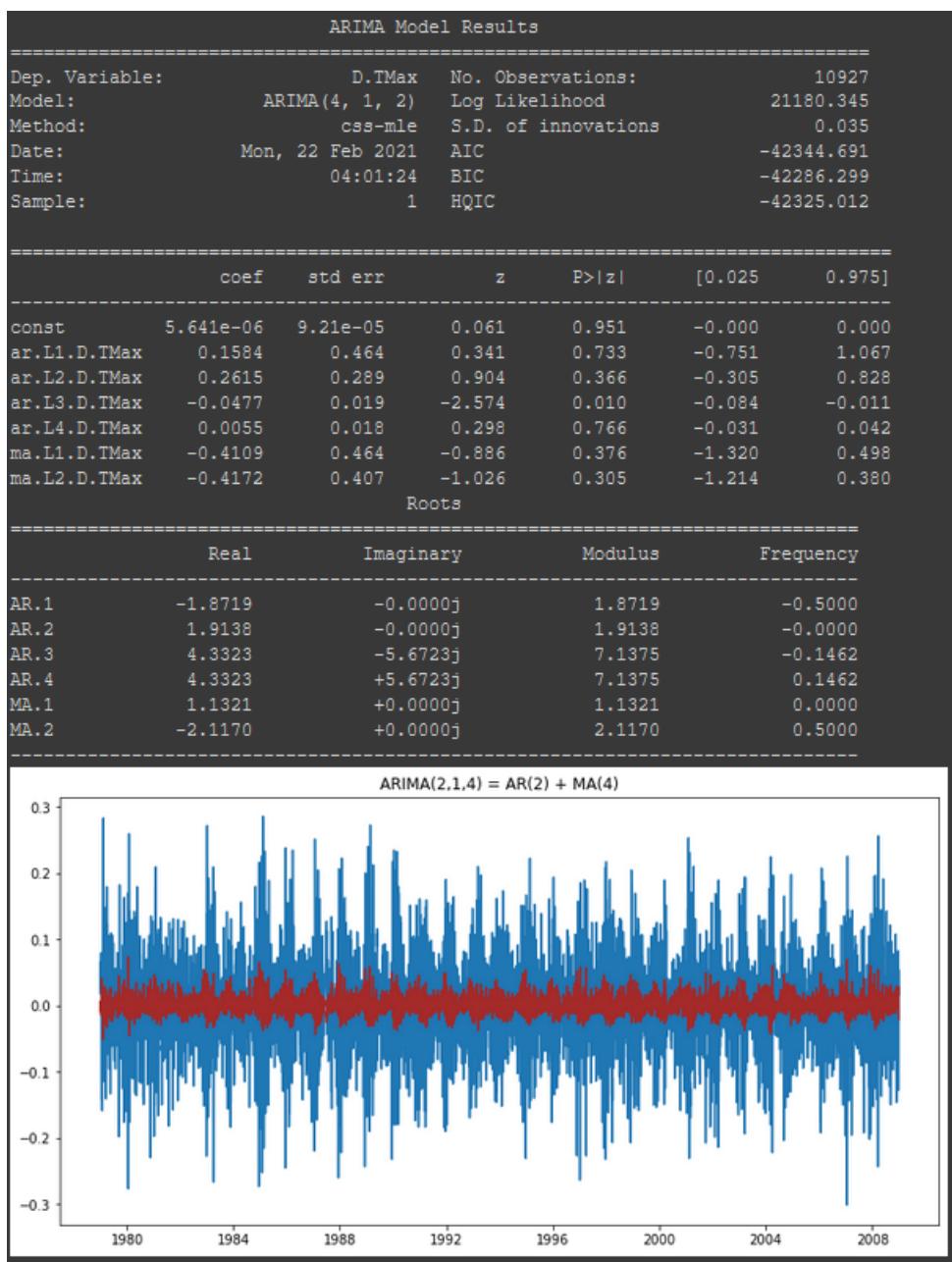
## MODELO 1: AUTO REGRESION (AR)



## MODELO 2: PROMEDIO MOVIBLE (MA)

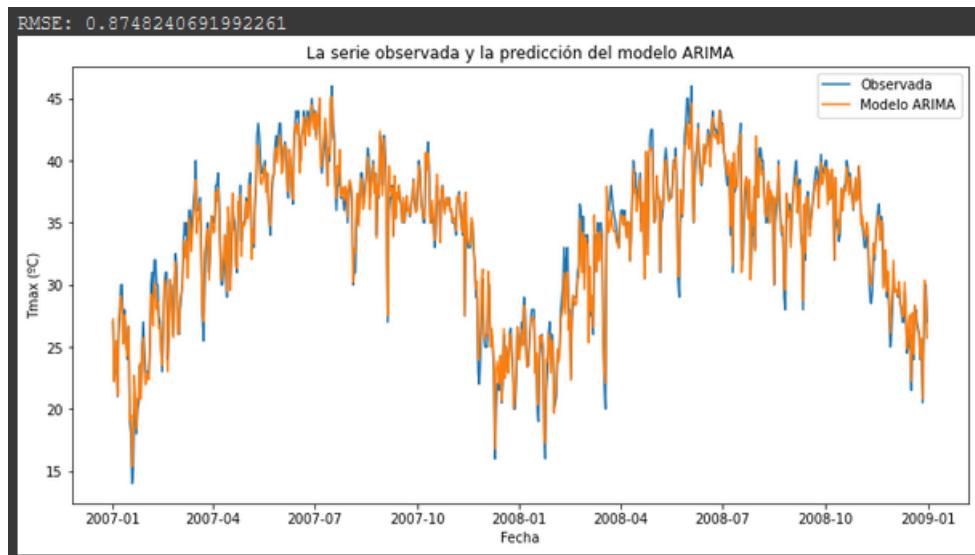


### **MODELO 3: ARIMA (MA + AR)**



Como se puede observar, en cuanto al criterio de AIC, el mejor modelo es el ultimo, el de ARIMA, que es con el cual nos quedaremos.

Como ultimo paso, se regreso la escala del df a la original y se grafico un periodo de tiempo mas corto, de 2007 a 2009, para observar mejor que tan acertado es el modelo, que como se puede ver por el siguiente grafico, es bastante, o al menos eso me parece a mi.



### **3. CONCLUSION**

En general, esta es una actividad que me parece muy interesante, ya que es una en la que comenzamos a usar conceptos y metodos estadisticos realmente avanzados, o al menos asi lo pienso. Sin embargo, y esto se puede observar en la ultima parte del codigo, me costo mucho entender que era lo que estaba pasando, ya que aun investigando los conceptos, no podia salir de mi confucion. Sin embargo, la primera parte de la actividad pienso que es la que mejor entendi, mientras que la segunda parte no la entendi mucho, pero pienso investigar mas al respecto, ya que quisiera entenderlo para poder hacer una investigacion personal que tengo en mente.

En cuanto a la tendencia de las temperaturas, se puede observar que estas han ido aumentando en los ultimos anos. Se puede observar un aumento bastante obvio en la grafica de tendencia.

Por otro lado, mientras que el modelo ARIMA aproxima bastante bien los valores de las temperaturas que se han encontrado, no se puede negar que aun hay unas partes donde se nota que falla, y esto aumenta increiblemente cuando se quiere ir hacia atras del inicio de los datos o mas alla del final de los mismos, donde comienza a fallar despues de no mucho mas de 7 dias, lo cual nos da una idea de que tan poco conocemos acerca de la prediccion del clima, uno de los misterios y problemas modernos.

Si quisiera mejorar algo acerca de la actividad seria el poder calcular exactamente el valor de  $p$  y  $q$ , ya que esta vez lo que se hizo fue aproximarlos mediante nuestro ojo, lo cual pienso que se puede mejorar, aunque honestamente no se como se hace.