**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE**

**Departamento de Matemáticas y Física**

**Sustentabilidad y tecnología**

**PROYECTO DE APLICACIÓN PROFESIONAL (PAP):**

**Programa de modelación matemática para el desarrollo de planes y proyectos de negocio**

****

**4J09   
Ciencia de datos aplicado a información de negocios y desarrollo social**

**Pronóstico general del sistema eléctrico (GCROC, Gerencia de Control Regional de Occidental)**

**PRESENTA**

**Ing. Financiera. Diana Laura Ramírez Hinojosa**

**Profesor PAP: Pablo Benavides Herrera**

**Tlaquepaque, Jalisco, noviembre 2019.**

Tabla de contenido

[1. Resumen 3](#_Toc23416703)

[2. Introducción 3](#_Toc23416704)

[3. Identificación de la problemática y reconocimiento del contexto 3](#_Toc23416705)

[3.1 CENACE 3](#_Toc23416706)

[3.2 Problema 4](#_Toc23416707)

[3.3 Contexto 4](#_Toc23416708)

[4. Planeación de una alternativa 5](#_Toc23416709)

[4.1 Objetivos 5](#_Toc23416710)

[4.2 Metodología 5](#_Toc23416711)

[4.3 Producto 5](#_Toc23416712)

[5. Validación de la alternativa 5](#_Toc23416713)

[5.1 Conceptos teóricos 5](#_Toc23416714)

[6. Implementación de la propuesta/alternativa 7](#_Toc23416715)

[7. Valoración de productos, resultados e impactos 10](#_Toc23416716)

[Competencias y aprendizajes relevantes 11](#_Toc23416717)

[8. Bibliografía y otros recursos 12](#_Toc23416718)

[9. Anexos Generales 13](#_Toc23416719)

[9.1 Bitácora semanal 13](#_Toc23416720)

# 1. Resumen

La problemática del CENACE es hacer un pronóstico con el menor error posible de la demanda de energía en la zona occidental del país. Este proyecto logró realizar un pronóstico a 5 días con .70% de MAPE. El alcance es hacia días no festivos, que no se consideran atípicos para el modelo. Sin embargo, los días festivos no son agarrados por el modelo, por lo que se buscará el siguiente semestre, trabajar en los análisis por separado de estos días.

Se espera que, con este proyecto, el CENACE se pueda enfocar en pronosticar los días que ellos consideran atípicos, teniendo ya un modelo que pronostique el resto de los días.

# 2. Introducción

El reporte que se presenta a continuación expone lo realizado en el PAP de ciencia de datos durante el semestre de otoño 2019. Se pretende explicar primero que nada el contexto en donde se desarrolla, la problemática a solucionar, los objetivos a alcanzar, el producto esperado, la teoría necesaria para entender el proyecto, para terminar con el proceso que se desarrolló a lo largo del semestre y los aprendizajes obtenidos. Se tuvo una bitácora, que será agregada como anexo, que da cuenta del trabajo realizado semana a semana.

La metodología seguida durante el semestre fue a base de prueba y error, al ir agregando cosas al modelo que lo pudieran mejorar, sin embargo, la idea central fue tener un modelo que sea explicable y que todo tuviera un fundamento, evitando el concepto de “Blackbox” en el proyecto.

# 3. Identificación de la problemática y reconocimiento del contexto

## 3.1 CENACE

El Centro Nacional de Control de Energía es un organismo público descentralizado cuyo objeto es ejercer el Control Operativo del Sistema Eléctrico Nacional (CENACE, 2019). La misión fundamental es “Sustentar eficientemente las transacciones de los productos de la electricidad entre generadores y consumidores” (CENACE, 2019).

Imagen 1. Gerencias del CENACE. Obtenida de: <https://www.cenace.gob.mx/Paginas/Publicas/Info/DemandaRegional.aspx>

Este organismo se divide en 9 gerencias, y con la que se está trabajando es la de Occidente. La gerencia de Occidente se conforma por 9 estados de la República que son: Jalisco, Nayarit, Zacatecas, Aguascalientes, Colima, Guanajuato, Michoacán, Hidalgo y San Luis Potosí.

El departamento de Evaluación y Estadística se dedica a hacer pronósticos de la demanda de energía, lo cual sirve para abastecer de manera correcta a todas las localidades dentro de los estados antes mencionados. El error del pronóstico que ellos manejan a nivel gerencial es entre 1% y 3% en MAPE (más adelante se detallará acerca de esta medida).

## 3.2 Problema

Uno de los problemas que la empresa con la que el PAP trabaja, el CENACE, presentó al iniciar el semestre, fue el de pronosticar la demanda de energía a nivel gerencial para reducir el MAPE que ellos manejan.

¿Por qué se busca constantemente reducir el error? Para la empresa es vital, ya que un 1% de error les genera a ellos 60 millones de pesos en pérdidas. Tal y como la persona con la que se tiene el contacto de la empresa, Camilo, explicó en la reunión, el objetivo no es necesariamente cambiar el modelo, si no mejorarlo con las innovaciones que los estudiantes presenten.

## 3.3 Contexto

La magnitud de este problema radica a un nivel nacional, se está hablando de 9 estados de la República, en donde un mal pronóstico crearía dos de las opciones siguientes:

* Pronosticar una demanda menor a la real, llevando a una falta de energía en diferentes zonas, energía que ya como mexicanos, muchas veces damos por sentado.
* Pronosticar una demanda de energía mayor a la real, dando como resultado una mayor producción de la necesaria, es decir, pérdidas para la empresa.

Ambas situaciones son malas opciones para el CENACE, y aunque el error actual es muy bajo, siempre hay espacio para reducirlo, y mejorar las condiciones actuales del pronóstico.

En retrospectiva, todos los habitantes de los nueve estados antes mencionados son los beneficiarios de la resolución del problema, y los afectados en caso de que algo fallara. El trabajo del CENACE es fundamental para un bien tan necesario como la energía.

# 4. Planeación de una alternativa

## 4.1 Objetivos

* Analizar la serie de tiempo de la demanda de energía.
* Crear un modelo para pronosticar demanda de energía.
* Reducir el error de los pronósticos del modelo creado.

## 4.2 Metodología

Los pasos que se esperan seguir son: Primero analizar la serie de tiempo, entender qué es lo que pasa dentro de ella. Después, buscar las variables que deben entrar en el modelo de regresión. Esto involucra las conocidas variables dummies, y otras variables que vayan surgiendo conforme al análisis que resulten de interés, por ejemplo, la frecuencia. Al poder establecer esto, se buscará crear el modelo de regresión, y a partir de ese momento, se buscarán errores y posibles mejoras que bajen el MAPE de los datos. Es un proceso que involucra una dedicación constante, ya que la serie de tiempo debe ser analizada desde todos los puntos de vista posible, en estos casos, de los factores más importantes a estudiar son los días festivos, y cómo afectan a la misma serie.

## 4.3 Producto

El producto que se espera tener es un programa, amigable para el usuario, que permita dar como resultado el pronóstico de los siguientes días de la demanda de energía. Se espera se pueda predecir, con el menor error, de 4 a 14 días, que es lo que la empresa espera.

# 5. Validación de la alternativa

## 5.1 Conceptos teóricos

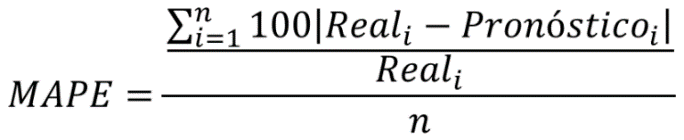
**MAPE.** Sus siglas vienen de error porcentual medio absoluto en inglés, y lo que da es la desviación en términos porcentuales. Esta medida que se obtiene como se ve en la Imagen 2, es lo que el CENACE pide para poder validar el modelo que se entrega. El MAPE se puede entender de dos maneras: Sobre la regresión y sobre el pronóstico. El MAPE de la regresión se refiere al error del modelo que creó a partir de los datos, al comparar los datos de la regresión con los reales. Ya con este modelo, se predicen la demanda de energía futura, y es cuando entra el MAPE sobre pronóstico, es decir, sobre datos que la regresión no tiene.

Imagen 2: Medición de error de pronóstico. Obtenido de: <https://ingenioempresa.com/medicion-error-pronostico/>

Todos los cambios que durante en el semestre se fueron agregando o quitando, fueron en base al error obtenido en el MAPE, ya que, si un cambio lo aumentaba, no constituía una mejora. De la misma manera, si el MAPE reducía al agregar un paso en el modelo, se considera una mejora.

**Variables ficticias**. Conocidas como dummies, se utilizan para explicar variables cuantitativas, ya que las regresiones aceptan solamente números, es decir, variables cuantitativas. Sirven para que el modelo distinga una característica que puede tener dos o más opciones, por ejemplo, el sexo de una persona. ¿Cómo se utiliza? Con números binarios. Tan simple como agregar una variable que contenga un 1 si es mujer y un 0 si es hombre, o viceversa. En el caso de la demanda de energía, había factores a considerar como los días de la semana, los meses del año, y los días festivos. Esto llevaba a crear muchas variables para que el programa distinguiera estos parámetros, quedando de la siguiente manera:

* Días de la semana: Seis variables dummies
* Mes del año: Once variables dummies
* Días festivos: Diecisiete variables dummies
  + Navidad
  + Año nuevo
  + Jueves santo
  + Viernes santo
  + Sábado santo
  + Domingo santo
  + Independencia de México
  + Asueto por día de la constitución
  + Día del trabajo
  + Día de la madre
  + Día de la virgen de Zapopan
  + Día de la virgen de Guadalupe
  + Día de muertos
  + Día de elecciones
  + Toma de poder presidencial
  + Asueto de la revolución mexicana
  + Asueto natalicio Benito Juárez

**Serie de Fourier**: El teorema de Fourier establece lo siguiente:

*Cualquier función periódica, con periodo T, se puede representar como suma de sinusoides de frecuencias f, 2f, 3f, .., llamadas armónicos. (La relación entre el periodo y la frecuencia es f=1/T*) (Barrière, 2011).

En pocas palabras, la serie de Fourier busca descomponer una función en sumas de senos y cosenos, con frecuencias de las señales de la función inicial. En el modelo que se buscó desarrollar, una parte fundamental fue el análisis de las frecuencias para poder determinar los armónicos a usar en l descomposición de la demanda de energía. Es importante recalcar que este teorema se puede utilizar por la forma de la serie, ya que es periódica.

**Análisis de frecuencia:** Como ya se mencionó, es el paso anterior para usar el teorema de Fourier. Este análisis da como resultado cada cuántos datos hay un componente desconocido que se repite, cuantas veces sea necesario. ¿Qué significa esto? Al tener una serie diaria con frecuencia 7, habla de que cada 7 días hay algo parecido en los datos, y esto se repite en toda la serie. Si una de sus frecuencias es 30, cada mes hay algo parecido, y así sucesivamente.

**Producto Kronecker**: Es una operación sobre dos matrices en donde se multiplica cada término de una matriz por la segunda matriz. Para este proyecto, se utilizó una versión simplificada en donde por filas, cada término de una matriz (dummies) se multiplicaba por la matriz de Fourier. De esta manera, se buscó captar el efecto de las dummies en conjunto con la serie de Fourier que fue creada.

**AR, ARIMA, SARIMA:** Se refieren a modelos con el enfoque de análisis en series de tiempo, y que han sido usadas para la determinación de proyecciones de variables que son susceptibles a cambios según el tiempo, y también buscan detectar los efectos de dichos cambios (Chávez, 1997).

# 6. Implementación de la propuesta/alternativa

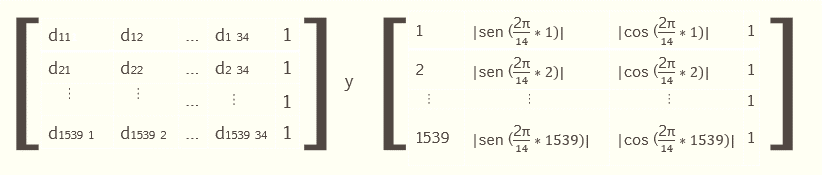
Los pasos que el programa realiza son los siguientes:

1. Importar librerías necesarias y la serie de datos, al igual que la tabla creada de los días festivos que se mencionaron anteriormente
2. Crear las variables dummies para meses, días de la semana, y los días festivos.
3. Observar descomposición de los datos. Entender que cada 7 días es el componente repetitivo más importante.

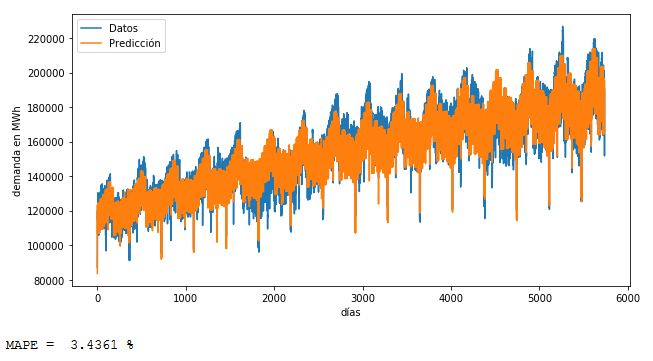
Imagen que contiene parado, tabla, viendo, agua

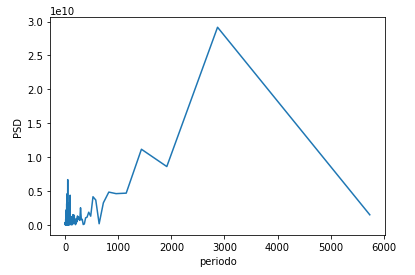
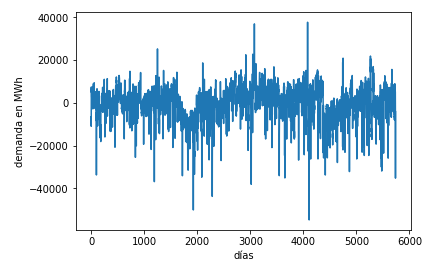
Descripción generada automáticamente

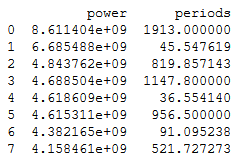
1. Tener dos tablas, una de las variables dummies y el componente repetitivo, otra de una serie de los días: 1,2,3,…1539.



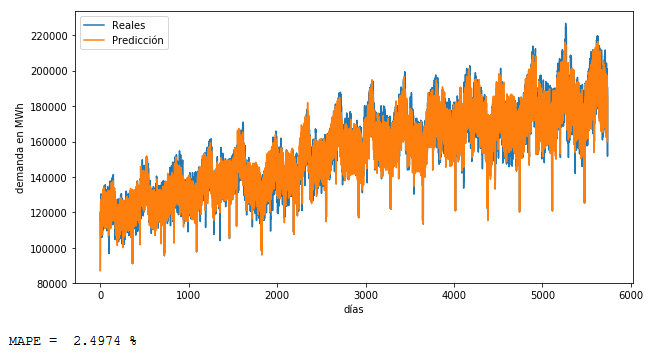
1. Elaborar producto kronecker por líneas de ambas tablas
2. Hacer una regresión lineal del producto kronecker para la serie de datos original.



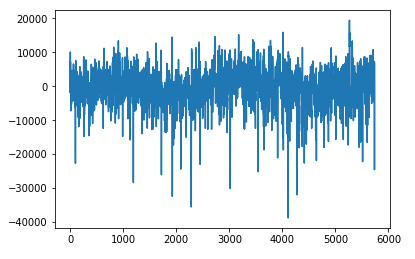
1. Graficar los residuales de la regresión, y hacerles un análisis de frecuencias
2. Elaborar según el teorema de Fourier, una variable de seno y coseno por cada frecuencia relevante



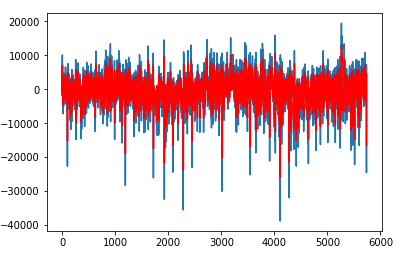
1. Hacer el producto kronecker de las variables dummies y las variables senos y cosenos
2. Hacer una regresión de los datos con el producto kronecker obtenido



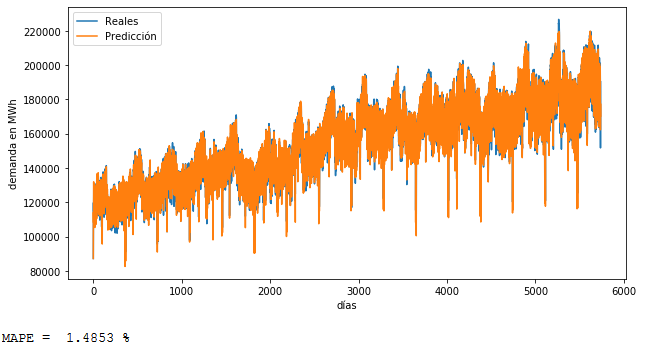
1. Volver a graficar residuales



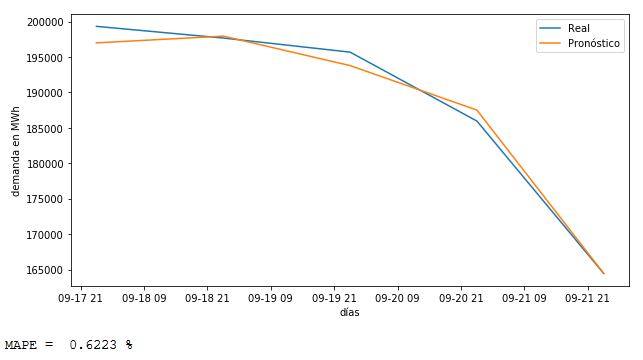
1. Agregar un ARIMA(1,0,4) a los residuales

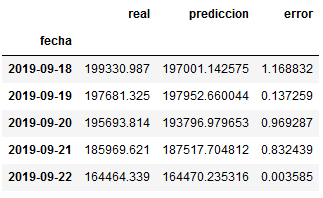


1. Hacer una regresión, sumando los datos ajustados del ARIMA



1. Hacer una regresión sin el ARIMA con parametrización Lasso, para después hacer el pronóstico con los últimos n días y sumarle la predicción de los residuales ARIMA, comparándolos con los datos reales, para obtener el MAPE.





# 7. Valoración de productos, resultados e impactos

## 7.1 Mejoras y aspectos para el siguiente PAP

Como se mencionó al inicio de este reporte, se espera que en el siguiente PAP, se trabaje en un modelo para los días que esta regresión no alcanza a cubrir, los días considerados como “atípicos” para el modelo.

Gracias a lo obtenido durante el semestre, se espera que por el momento, el CENACE se pueda enfocar en predecir los días “atípicos” para el modelo, teniendo en cuenta que ya hay un modelo para los días explicables.

## 7.2 Competencias y aprendizajes relevantes

Este PAP me ha hecho crecer mucho, y puso en duda muchas veces mi conocimiento teórico y práctico. Aprendí básicamente lo indispensable para analizar una serie de tiempo, y a pesar de que constituyó a veces un reto, logré sacarlo adelante, y me siento orgullosa de lo que he aprendido y lo que he logrado.

En mis aprendizajes profesionales están las series de Fourier, el análisis de frecuencia, el análisis de residuales, de manera no sólo teórica, sino todo lo que implicaba hacer un programa lo suficientemente automatizado para hacer los cálculos necesarios. Entendí que me hace falta mucha práctica para crecer profesionalmente, pero considero que este PAP hizo crecer mis conocimientos de manera exponencial.

En el ámbito social, me ayudó a concientizar algo que creo que la mayoría de mexicanos damos por sentado, la energía. Sabemos que si pagamos, tendremos energía, así de fácil, pero no vemos todo lo que hay detrás. Yo no sabía, por ejemplo, que el pronóstico era necesario, y entender las posibles consecuencias tanto para la empresa como para la sociedad, me hicieron pensar en que es mejor no decidir por una de estas dos entidades, sino mejor enfocarme en intentar ayudar a que ambas partes estén satisfechas. De alguna manera, que las personas no necesiten investigar qué hay detrás de pagar un recibo y tener energía, porque si eso pasa, es porque la demanda no está siendo atendida completamente.

Personalmente, me enfrenté a retos, y aprendí a no darme por vencida tan fácilmente, ya que buscando se encuentra. Por el otro lado aprendí que ser obstinada en algunos casos no lo vale, si no se puede de una manera, se debe encontrar otra diferente para lograr el objetivo. Me quedo con muchos aprendizajes en este PAP, creo que mi conocimiento creció, y estuvo en una etapa de mi vida en donde crecí personalmente, por lo que mejoré constantemente mi visión de lo que podía llegar a aprender a lo largo de todo el semestre.

# 8. Bibliografía y otros recursos

Quiénes Somos en “CENACE.GOB” 2019. Disponible en: <https://www.cenace.gob.mx/CENACE.aspx> [26 de octubre 2019]

Camacho, Máximo (s.f.) Variables Ficticias en “Universidad de Murcia”. Disponible en: <https://www.um.es/econometria/adeydcho/transpar/07-ficticias_Maximo.pdf> [28 de octubre 2019]

Berrière, Lali (2011). Serie de Fourier en “Universitat Politècnica de Catalunya ». Disponible en: <https://mat-web.upc.edu/people/lali.barriere/as/serie-fourier.pdf> [31 de octubre 2019]

Chávez, Nicolás (1997). Modelos ARIMA en “SciELO”. Disponible en: <http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2077-33231997000100005> [3 de noviembre 2019]

# 9. Anexos Generales

## 9.1 Bitácora semanal

**Semana 1:** Familiarización con la empresa CENACE, analizar los posibles proyectos en base a la información que se nos dio.

**Semana 2:** Análisis de los datos. Se busca adecuar el modelo, utilizando variables dummies de días de la semana y los días festivos que el CENACE considera, y la información del estado de tiempo diaria, con modelo de Machine Learning, dando como resultado un MAPE de 12%

**Semana 3:** Se inicia a considerar la serie de datos para predicción de la misma, sin tomar en cuenta los datos del estado del tiempo, con un modelo ARIMA, que da como resultado un MAPE de 10%. También se busca hacer un modelo SARIMA, debido a la estacionalidad, pero se decide considerar una regresión lineal simplemente.

**Semana 4:** Se buscar hacer una regresión lineal, por lo que se inicia a trabajar en las variables de entrada. Se establecen las dummies de:

* Día de la semana
* Mes del Año
* Días festivos relevantes: Navidad y Año Nuevo

Se analiza la serie de datos, y debido a la forma que tiene, y a la estacionalidad, se decide hacer un modelo en base a series de Fourier, utilizando senos y cosenos de las frecuencias más notorias en el modelo. Para hacer esto, se hizo un análisis de las frecuencias dibujando un periodograma. Ya teniendo todo en una tabla (Fourier y Dummies) se implementó el método simplificado del producto Kronecker. Se hizo el producto cruzado de todas las variables. El MAPE se situó, según las frecuencias elegidas, entre 4% y 6%

**Semana 5:** Debido al error generado, se agregó como variable dummie de día festivo el viernes santo y el Día de la independencia. Se volvieron a calcular las frecuencias, ya que en este punto, se tuvo la información actualizada por parte del CENACE hasta el 22 de septiembre. El MAPE para este momento es de 1.87%.

Hasta este punto, el MAPE que se estuvo calculando fue de la regresión como tal. Así que se genera un código para predecir la demanda de energía de los últimos n días, para poder comparar con los datos reales, y así poder tener un MAPE sobre el pronóstico. Se calcula el error sobre pronóstico entre 2% y 3%.

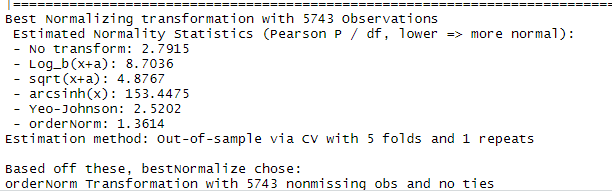
**Semana 6:** Al ver que posiblemente algunas de las frecuencias como 7 y 30, se estuvieran repitiendo con las variables dummies de días y meses, se decide hacer el análisis de las frecuencias con las variables dummies únicamente. De esta manera, se puede ver qué frecuencias son las que las variables dummies no alcanzan a cubrir, para ponerlas en el modelo de senos y cosenos, y evitar muchas variables repetidas en el modelo de regresión lineal. El MAPE decrece considerablemente hasta 1.6% sobre pronóstico a 5 días, sin embargo, el MAPE de la regresión aumenta a 3% aproximadamente.

**Semana 7:** Se decide analizar otros posibles días festivos que mejoren el desempeño de la regresión sin éxito. Se implementa un modelo de regresores que tome en cuenta la demanda de energía del día anterior, aumentando el MAPE, por lo que se decide no tomar en cuenta. Se busca hacer una transformación de los datos agregando el logaritmo natural, y al terminar regresar con exponente e el dato de la predicción para poder comparar con los datos reales, pero aumenta también el MAPE, por lo que no se toma en cuenta.

**Semana 8:** Se hace un análisis sobre la estacionalidad de la serie. El resultado se basó en dos factores que mejoraron el modelo: El primero es que la estacionalidad se basaba en el 7, es decir, cada semana. Esto fue obtenido de una descomposición de la serie Y el segundo: La forma de la serie no seguía senos y cosenos normales, sino senos y cosenos absolutos. Para agregar la estacionalidad, se metió en la tabla de las dummies el seno y coseno absoluto de 14 (para que al poner ambas, quedara igual que la descomposición. Y en las funciones de senos y cosenos para la regresión, se cambiaron a valores absolutos, dejando el MAPE en .78% para 5 días.

**Semana 9:** Se busca analizar las fechas de días especiales por aparte, ya que los MAPES en esos días aumentan hasta 20%, pero no se encontró algo relevante. Se agrega al modelo una transformación de datos en Box Cox, para ajustar la distribución de los datos a una normal. Al analizar los histogramas de los datos antes y después de esta transformación, se determinó que no era buena idea, dado que ambas eran bimodales, y la transformación no lo mejoraba.

**Semana 10:** Se decide buscar otras transformaciones, por lo que utilizando la función bestNormalize en R, pero ahora aplicada a los residuales del modelo de regresión lineal. Box Cox no se puede usar ya que los datos de entrada deben ser todos positivos.



Al buscar un modelo para los residuales como un ARIMA, se notó que ya sea con la transformación de orderNorm (la mejor según el coeficiente de Pearson), el mejor modelo es (2,0,0), es decir un modelo autoregresivo de orden 2.

**Semana 11**: Se busca modificar el AR 2 que se obtuvo por un modelo SARIMA, ya que al graficar los residuales de los residuales con AR, aún había algún dato periódico. Al hacer la regresión con todo lo anterior, el resultado del MAPE bajó considerablemente (de .78 %a .48%), pero al pronosticar datos de la serie de tiempo que no fueran los últimos 5 como normalmente se hacía, el MAPE incrementó.

**Semana 12:** Esta semana se trabajó en el reporte PAP, en la identificación de la problemática, la planeación de la alternativa y la validación de la alternativa.

**Semana 13:** Después del análisis de lo llevado a cabo durante el semestre, se encontró un error fundamental para la predicción, el MAPE aumentó debido a esto a 1.8%. Se trabajó en el reporte PAP, introducción, identificación del problema.

**Semana 14:** Durante esta semana se trabajó para agregar el AR en los residuales del modelo. Se corrigieron errores pasados. Además, se hizo una regresión Lasso para reducir el número de variables relevantes para el modelo, dejando el pronóstico a 5 días de .70%

Semana 15, se probaron modelos ARIMA encontrado un mejor pronóstico que el AR, siendo este (1,0,4). Se agrega al modelo y el MAPE final queda en .62% para 5 días.