UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

PRONÓSTICO A CORTO PLAZO DE AFLUENCIA DE PASAJEROS UTILIZANDO TÉCNICAS DE DATA MINING: METRO S.A.

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTION

DE OPERACIONES

DENISSE FABIOLA GARNICA PÉREZ

SANTIAGO DE CHILE

2011

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

PRONÓSTICO A CORTO PLAZO DE AFLUENCIA DE PASAJEROS UTILIZANDO TÉCNICAS DE DATA MINING: METRO S.A.

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTION

DE OPERACIONES

DENISSE FABIOLA GARNICA PÉREZ

PROFESOR GUIA:

RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN EVALUADORA

GASTÓN L´HUILLIER CHAPARRO
FABIÁN MEDEL GARCÍA
HÉCTOR ALTAMIRANO GUZMÁN

SANTIAGO DE CHILE

SEPTIEMBRE 2011

RESUMEN EJECUTIVO

La empresa Metro S.A. es parte fundamental del sistema de transporte público en la ciudad de Santiago, sin embargo, desde la puesta en marcha de Transantiago en Marzo del año 2007, la demanda de pasajeros creció explosivamente y es por esta razón, que no se contaba con información de calidad para poder tomar decisiones respecto de frecuencias de trenes, demanda de viajes, o la afluencia de pasajeros en las estaciones. Esta información es crucial para poder entender el tamaño de los andenes, trenes, cantidad de boleterías, incluso la densidad promedio en los vagones, etc.

El objetivo de este trabajo es poder probar la factibilidad de pronosticar la afluencia de pasajeros con técnicas de Minería de Datos con un error aceptable (inferior al 10%). Para que metro pueda utilizar esta metodología en el futuro para obtener información util respecto de la afluencia de pasajeros.

Para este estudio se seleccionaron cuatro estaciones y para cada una de ellas, se consideró un pronóstico de su afluencia a muy corto plazo (diario) y otro a corto plazo (mensual). Posteriormente, se seleccionó un conjunto de ocho métodos para el estudio de series de tiempo. Regresiones (lineales, logísticas, etc.), Suavización Exponencial, Redes Neuronales y Support Vector Regression. Se aplicaron transformaciones para eliminar estacionalidades y el efecto de Transantiago.

Para los pronósticos de afluencia de corto plazo todos los métodos seleccionados presentan errores inferiores al 10%. Sin embargo, el método de promedios móviles con horizonte de dos periodos presenta errores bajo el 3.3% en todas las estaciones estudiadas. Estos son excelentes resultados, por lo cual se recomienda ampliamente el uso de estas técnicas para este caso. Para el muy corto plazo, los modelos se comportan incluso mejor, con errores inferiores al 2%, por lo cual se puede decir que es posible utilizar estos métodos con gran confianza en ambos casos.

Este trabajo es para Yuyi, por ser el pilar más importante en mi estadía en Chile y en la vida, por su ayuda y cariño invaluables, y el sacrificio que toda renuncia representa.

AGRADECIMIENTOS

A mi padre, por darme el amor y la felicidad que son el mejor aliciente para seguir adelante y no temer a las adversidades. A mi madre, por creer siempre en mí y entregarme su amor incondicional, este emprendimiento no habría sido posible sin su constante apoyo. A mi hermano, porque nunca deja que desfallezca, es mi fuerza y la persona que da sentido a mi vida.

A mi profesor guía y Richard Weber por sus enseñanzas, orientación y constante aliento durante este proceso.

A Lucia, Pablo y Gérard, por ser tan incondicionales conmigo, su amistad, confianza, fuerza y apoyo en todo momento hacen que me sienta siempre en familia. A Julie por el apoyo entregado y las gestiones realizadas.

A la empresa Metro S.A. por toda la cooperación brindada para realizar este trabajo.

TABLA DE CONTENIDO

Página

INDICE DE FIGURAS

Página

INDICE DE TABLAS

Página

INTRODUCCIÓN

Metro es una Sociedad Anónima cuyo principal objetivo es entregar un servicio rápido, seguro y de excelencia a más de dos millones y medio de pasajeros diariamente, intentando, al mismo tiempo, mejorar su calidad de vida al conectarlos con la ciudad; de este modo, Metro se ha constituido en la columna vertebral del sistema de transporte público de Santiago (Transantiago).

Durante el año 2006 se concretaron tres proyectos importantes en el marco de la expansión de la red de Metro: extensión de las Líneas 2 y 4, e inauguración de la Línea 4A, los cuales permitieron aumentar en 17 km. la cobertura de la red, con 14 estaciones que incrementaron en un 24% los viajes en metro, completando un total de 84 km. de red con 92 estaciones a disposición de los clientes [**1**].

Este crecimiento trajo consigo grandes cambios a la empresa debido al aumento en la afluencia de pasajeros, al igual que lo ocurrido a partir de la fusión con Transantiago que a contar del 2007 dobló la cantidad de pasajeros recibidos diariamente, produciendo variaciones significantes en el comportamiento del mercado y que, en consecuencia, obligó a incorporar una importante dotación de personal e implementar procesos de capacitación.

Al momento de iniciar este estudio Metro tenía como próximos proyectos, la ampliación de su red con dos nuevos tramos, con la extensión de las lineas 1 y 5, que entraron en funcionamiento entre los años 2009 y 2010.

Estos acontecimientos sumados a las decisiones tomadas durante el año 2011 en cuanto a la construcción de las lineas 3 y 6, tienen gran relevancia en la justificación de la presente investigación, por ello es tan importante recalcarlos, pues sus resultados serán empleados en la programación de los trenes, sus movimientos y el desarrollo de las tablas de tiempos en cada una de las líneas que conforman la red de Metro.

CAPÍTULO 1: Descripción del problema

1.1. Motivación del estudio

El servicio de transporte de pasajeros de Metro S.A. es indispensable para la ciudad de Santiago, por ello es de suma importancia que sea capaz de realizar su operación diaria de la manera más eficiente posible. En este sentido, un aspecto crítico a conseguir es que la programación de los trenes y el servicio en general sean óptimos para brindar un servicio de calidad a los pasajeros, para lo cual es necesario conocer a priori la cantidad de pasajeros que harán uso del servicio diariamente (afluencia de pasajeros), pues a partir de esta información es posible generar las tablas de tiempo y movimientos, así como la programación de los trenes de cada una de las cinco líneas que se encuentran en operación.

Actualmente, se realizan cada doce meses predicciones trianuales de afluencia de pasajeros sujetas a modificaciones no oficiales. Estos cambios en los pronósticos, al no ser reconocidos de manera formal, generan errores porque su contenido no es entregado a todas las áreas de Metro lo que provoca fallas en las programaciones.

Por otro lado, el inicio de Transantiago en febrero de 2007 ha tenido un fuerte impacto en el sistema de transportes de la ciudad y, como consecuencia, en el Metro debido al incremento en aproximadamente un 50% en la afluencia de pasajeros. Esta alza explosiva provocó que las predicciones realizadas para el 2007 tuviesen un significativo margen de error, ya que los datos históricos no reflejaban el comportamiento real del sistema. Es por esta razón la empresa tiene la necesidad de contar con herramientas de mayor precisión en el pronóstico para hacer frente a los cambios y, de este modo, determinar cuáles son los patrones de comportamiento que generan variaciones en la afluencia de pasajeros en determinados períodos de tiempo. Realizar esto de manera precisa (con un bajo nivel de error), permitirá a la empresa cumplir con su nivel de servicio, gracias a una correcta programación de los trenes y sus tiempos de movimiento.

Por último, la empresa detectó la necesidad de realizar dichos pronósticos en un breve plazo para reaccionar a los cambios en la demanda de manera oportuna y rápida. Asimismo, para que Metro se embarque en nuevos proyectos de manera exitosa (extensiones de red y servicios *express*) necesita conocer todas las variables antes mencionadas.

En conclusión, esta investigación pretende explotar y buscar indicios sobre la utilidad de diversos métodos matemáticos como herramientas de pronóstico a corto plazo, para evaluar cuál es el más exacto en la entrega de sus resultados.

1.2. Planteamiento del problema

A partir de la puesta en marcha de Transantiago, el indicador de densidad en hora punta pasó de 4,72 pasajeros/m^2 promedio en el año 2006, a un promedio de 5,5 pasajeros/m^2 el 2007 (siendo el máximo registrado de aproximadamente 6,4 pasajeros/m^2), ésto debido al importante incremento en el número de pasajeros, llegando Metro a recibir un total de 270 millones de personas adicionales, es decir, 81,5% más que el 2006, totalizando 600 millones de viajes durante el año 2007 [**2**].

Este cambio explosivo en la demanda ha tenido graves repercusiones en el servicio brindado, por ello Metro determinó la necesidad de establecer un sistema eficiente que proyecte en un corto plazo la afluencia de pasajeros (principal problema enfrentado en la puesta en marcha de Transantiago) para mejorar el servicio, pues esta carencia es la causante de muchos problemas a nivel de planificación y entrega oportuna de respuestas.

Paralelamente, se encuentra la dificultad de no contar con información disponible sobre los patrones que influyen en las variaciones de la demanda, ya sea a nivel estacional, diario o según horario para las distintas estaciones.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar modelos de proyección de afluencia de pasajeros de corto plazo, entregando evidencias sobre su utilidad, para brindar a la empresa información de vital importancia en el desarrollo de la planificación diaria agregada.

1.3.2. Objetivos específicos

- Efectuar un levantamiento del estado del arte en técnicas de *Data Mining* para pronósticos de series de tiempo y de las metodologías existentes para este proposito.

- Determinar las técnicas de *Data Mining* que permiten obtener la mejor proyección de afluencia de pasajeros, evaluando su desempeño predictivo.

- Evaluar los modelos utilizados y comparar sus resultados para establecer qué tipo de modelo es capaz de generar un mejor pronóstico de la afluencia de pasajeros en el muy corto y corto plazo.

- Analisar los resultados para entender los factores que alteran el comportamiento de la afluencia de pasajeros en el tiempo.

1.4. Justificación del estudio

La afluencia de pasajeros es la base de la cadena productiva de Metro debido a que permite realizar planeación agregada, desarrollar programas de circulación de trenes, tableros de servicio de conducción y tableros de rotación de horario, convirtiéndola en información de vital importancia para el desarrollo de los nuevos proyectos. Todo lo anterior justifica la necesidad de generar los niveles de oferta de transporte que serán entregados al finalizar la presente investigación.

Por otra parte, identificar los patrones que producen variaciones en la demanda, ya sea a nivel estacional, diario o según horario, es de vital importancia, pues Metro no

cuenta con esta información que le permitiría enfrentar de manera anticipada los cambios en la demanda y, por ende, aumentar la calidad del servicio entregado a sus clientes.

1.5. Alcances del estudio

En el ámbito organizacional de Metro, el presente estudio se enmarca dentro de la Gerencia de Operaciones, quedando bajo la supervisión directa de la Subgerencia de Ingeniería de Operaciones, y busca entregar evidencias respecto a la utilidad de las técnicas de *Data Mining* como herramienta para el pronóstico de afluencia de pasajeros a corto plazo, así como de las fortalezas que la identificación de patrones de comportamiento trae consigo.

Para ello, la investigación será estructurada siguiendo parámetros espaciales, pues se trabajará sólo con algunas estaciones representativas, a partir de las cuales se inferirán los patrones de comportamiento de la demanda y el pronóstico de afluencia de pasajeros, por esto, en conjunto con la empresa, se escogieron las siguientes estaciones de la Línea 1:

- Estación San Pablo
- Estación Escuela Militar
- Estación Pedro de Valdivia
- Estación Universidad de Chile

Las dos primeras fueron escogidas por su condición de estaciones terminales, la estación Pedro de Valdivia debido a su condición de punto neurálgico de la Línea 1 (según el criterio de los funcionarios de Metro), y la estación Universidad de Chile por estar ubicada en el centro de Santiago.

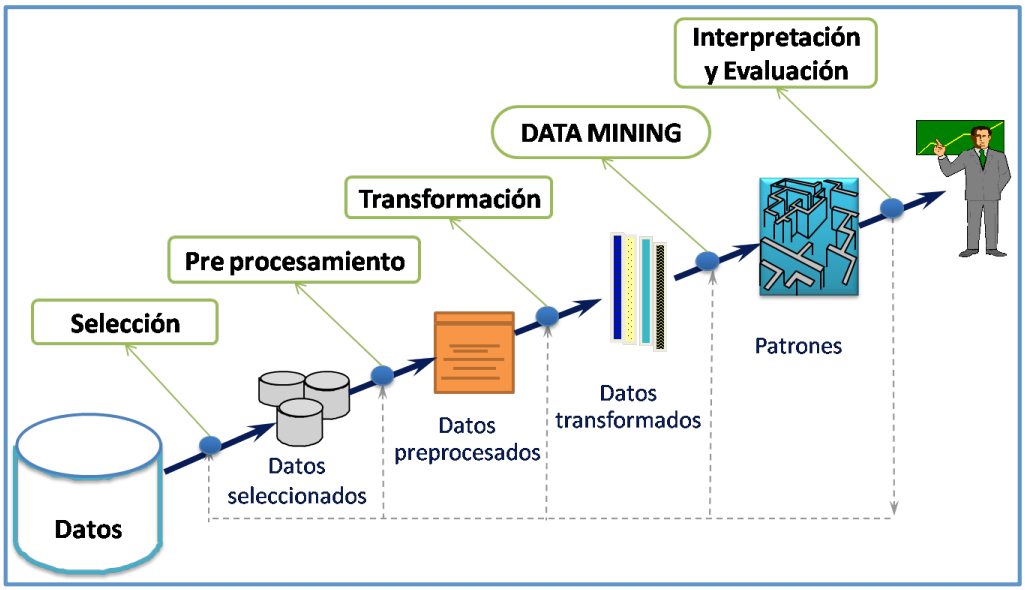
Por otro lado, el principal elemento a utilizar, en tanto parámetro de comparación, es la información relativa a la afluencia diaria de pasajeros contabilizada cada 15 minutos (paso por torniquete) a contar del año 2002 en las cuatro estaciones antes mencionadas.

Se realizará la proyección a corto plazo utilizando la afluencia histórica desde el año 2002 al año 2009. Utilizando este último para poder evaluar la calidad de los pronósticos generados. Se espera poder evaluar más de dos modelos o técnicas de pronóstico, para poder establecer cuál es la que permite generar mejores resultados.

1.6. Metodología utilizada

El proceso de descubrir conocimiento en Bases de Datos -*Knowledge Discovery in Database (*KDD)- [**3**] desarrolla las ideas centrales del procesamiento de datos, con el objetivo de extraer e interpretar patrones [**4**,**5**,**6**,**7**] para generar a partir de ello un nuevo conocimiento respecto del estudio, en este caso, los pronósticos a corto plazo. Por estas características el Proceso KDD (Figura 1) cumple con los requisitos para convertirse en la base metodológica del estudio aquí presentado.

Figura 1: Proceso KDD *(Knowledge Discovery in Databases).*



Fuente: Elaboración propia.

El Proceso KDD parte con una comprensión básica del dominio en que surge el fenómeno analizado y la identificación del objetivo deseado, para luego continuar con las siguientes etapas, cuya aplicación podría ser recursiva [**3**,**5**,**8**]:

- Levantamiento de datos y recolección de casos favorables para la investigación.
- Manejo de datos: comprobar que sean correctos, no ambiguos, consistentes y completos.
- Pre-procesamiento de datos: identificar valores perdidos y fuera de rango para efectuar transformaciones o tomar las medidas correctivas necesarias como, por ejemplo, eliminación de aquellos datos irrelevantes para la investigación.
- Data Mining: determinar las técnicas predictivas adecuadas para la investigación, e identificar los patrones de comportamiento.
- Desarrollo, resolución y validación del Modelo.
- Análisis de resultados (interpretación y evaluación).

Basados en el proceso de KDD, varias metodologías han sido creadas para facilitar el trabajo en esta área. Dos de las metodologías más utilizadas en el mundo son CRISP-DM (Cross Industry Standard Process for Data Mining) [**9**] y SEMMA (sample, explore, modify, model, assess) [**10**]. Aunque en el sitio oficial de SAS[1], se indica que un error común es usar SEMMA como una metodología, puesto que más bien, es una organización lógica de todas las funcionalidades incluidas en SAS. Sin embargo, hay numerosos trabajos que utilizan SEMMA como una metodología formal [**10**,**11**,**12**].

[1] http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

En este trabajo se optó por usar el proceso de KDD, en lugar de CRISP-DM, SEMMA u otras metodologías ya que es un proceso simple y claro de explicar a las personas de Metro S.A.

1.7. Esquema de Solución

En *Data Mining* se han incorporado distintas herramientas matemáticas para la extracción de información desde grandes bases de datos. Aquí se encuentran técnicas como: árboles de decisión, reglas de asociación, modelos econométricos y redes neuronales, entre otras.

Los modelos econométricos son sistemas matemáticos que generan patrones a partir de los datos recibidos, con el fin de minimizar los errores asociados. Estos han probado tener capacidades suficientes para resolver los problemas de predicción [**13**].

Una de las aplicaciones de estos modelos es el pronóstico de series de tiempo, donde no se busca explicar la variable de interés, sino sólo pronosticarla según su pasado comportamiento. Los modelos de series de tiempo siempre son más precisos en sus proyecciones que los modelos econométricos más complejos, algunos de los cuales poseen múltiples ecuaciones y decenas de variables [**14**], es por esto que, para efectos de la presente investigación, se evaluará su aplicación en el pronóstico de afluencia de pasajeros, el cual permitirá a la empresa realizar la planeación agregada diaria.

1.8. Resultados esperados

Se busca desarrollar un sistema de pronóstico de afluencia de pasajeros para Metro S.A., que se convierta en el sustento principal de la planeación agregada de su cadena productiva. Es por esto que en el transcurso de la investigación se desarrollarán los siguientes procesos:

- Se estudiarán los patrones de comportamiento de la afluencia de pasajeros, buscando que la empresa sea capaz de enfrentar los cambios de manera oportuna y rápida.
- Se construirán modelos de series de tiempo para la proyección a corto plazo de la afluencia de pasajeros.
- Se encontrará la relación existente entre los distintos patrones de afluencia, con el fin de determinar si esta relación aporta a la proyección de corto plazo.
- Se comparará y elegirá el modelo más adecuado para la proyección de afluencia de pasajeros.

Con todo lo mencionado se pretende aportar evidencias respecto al potencial de las técnicas de *Data Mining* para el pronóstico de afluencia de pasajeros e identificación de los factores de comportamiento de la demanda.

Capítulo 2: Marco Teórico

2.1. *Data Mining*: Definiciones y distinciones

Data Mining es un elemento del proceso de descubrimiento de conocimiento en Bases de Datos (*Knowledge Discovery in Database,* KDD), definido como un procedimiento estándar bien estructurado, íntimamente conectado con administradores, tomadores de decisión y aquellos envueltos en el despliegue de resultados. Esta técnica supone la aplicación iterativa de datos y reúne ventajas de diversas áreas de la Ingeniería, a saber, estadística, lógica, computación y procesamiento masivo, entre otras.

Los objetivos de KDD, metodología utilizada para el desarrollo de esta memoria, se definen por el uso previsto del sistema, distinguiéndose dos objetivos [**3**]:

a. **Verificación:** el sistema se limita a comprobar la hipótesis del usuario.

b. **Descubrimiento:** el sistema encuentra de forma autónoma nuevos patrones. Este objetivo el que más ayuda en la predicción, ya que el sistema busca esquemas de comportamiento para pronosticar la futura conducta de algunas entidades.

A partir de la definición de estos objetivos, se puede advertir que *Data Mining* es una metodología que permite descubrir nuevas correlaciones significativas, ajuste de modelos, identificación de relaciones entre datos o descubrimiento de patrones de comportamiento, a través de la observación y tamizado de grandes cantidades de datos almacenados. Es un proceso no trivial de identificación válido, novedoso, potencialmente útil y entendible, que permite descubrir patrones comprensibles que se encuentran ocultos en los datos [**3**], usando tecnologías de reconocimiento, como técnicas estadísticas y matemáticas [**15**].

Por otro lado, desde un punto de vista empresarial, podría asumirse la siguiente definición de *Data Mining*: es la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión [**4**,**5**,**6**,**7**,**8**]. Organizaciones innovadoras de todo el mundo ya utilizan *Data Mining* para localizar y atraer a los clientes de mayor valor, volver a configurar su oferta de productos, aumentar las ventas y reducir al mínimo las pérdidas debidas a un error o fraude [**16**].

Asimismo, *Data Mining* es considerada por la literatura como una tecnología basada en prácticas como la recolección masiva de datos, las cuales pasan por diversas etapas y áreas integradas, apoyándose en el uso de algoritmos concretos, confiables y entendibles, para obtener patrones de comportamiento a partir de los datos pre-procesados [**17**].

Por lo tanto, el principal objetivo de *Data Mining* es investigar los antecedentes presentes en las bases de datos que podrían almacenar información de muchos años. Para este propósito utiliza herramientas que se combinan fácilmente y pueden ser analizadas y procesadas con rapidez, las cuales, además, permiten extraer datos relevantes de archivos corporativos o registros públicos mediante instrumentos

indagatorios, para efectuar preguntas *ad-hoc* y obtener respuestas expeditas que lleven al descubrimiento de resultados valiosos y, en algunos casos, inesperados [**17**].

Entonces, es posible afirmar que *Data Mining* aumenta considerablemente la capacidad y precisión de análisis de importantes volúmenes de datos, impactando de manera directa en las acciones de la empresa al descubrir nueva información, optimizar valores y dar fiabilidad a las hipótesis planteadas.

En esta situación la técnica utilizada se denomina **modelado**, a saber, el acto de desarrollar el modelo de una situación puntual, donde ya se conocen las respuestas, y aplicarlo a una situación desconocida, pero siempre restringido por las exigencias establecidas para la obtención de resultados como, por ejemplo, la elaboración de patrones de comportamiento que pueden ser utilizados en el proceso de pronóstico. Dichos resultados se pueden ser categorizar en cinco tipos [**4**,**5**,**15**]:

- **Descripción**. Sugiere las posibles explicaciones para los patrones y tendencias descubiertas; en esta tarea las técnicas de *Data Mining* permiten una interpretación y explicación intuitiva por parte del analista. A menudo es posible lograr una descripción de alta calidad al emplear un análisis exploratorio y un método gráfico de exploración de datos.

- **Estimación**. Tiene como variable objetiva una cifra numérica cuyos modelos son construidos usando registros completos, los cuales proporcionan tanto el valor de la variable objetivo como el de las variables que permiten predecir. Para nuevas observaciones se realizan estimaciones de la variable objetivo, basadas en los valores de las variables predictivas.

- **Predicción**. Para el proceso de predicción los resultados están en el futuro. Cualquiera de los métodos y técnicas usados para clasificar y estimar también pueden ser empleados en un proceso predictivo pero en circunstancias apropiadas. Estos incluyen métodos tradicionales estadísticos de estimación de punto y estimaciones de intervalos de confianza, regresión lineal simple, correlación y múltiple regresión, además de algunos métodos de descubrimiento de conocimiento, como redes neuronales y árboles de decisión, entre otros.

- **Clasificación**. En ella hay una variable objetivo categórica. Los modelos de Minería de Datos examinan un gran conjunto de registros, y cada uno de ellos contienen información, tanto de la variable objetivo como de las variables predictivas.

- ***Clustering*** **(agrupamiento)**. Se refiere al agrupamiento de registros, observaciones o casos en subclases de objetos similares. Un clúster es una colección de registros similares, pero que diferencian a los registros en otros clúster. En *Clustering* no se define una variable objetivo ya que el propósito no es clasificar, estimar o predecir el valor de ésta, sino segmentar el conjunto de datos en subgrupos relativamente homogéneos, tratando de maximizar la semejanza entre los registros dentro de un subgrupo, y minimizar la semejanza con los registros de otros subgrupos. Usualmente el *clustering* se usa en tanto paso preliminar para el proceso de Minería

de Datos, donde los clúster resultantes son usados como datos de entrada para otras técnicas diferentes, como por ejemplo, redes neuronales.

- **Asociación**. Es el proceso de descubrir atributos similares, es decir, establecer reglas para cuantificar la relación entre dos o más atributos. Las reglas de asociación son de la forma: *"Si X es el antecedente, entonces Y es la consecuencia"*, en conjunto con la medida del soporte y la confianza asociada con la regla.

En conclusión, en *Data Mining* se recopilan datos esperando que de ellos surjan las hipótesis necesarias para inferir una descripción o explicación de su existencia. De esta manera, para validar la hipótesis inspirada por los mismos datos, se debe presentar un enfoque exploratorio y no uno confirmador.

Contando con bases de datos grandes y de buena calidad, *Data Mining* puede generar nuevas oportunidades de negocios al proveer las siguientes capacidades [**17**]:

- Predicción automatizada de tendencias y comportamientos. *Data Mining* automatiza el proceso de búsqueda de información predecible en grandes bases de datos. Por ejemplo, identifica segmentos de población que probablemente respondan de manera similar a determinados eventos.

- Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de *Data Mining* barren las bases de datos e identifican modelos previamente escondidos en un solo paso.

Para adentrarse en un proyecto de *Data Mining*, independiente de la técnica específica de extracción de conocimiento, se deben seguir los siguientes pasos (revisar figura 2):

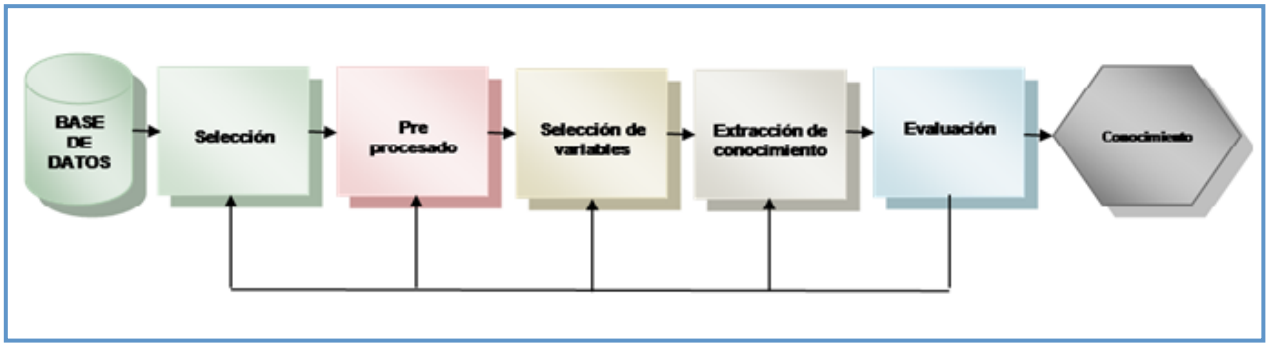
a. **Filtrado de datos**. La manera en que los datos están organizados dentro de las bases, en general, no es idónea ni permite su utilización como materia prima para un algoritmo de *Data Mining*, por esta razón, es necesario realizar un pre-procesamiento de ellos, es decir, evaluar la calidad de los datos, realizarles una limpieza, filtrarlos o transformarlos según los requerimientos de los resultados, eliminando datos incorrectos, inválidos o desconocidos. El objetivo principal es minimizar la 'basura' que entra al modelo. Es importante hacer una descripción de los datos a partir de esto, por ejemplo, un resumen de sus atributos estadísticos (como medias y desviaciones estándar), un examen visual usando tablas y gráficos, y buscar vínculos entre los datos (en tanto valores que a menudo se presentan juntos) [**16**].

b. **Selección de variables**. A veces, pese a realizar un pre-procesado, igualmente se cuenta con una gran cantidad de datos, para esto una óptima solución es seleccionar sólo una característica, es decir, elegir las variables con más influencia en el problema pero que no afectan la calidad del modelo. Los métodos de selección de variables son: los basados en la elección de los mejores atributos del problema, los que buscan variables independientes mediante Test de Sensibilidad, algoritmos de distancia o heurísticas.

c. **Extracción de conocimiento**. Con el uso de cualquier técnica de *Data Mining* es posible obtener un modelo de conocimiento que represente los patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre ellas.

d. **Interpretación y evaluación.** Con la obtención de uno o más modelos, se procede a su validación, o bien, a compararlos en busca del que se ajuste mejor al problema.

Figura 2: Fases para el desarrollo de un proyecto de Data Mining.



Fuente: Elaboración propia.

Es válido recalcar que *Data Mining* no es un artilugio mágico, es un error creer que la Minería de Datos es un conjunto de herramientas aisladas, aplicadas en el departamento de análisis y relacionados sólo inconsecuentemente con el negocio establecido o el esfuerzo de la investigación, requiere de manera significativa la intervención humana en cada etapa, incluso luego que el modelo se haya desplegado, pues a menudo la introducción de nuevos datos requiere de una actualización.

La supervisión de calidad y otras medidas evaluativas deben ser realizadas por analistas humanos, de este modo, el proceso permitirá descubrir patrones de comportamiento, pero depende del analista encontrar su origen [**15**]. Por otro lado, las relaciones de predicción encontradas no son necesariamente las causas de una acción o comportamiento. Para asegurar resultados significativos, es vital la comprender los datos [**16**].

2.2. *Data Mining* Predictivo

El objetivo de *Data Mining* es producir nuevos conocimientos con los que el usuario pueda operar. Para ello es necesaria la construcción de un modelo basado en datos recogidos a partir de diversas fuentes: operaciones corporativas, historial de clientes, información demográfica, datos de control de procesos y bases de datos externas (por ejemplo, información de la oficina de crédito o datos meteorológicos). El resultado del modelo es una descripción de los patrones y relaciones existentes en los datos, que pueden sesr utiliza dos con cierto grado de confianza en el proceso de predicción [7,16].

Para evitar confundir los diferentes aspectos de *Data Mining*, es posible plantear la siguiente pregunta [**16**]: ¿cuál es el objetivo de la aplicación de la herramienta? Una posible respuesta sería la búsqueda de patrones en los datos que ayude a mantener los buenos clientes, de esta manera es posible construir un modelo para predecir la rentabilidad del cliente y otro para identificar a los clientes que tienden a dejar la empresa.

Con esto claro, entonces, lo siguiente es establecer los diversos tipos de modelos de *Data Mining* predictivo que existen. Si bien existe un gran número de modelos, éstos pueden clasificarse en las siguientes grandes familias [**4**,**5**,**15**,**18**]:

a. **Clasificación.** Una de las aplicaciones más populares de Minería de Datos corresponde a esta categoría. Para esto, en primer lugar, se requiere un conjunto conocido de categorías y, en segundo lugar, un conjunto de puntos que se sabe pertenecen a una de estas categorías pero se desconoce específicamente a cuál. En este caso los algoritmos de minería de datos toman un elemento y son capaces de clasificarlo dentro de una o más categorías conocidas con un cierto margen de error. Aplicaciones de este tipo son las de reconocimiento de caracteres en las impresoras (OCR), así como las herramientas para reconocimiento de escritura manuscrita.

b. **Clustering.** Como ya se mencionó, estos modelos provienen del concepto inglés clúster que en español significa 'agrupamiento'. En este caso, y a diferencia de los modelos de clasificación, no se conocen las categorías a priori. En consecuencia, estos algoritmos buscan generar clústers o grupos donde las observaciones o elementos estén muy relacionados entre sí. Para esto normalmente los algoritmos utilizan una medida de similitud. Los algoritmos más usuales son k-means, k-medoids, Self Organizing Map (o SOM), *clustering* aglomerativo, clúster jerárquico, entre otros.

c. **Regresión.** Utiliza valores tomados de una nube de puntos en el espacio para intentar calzar una ecuación, que en el caso más simple puede ser una recta (también llamada regresión lineal). Aunque existen regresiones polinomiales y exponenciales, ente otras, este tipo de modelos son los de más simple aplicación y normalmente se encuentran en muchos de los programas estadísticos o de Minería de Datos, como R, SPSS, SAS, etc. Usualmente se utiliza el factor de correlación lineal o el factor R cuadrado para comprobar la calidad de la regresión. Una vez practicado el modelo es posible predecir uno o más periodos posteriores a los datos utilizados para el entrenamiento.

d. **Series de Tiempo.** Este tipo de modelos predice valores futuros que son desconocidos, basados en la historia acumulada de una o más variables. A diferencia de las regresiones, en una serie de tiempo el orden de las observaciones es importante porque se trata de valores que ocurren en un orden determinado. Los modelos más tradicionales para analizar series de tiempo son los promedios móviles, modelos ARIMA, X11 y Suavización Exponencial. El modelo obtenido en este caso también es utilizado para predecir el valor de la variable objetivo en los meses, días, años o períodos para los cuales sea necesario. Normalmente, el error de estos modelos es calculado utilizando la del promedio de la desviación sobre el

promedio o *Medium average deviation* (MAE), también es muy usado el *Mean Average Deviation* cuyas siglas son, lamentablemente, MAE también. Asi, mismo se puede obtener el MAE porcentual, a este indicador se le conoce como MAPE (mean absolute percentage error). [**18**,**19**]

2.3. Técnicas de Series de Tiempo

Si bien existen estas cuatro familias de Minería de Datos predictivas, los datos con los que se cuenta para esta investigación son series de tiempo. Por ende, y dado que la afluencia de pasajeros tiene una fecha y hora específicas, los modelos a aplicar en primer lugar son los de series de tiempo.

Aunque existen más métodos de pronóstico, por simplicidad se presentan sólo los considerados más usuales y sencillos de llevar a cabo: Promedios Móviles y Suavización Exponencial.

Estos pueden utilizarse en los siguientes contextos: cuando hay información disponible de la variables que se pronostica, la información puede ser cuantificada, si se considera razonable que el patrón de comportamiento del pasado continuará en el futuro, si se cuenta con una base de datos histórica y se desea pronosticar una variable considerando su comportamiento pasado; entonces, es posible utilizar el Método de Promedios Móviles o el Método de Suavización Exponencial, conocidos también como Métodos de Series de Tiempo[2].

2.3.1. Método de Promedios Móviles

La utilización de esta técnica supone que la serie de tiempo es estable, es decir, los datos que la componen se generan sin variaciones importantes entre un dato y otro (error aleatorio=0)[3], lo cual significa que el comportamiento de los datos, aunque muestre un crecimiento o un decrecimiento, lo hará con una tendencia constante.

Al usar el Método de Promedios Móviles [**4**,**15**,**18**,**20**] se supone que todas las observaciones de la serie de tiempo son igualmente importantes para la estimación del parámetro a pronosticar (en este caso la afluencia de pasajeros). De esta manera, se utiliza como pronóstico para el siguiente periodo el promedio de los n valores pertenecientes a los datos más recientes de la serie de tiempo. Utilizando una expresión matemática, se obtiene:

$$Promedio\ Movil = \frac{\sum(n\ valores\ más\ recientes\ de\ la\ serie\ de\ tiempo)}{n}$$

Ecuación 1

[2] Una serie de tiempo es un conjunto de observaciones respecto a una variable, medidas en puntos sucesivos en el tiempo o a lo largo de periodos sucesivos. Un análisis de una secuencia de datos se conoce como análisis de series de tiempo de una variable.

[3] El error aleatorio muestra el grado de confiabilidad con que se van a comportar los datos. La variación del error puede ser de 0 a 1, en donde, un error aleatorio = 0 muestra una total confiabilidad del comportamiento de los datos, y un error aleatorio = 1 muestra que los datos no son confiables en su comportamiento.

En la Ecuación 1, el término n indica que, conforme se tiene una nueva observación de la serie de tiempo, se reemplaza la más antigua de la ecuación y se calcula un nuevo promedio. El resultado es un desplazamiento del promedio (un periodo en el futuro), y en la medida que se obtienen nuevos datos, se sustituyen en la fórmula y generan una modificación del valor promedio.

No existe una regla específica que indique cómo seleccionar la base del promedio móvil. Si la variable a pronosticar no presenta variaciones considerables, esto es, si su comportamiento es relativamente estable en el tiempo, se recomienda que el valor de n sea grande. Por el contrario, es aconsejable un valor pequeño de n si la variable muestra patrones cambiantes. En la práctica, los valores de n oscilan entre 2 y 10 [**3**].

El método de promedios móviles es muy útil cuando se tiene información no desagregada, y cuando no se conoce otro método más sofisticado que permita predecir con mayor confianza.

2.3.2. Método de Suavización Exponencial

Otro método para realizar un pronóstico es el método de suavización exponencial [**4**,**15**,**18**,**20**]. A diferencia de los promedios móviles, este método predice otorgando una ponderación a los datos dependiendo del peso que tengan dentro del cálculo. Esta ponderación se lleva a cabo otorgando un valor a la constante de suavización α, y puede ser mayor que cero y menor a uno.

El método de Suavización Exponencial supone que el proceso es constante, al igual que el método de Promedios Móviles. Sin embargo, es un promedio ponderado de los valores reales y los valores pronosticados, a diferencia del otro método, en donde los datos para calcular el promedio tienen la misma ponderación. De manera particular, esta técnica considera que las observaciones recientes tienen mayor valor y le otorga mayor peso dentro del promedio.

$$F_0 = X_0$$
$$F_t = \alpha X_{t-1} + (1-\alpha)F_{t-1}$$

Ecuación 2

La Suavización Exponencial utiliza un promedio móvil ponderado de los datos históricos de la serie de tiempo como pronóstico; es un caso especial de Promedio Móvil en donde se selecciona un solo valor de ponderación[4]. El modelo básico de suavización exponencial se presenta la Ecuación 2.

En esta expresión se presentan el pronóstico del periodo actual y el anterior; y X_{t-1} es el valor real de la variable a predecir en el periodo t. El parámetro α es la constante de suavización exponencial $\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}$.

[4] La ponderación se determina considerando el peso que se le asigna al valor más reciente de la serie. Los pesos o ponderaciones para los demás valores se determinan automáticamente, haciéndose más pequeños conforme las observaciones se alejan del presente.

La utilización de esta ecuación implica algunas especificaciones. El cálculo de F_t está ligado con los dos períodos anteriores. En otras palabras, el pronóstico de suavización exponencial en determinado periodo es F_t es igual al valor real de la serie de tiempo en el periodo anterior X_{t-1}, por la constante de suavización α, más $(1-\alpha)$ por el pronóstico del periodo anterior F_{t-1}.

A pesar de que la suavización exponencial entrega un pronóstico que es un promedio ponderado de todas las operaciones anteriores, no es necesario guardar todos los datos del pasado a fin de calcular el pronóstico para el periodo siguiente. De hecho, una vez seleccionada la constante de suavización α, sólo se requiere de dos elementos de información para calcular el pronóstico. La Ecuación 2 muestra que con un α dado se puede calcular el pronóstico para el periodo t, simplemente conociendo los valores reales y pronosticados de la serie de tiempo para el periodo t, es decir, X_{t-1} y F_{t-1}.

La elección de la constante de suavización α es crucial en la estimación de pronósticos futuros. Si la serie de tiempo contiene una variabilidad aleatoria sustancial, se preferirá un valor pequeño como constante de suavización. La razón de esta aseveración es que gran parte del error de pronóstico es provocado por la variabilidad aleatoria, por lo que un valor pequeño de α permite un pronóstico mejor. Por el contrario, para una serie de tiempo con una variabilidad aleatoria relativamente pequeña, valores más elevados de la constante de suavización tienen la ventaja de ajustar con rapidez los pronósticos cuando ocurren errores de pronóstico, y permiten, por lo tanto, que el pronóstico reaccione con mayor rapidez a las condiciones cambiantes. En la práctica, el valor de α está entre 0.01 y 0.90 [**3**].

2.4. Métodos de Regresión

Los métodos de regresión también serán aplicados en este trabajo debido a su naturaleza. Consisten en pasar una ecuación, que puede ser lineal, exponencial, logarítmica, etc., por una nube de puntos. A continuación, se explica cada uno de los modelos de regresión aplicados en la investigación.

2.4.1. Método de Regresión Lineal

El modelo lineal relaciona la variable dependiente Y con K variables explicativas Xk (k = 1,...K), o cualquier transformación de éstas que generen un hiperplano de parámetros βk desconocidos [**15**]. En este caso, por ejemplo, la afluencia de pasajeros en la Estación Escuela Militar depende del día de la semana que se contabilice. Se puede afirmar que la variable Y (la afluencia de pasajeros en la estación) es dependiente del día de la semana X. A esto se le llama un modelo de regresión simple donde:

$$Y = f(x)$$

Ecuación 3

Si consideramos que la relación f, que liga Y con X, es lineal, entonces la [5,21,22] se puede escribir así:

$$Y_t = \beta_1 + \beta_2 X_1$$

Ecuación 4

En cualquier caso, las relaciones del tipo anterior raramente son exactas, más bien se trata de aproximaciones en las que se han omitido muchas variables de importancia secundaria, por esta razón se debe incluir un término de perturbación aleatoria ut que refleje todos los factores –distintos de X- con influencia sobre la variable endógena, pues ninguno de ellos es relevante individualmente. Con ello, la relación quedaría de la siguiente forma:

$$Y_t = \beta_1 + \beta_2 X_1 + \mu_1$$

Ecuación 5

El objetivo principal de la regresión es la determinación de β_1 y β_2 a partir de la información contenida en las observaciones que se dispone. Esta estimación se puede llevar a cabo mediante diversos procedimientos.

En ocasiones la relación entre X e Y no es lineal, sin embargo, una transformación de estas variables sí lo es. Es decir, se puede aplicar logaritmos y el resultado, X vs Log(Y), puede guardar una relación lineal, en este caso, se dice que se ha linealizado la relación entre X e Y (a este modelo se le llama regresión logarítmica). A continuación, se enunciarán las ecuaciones de diversos modelos de regresión lineal (ver Tabla 1).

Tabla 1: Modelos de Regresión Lineal

MODELO	ECUACIÓN	O TAMBIÉN
Logarítmica	$Y = \beta_0+\beta_1 \ln(X)$	
Inversa	$Y = \beta_0 +\beta_1 . X$	
Cuadrática	$Y = \beta_0 + \beta_1 X + \beta_2 X_2$	
Cúbica	$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3$	
Potencia	$Y = \beta_0 X^{\beta 1}$	$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(X)$
Compuesto	$Y = \beta_0 \beta_1 X$	$\ln(Y) = \ln(\beta_0) + X \ln(\beta_1)$
Curva-S	$Y = \exp^{\beta 0} + \beta_1/t$	$\ln(Y) = \beta_0+\beta_1/t$
Crecimiento	$Y = \exp^{(\beta 0+\beta 1 X)}$	$\ln(Y) = \beta_0 + \beta_1 X$
Exponencial	$Y = \beta_0 \exp^{\beta 1} X$	$\ln(Y) = \ln \beta_0 + \beta_1 X$

Fuente: Elaboración propia .

2.4.2. Método de Regresión Polinomial

Cuando la relación entre las variables dependientes e independientes es no lineal, puede ser de gran ayuda incluir términos polinomiales para explicar la variación de la variable dependiente.

Los términos polinomiales se incluyen, para todos los efectos, en este modelo, por ello se puede realizar una estimación y, a partir de estos resultados, determinar si la curva polinomial que se asume es estadísticamente significativa para los datos de la investigación [**20**].

2.5. Método de Redes Neuronales Artificiales

Las Redes Neuronales Artificiales o RNA están compuestas por un gran número elementos de procesamiento altamente interconectados (Neuronas) que trabajan al simultáneamente para dar solución a problemas específicos [**4**,**15**]. Las RNA, tal como las personas, aprenden de la experiencia. Se trata de un modelo capaz de manejar las imprecisiones e incertidumbres que surgen cuando se intenta resolver problemas relacionados con el mundo real (reconocimiento de formas, toma de decisiones, etc.), ofreciendo soluciones robustas y de fácil implementación. Las RNA están compuestas de muchos elementos sencillos que operan en paralelo, el diseño de la red está determinado mayormente por las conexiones entre sus elementos, al igual que las conexiones de las neuronas cerebrales. Han sido entrenadas para la realización de funciones complejas en variados campos de aplicación. Hoy pueden ser empleadas para resolver problemas complejos en sistemas computacionales o situaciones del ser humano.

Las Redes Neuronales fueron concebidas originalmente con el objetivo de modelar la biofisiología del cerebro humano [**23**,**24**,**25**], es decir, entender y explicar cómo funciona y opera. La meta era la creación de un modelo capaz en emular el proceso humano de razonamiento. La mayor parte de los trabajos iniciales en redes neuronales fue realizada por fisiólogos y no por ingenieros [**26**].

Las RNA pueden tener factores de peso fijos o adaptables. Aquellas con pesos adaptables emplean leyes de aprendizaje para ajustar el valor de la fuerza de interconexión con otras neuronas. Si las neuronas utilizan pesos fijos, entonces su tarea debe estar previamente definida. Los pesos son determinados a partir de una descripción completa del problema. Por otra parte, los pesos adaptables son esenciales si no se conoce previamente cuál debe ser su valor correcto.

En los años 50's y 60's, se efectuaron diversos intentos por adaptar los patrones de Redes Neuronales [**26**] a la generación de aprendizaje. Rosenblatt diseñó el modelo Perceptron, el cual contenía tres tipos de neuronas:

a. Sensoriales: tomaban entradas desde afuera de la red.
b. Asociativas: eran sólo internas.
c. De respuesta: propagaban señales afuera de la red hacia el mundo externo

La distinción entre estas tres neuronas es importante aún, sin embargo, ahora se

refieren como unidades de entrada, de salida y ocultas. Rosenblatt desarrollo métodos para alterar los niveles sinápticos buscando que la red aprendiera a reconocer tipos de entradas. Por ejemplo, produjo una red que aprendió a responder a líneas verticales pero no a horizontales (pues se sabe que neuronas especializadas en la visión actúan de esta forma).

Como muchas Redes Neuronales posteriores, el rasgo más importante del Perceptron de Rosenblatt fue aprender a clasificar sus entradas; lo cual contrasta con la clásica ciencia computacional donde el programador escribe un programa que le indica a la computadora cómo clasificar sus entradas.

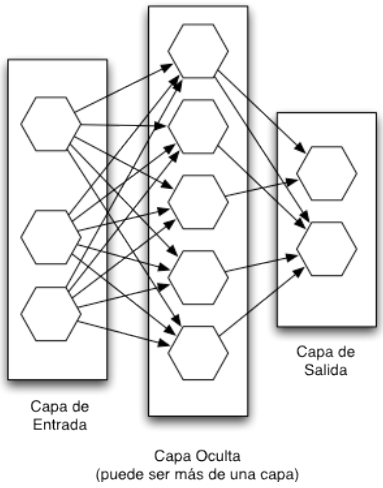
En 1969, publicaron un libro que influenció notoriamente el pensamiento sobre Redes Neuronales [**27**]. En primer lugar, demostraron que para efectuar ciertas importantes tareas de clasificación en geometría requerían de un incremento arbitrario en el tamaño del Perceptron, mientras aumentaba el tamaño de la retina. En segundo lugar, concluyeron que los Perceptrones eran incapaces de aprender a resolver cualquier problema linealmente inseparable, porque dificultades muy simples son linealmente inseparables (por ejemplo, el XOR).

Posterior a los años 80's el campo de la Inteligencia Artificial se hizo popular. En él la inteligencia se modela *top-down* con algoritmos diseñados para modelar procesos mentales de alto nivel, como la asociación de conceptos, deducción, inducción y razonamiento.

2.5.1. Método de *Back Propagation*

Las Redes Neuronales se pusieron de moda en 1986, cuando Rumelhart y McClelland demostraron que algunas de las distinciones imposibles para Perceptrons simples, pueden ser resueltas por redes multi-nivel con funciones de activación no-lineales usando un procedimiento [**28**]simple de entrenamiento: los algoritmos *back-propagation* [**15**,**27**].

Figura 3: Red Neuronal Artificial (RAN)



Fuente: Elaboración Propia.

Back-propagation y sus derivados son los métodos más importantes y difundidos de entrenamiento de redes usados en la actualidad. El término causa mucha confusión pues, estrictamente hablando, *back-propagation* se refiere al método para computar el gradiente de error de una red *feed-forward*. Básicamente el entrenamiento de estas redes consiste en:

a. La pasada hacia adelante (*forward pass*): tanto las salidas como el error en las unidades de salida es calculado.
b. La pasada hacia atrás (*backward pass*): el error de las unidades de salida es usado para alterar los pesos en ellas mismas. Luego el error en los nodos ocultos es calculado (mediante la propagación hacia atrás -*back-propagation*- del error en las unidades de salida mediante los pesos), y los pesos en los nodos ocultos son alterados usando estos valores.

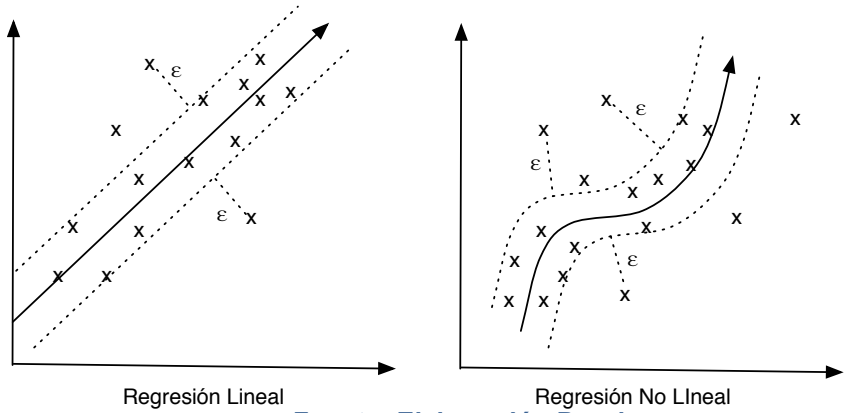
Para cada dato que se desee aprender se ejecutan una pasada hacia adelante y otra hacia atrás. Estos se repiten una y otra vez hasta que el error se encuentre a un nivel suficientemente bajo.

2.6. Modelo de *Support Vector Regression*

Este modelo, también llamado SVRs por sus siglas en inglés, se basa en el concepto de planos de decisiones que definen los límites de decisión; utiliza *Support Vector Machines* (SVM) para realizar una regresión [29,30]. Normalmente los SVM son utilizados como clasificadores, es decir, para separar elementos de una clase de elementos pertenecientes a otra. No obstante, las SVMs pueden emplearse para efectuar regresiones, ocasión en las que se les llama SVR.

El objetivo primordial al usar SVR es buscar y optimizar los límites dados para la regresión. Se fundamenta en el concepto de una función de pérdida que ignora el error, el cual usualmente es colocado a una cierta distancia de las observaciones reales. En la Figura 4 se aprecia un ejemplo de regresión lineal y no lineal usando SVRs, donde se observa que el error es cero para todos los elementos al interior de una banda.

Figura 4: Support Vector Regression (SVR) vs Clasificador



Fuente: Elaboración Propia.

Hay situaciones en que el modelo lineal no es el más adecuado, como en el caso antes mencionado: sería como intentar pasar una recta en la nube de puntos de la derecha, donde claramente el modelo lineal produciría grandes errores (i.e. varios elementos fuera de la banda). Las SVRs permiten, de este modo, pasar un hiperplano que no es lineal para clasificar los elementos de mejor manera.

Para esto, el método consiste en realizar un re-ordenamiento de los elementos del conjunto inicial, utilizando diferentes tipos de funciones, llamadas funciones de Kernel [**30**]. Es relevante destacar que al efectuar esta transformación, los elementos se vuelven linealmente separables en el nuevo espacio.

Debido a que éste es uno de los algoritmos más nuevos y ampliamente usados en el área de pronósticos y regresiones, también será utilizado para los diversos experimentos de esta investigación.

En la construcción de un hiperplano óptimo, una SVR usa un algoritmo iterativo de entrenamiento con el fin de minimizar la función de error [**30**]. De acuerdo a la forma de la función de error, los modelos SVR se pueden clasificar en cuatro grupos distintos [**31**]:

a. Clasificación SVM tipo 1 (conocido como la clasificación C-SVM)
b. Clasificación SVM tipo 2 (conocido como la clasificación de nu-SVM)
c. Regresión SVM tipo 1 (conocido como la regresión epsilon-SVM)
d. Regresión SVM tipo 2 (conocido como la regresión nu-SVM)

2.6.1. Regresión usando SVMs tipo 1

Para ejecutar una regresión los datos deben cumplir la siguiente ecuación:

$$y = f(x) + error$$

Ecuación 6

La tarea consiste, entonces, en encontrar una forma funcional para f que pueda predecir correctamente los nuevos casos que la SVM no ha presentado en la anterior. Esto se puede lograr mediante el entrenamiento del modelo de SVM sobre un conjunto de muestras, es decir, un conjunto de ejercicio, proceso que implica, al igual que la clasificación (véase más arriba), la optimización secuencial de una función de error. Dependiendo de la definición de esta función, dos tipos de modelos SVM se pueden reconocer:

$$error = \frac{1}{2} w^t w + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \xi'_i$$

Ecuación 7

El cual se debe minimizar de acuerdo a varias restricciones:

$$w^t \phi(x_i) + b - y_i \leq \varepsilon + \xi'_i$$

Ecuación 8

$$y - w^t \phi(x_i) - b_i \leq \varepsilon + \xi_i$$

Ecuación 9

$$\xi_i, \xi'_i \geq 0, i = 1, \ldots, N$$

Ecuación 10

2.6.2. Regresión usando SVMs tipo 2

En este caso, también se utiliza la Ecuación 6, pero la expresión para denotar el error del modelo es diferente.

$$error = \frac{1}{2} w^t w - C \left(\nu\varepsilon + \frac{1}{N} \sum_{i=1}^{N} (\xi_i + \xi'_i) \right)$$

Ecuación 11

$$(w^t \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i$$

Ecuación 12

$$y_i - (w^t \phi(x_i) + b) \leq \varepsilon + \xi'_i$$

Ecuación 13

$$\varepsilon, \xi_i, \xi'_i \geq 0, i = 1, \ldots, N$$

Ecuación 14

2.6.3. Funciones de Kernel

Por otra parte, como ya se mencionó, es necesario transformar el espacio y para esto se utilizan las funciones de Kernel que se indican a continuación [**28**]. Vale la pena destacar que, según la literatura, la función RBF es la más usada de todas estas funciones.

$$\phi = \begin{cases} x_i x_j & Lineal \\ (\gamma x_i x_j + coef)^{\alpha} & Polinomial \\ exp\left(-\gamma |x_i - x_j|^2\right) & RBF \\ \tan(\gamma x_i x_j + coef) & Sigmoide \end{cases}$$

Ecuación 15: Funciones de Kernel

2.7. Medición de la Calidad de los Pronósticos

En Estadística la desviación absoluta de un elemento en una colección de datos, corresponde a la diferencia absoluta entre ese elemento y un punto dado de la colección [**15**]. Comúnmente, el punto desde el cual se mide la desviación es un punto que calcula la tendencia central de la muestra, es decir, la mediana, la moda o la media.

La Ecuación 16 muestra la expresión general de la desviación absoluta, donde se tiene un set de n observaciones $\{x_1, x_2, x_3, \dots, x_n\}$ y $m(x)$ es una medida de tendencia central.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m(x)|$$

Ecuación 16

En particular, si se toma la media para el cálculo se estaría hablando de *Mean Absolute Deviation* o MAD, que es una medida comúnmente utilizada para calcular la calidad de un pronóstico en series de tiempo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Ecuación 17

Otra medida relacionada al MAD, es el *Mean Absolute Error* o MAE, que indica la diferencia absoluta entre el valor real y el valor pronosticado. Este indicador se enfoca mucho más en la salida de los modelos de pronóstico. La expresión del MAE se encuentra en la

Ecuación 17 y la misma es posible escribirla en términos porcentuales, lo cual da origen al MAPE o *Mean Absolute Percentage Error*, cuya expresión se encuentra en la

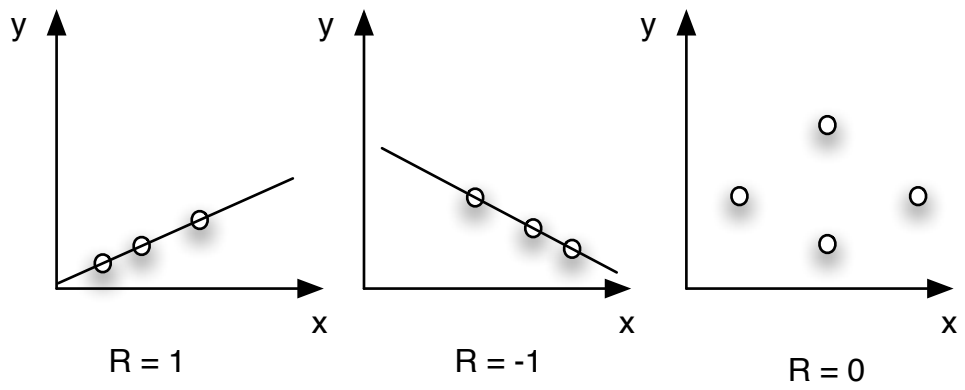
Ecuación 18.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{f_t - y_t}{f_t}\right|$$

Ecuación 18

Otro indicador comúnmente utilizado en modelos regresivos, es el factor de correlación. Éste mide la correlación existente entre dos variables, su valor varía entre -1 y 1. Un factor de correlación 1 indica que las variables están perfectamente correlacionadas; un valor -1 indica que las variables están inversamente correlacionadas de manera perfecta; y finalmente, un valor 0 indica que las variables no están correlacionadas. Un ejemplo se aprecia en la Figura 5.

Figura 5: Factor de Correlación Lineal



Fuente: Elaboración propia.

Claramente, el factor de correlación lineal no es un indicador de la calidad del pronóstico generado por un modelo. Sin embargo, y dado que normalmente las herramientas para generar pronósticos lo entregan (en el caso de las correlaciones), se guardará para mostrar más adelante la correlación de las variables.

Finalmente, en este trabajo se empleará el MAPE en la medición de la calidad global de los pronósticos calculados, dado que es posible estimar este indicador para todos los métodos utilizados; a diferencia del factor de correlación lineal que no se puede obtener, por ejemplo, para los promedios móviles.

Capítulo 3: Situación actual y recopilación metódica de datos

3.1. Información general

3.1.1. Descripción del Transantiago.

El nuevo sistema de transporte público urbano para la ciudad de Santiago de Chile es Transantiago. La primera etapa de este proyecto comenzó su operación el 22 de octubre de 2005, la cual finalizó el 10 de febrero de 2007, dando paso a la transición definitiva al nuevo sistema. El objetivo principal de su implementación era cambiar por completo la organización del transporte colectivo existente en la ciudad [**32**].

Transantiago en conjunto con Metro S.A. realizaron un rediseño completo del sistema de transporte público, basado en el uso de alimentadores y troncales, cambiando por completo los recorridos de las antiguas micros y estableciendo a Metro como el eje troncal de dicho cambio.

La puesta en marcha de Transantiago generó una serie de problemas, revelando importantes deficiencias y errores tanto en su diseño como en la implementación del proyecto. Esto generó una grave crisis a nivel metropolitano, social y político [**32**].

3.1.2. Descripción de la empresa

Metro S.A. es el ferrocarril más moderno de Sudamérica [**33**], es parte fundamental del sistema de transportes de la ciudad de Santiago abarcando casi la totalidad de la urbe. En el momento de levantamiento de información para este trabajo, la empresa contaba con 5 líneas de operación, 92 estaciones y una extensión de 84 km., diariamente recibía en sus instalaciones cerca de 1.300.000 pasajeros -dato válido hasta el año 2006-, y a partir de la puesta en marcha del nuevo sistema de transportes, esta cifra llegó a ser el doble entre los años 2007 y 2009, cuyo máximo llegó el 30 de mayo de 2008, recibiendo un total de 2.504.089 pasajeros. [**34**].

Durante el proceso de análisis de este proyecto se pusieron en marcha 16 nuevas estaciones que extendieron la red hacia el oriente, hasta la comuna de Las Condes, y al poniente hasta Maipú. Estos trabajos finalizaron durante el año 2009 y 2010, respectivamente, dejando al Metro con una extensión de 126 km y abarcando a 21 de las 36 comunas de Santiago [**32**].

Para conocer un poco más a cerca de la Empresa, a continuación se mencionan algunos hitos importantes de su historia y desarrollo. Metro S.A. comienza sus operaciones el 15 de septiembre de 1975 con el funcionamiento de su primera línea, cuyo tramo unía de manera subterránea las comunas de Lo Prado (estación San Pablo) y el centro de Santiago (estación La Moneda); posteriormente, dicha línea fue extendida hasta la comuna de Las Condes (estación Escuela Militar).

Asimismo, el año 1978 se inauguró la nueva Línea 2 en la red, tramo que unía las comunas de Santiago Centro (estación Los Héroes) con San Miguel (estación Franklin). En diciembre de ese mismo año, la línea se extendió hacia el sur a lo largo de Gran Avenida hasta Lo Ovalle [**33**]. Por otro lado, el año 1987 se dio origen a dos nuevas estaciones hacia el norte en la Línea 2: Santa Ana y Mapocho.

Posteriormente, el año 1997 se inauguró la Línea 5 con una extensión de 10,3 km. uniendo las comunas de Providencia (estación Parque Bustamante) y La Florida (estación Bellavista La Florida), luego se extendió dicha línea articulando las comunas de Providencia (estación Baquedano) con Santiago Centro (estación Santa Ana) [**33**].

Con la llegada de Ricardo Lagos a la Presidencia el año 2000, se diseñó la extensión de la Línea 5 al poniente hasta Quinta Normal -siguiendo el eje de calle Catedral-, y de la Línea 2 por el norte y el sur para unir ambos extremos de la Circunvalación Américo Vespucio [**35**].

El 22 de diciembre de 2006, se concretaron todas las extensiones hechas a la Línea 2 propuestas por el gobierno del presidente Lagos, uniendo las comunas de Recoleta (Estación Américo Vespucio Norte) con la Cisterna (estación La Cisterna).

Dentro de los últimos proyectos de ampliación del Metro, se encuentra la construcción de la Línea 4, la cual fue inaugurada en su totalidad en marzo del 2006, con 24,7 kilómetros y 22 estaciones que unen las comunas de Providencia, Las Condes, Ñuñoa, La Reina, Peñalolén, Macul, La Florida y Puente Alto. Finalmente, la Línea 4 se complementó con la inauguración de un ramal, la Línea 4A, que desde el 16 de agosto de 2006 conecta las Líneas 2 y 4 [**33**].

En resumen, las estaciones que conforman la red de Metro S.A. (ver Figura 6) están divididas en 5 líneas de funcionamiento [**33**]:

- **Línea 1.** Eje principal de la Red, recorre la ciudad de poniente a oriente, se identifica su recorrido con el color rojo y tiene una duración aproximada de viaje de 29,7 minutos por todo el tramo. Tiene una extensión de 16 Km. y 24 estaciones de servicio, atravesando las comunas de Lo Prado, Estación Central, Santiago, Providencia y Las Condes. Esta línea concentra cerca del 50,5% del total de los viajes [**1**].

- **Línea 2.** Recorre la ciudad en sentido longitudinal de norte a sur, se identifica su recorrido con el color amarillo y tiene una duración aproximada de viaje de 34,2 minutos a lo largo del tramo. Su extensión es de 20,6 Km. y 22 estaciones de servicio, atravesando las comunas de Recoleta, Santiago, San Miguel y La Cisterna. Esta línea concentra cerca del 17,8% de los viajes realizados en Metro [**1**].

- **Línea 4.** Se identifica con el color azul marino y tiene una duración aproximada de viaje de 40 minutos a lo largo del tramo. Tiene una extensión de 24,7 Km. y 22 estaciones de servicio, atravesando las comunas de Providencia, Las Condes, La Reina, Ñuñoa, Peñalolén, Macul, La Florida y Puente Alto.

- **Línea 4.** Es un ramal de la Línea 4, se identifica con el color celeste y tiene una duración aproximada de viaje de 12 minutos a lo largo del tramo. Tiene una extensión de 7,7 Km. y 6 estaciones de servicio, atravesando las comunas de La Florida, La Granja, San Ramón y La Cisterna.

- **Línea 5.** Se identifica con el color verde, su extensión es de 15,5 kilómetros y 18 estaciones de servicio, atravesando las comunas de Santiago, Providencia, Ñuñoa, Macul, San Joaquín y La Florida. Esta línea concentra cerca del 16% de los viajes totales de la red [**1**].

Por otro lado, la empresa cuenta con estaciones de intercambio modal, como son Quinta Normal, Vespucio Norte y La Cisterna, para combinar el servicio de ferrocarriles con otros medios de transporte. Además, con la inauguración de la Estación El Sol en la Línea 5 (extensión a Maipú), se construyó una nueva estación de intercambio modal.

Figura 6: Red Metro de Santiago (2009).

Fuente: www.metro.cl.

Metro S.A., a partir del 10 de febrero de 2007, se constituye en el articulador del nuevo sistema de transporte de la ciudad de Santiago, lo cual trajo consigo que la demanda se duplicara a partir de esa fecha causando grandes aglomeraciones, alcanzando una ocupación del servicio de más de 2,4 millones de pasajeros diarios, llegando a tener 6,4 pasajeros por metro cuadrado en los trenes.

Por estas razones, se debió efectuar mejoras en su infraestructura, como la compra de 11 nuevos trenes asignados a las distintas líneas, cambios en el horario de servicio, establecimiento de servicios expresos en las Líneas 4 y 5 durante las horas de mayor demanda y acondicionamiento de los trenes con los que ya contaba la empresa. [**33**].

Durante el proceso de este trabajo, Metro inauguró dos grandes proyectos anunciados durante la presidencia de Ricardo Lagos el 2005. La primera es, la extensión de la Línea 1 hacia el oriente, desde Escuela Militar hasta la estación Los Domínicos en la comuna de Las Condes, agregando tres nuevas estaciones y 4 kilómetros a la red. La segunda es la extensión de la red hacia el poniente, conectando las comunas de Maipú, Pudahuel, Lo Prado y Quinta Normal, agregando 11 estaciones al sistema. Ambas extensiones se pusieron en funcionamiento a fines del año 2010, fecha en la cual también se inauguró la Estación San José de la Estrella de la Línea 4. En la figura 7 se puede apreciar la extensión de la red de Metro prevista para el año 2010.

Figura 7: Extensión de la Red de Metro de Santiago (2010).

Fuente: www.metro.cl.

En cuanto a futuros proyectos, como expansiones o líneas nuevas, la empresa anunció, una vez evaluado el funcionamiento de Transantiago y su estabilización, la construcción de dos nuevas lineas la linea 3, que busca conectar la Reina y Huechuraba; y la linea 6, conectando Pedro Aguirre Cerda con Providencia. Añadiendo 28 estaciones a la actual Red.

Sin embargo, debido a los problemas de congestión existentes en la Línea 1 desde la implementación de Transantiago, la prioridad siempre es la construcción de una alternativa a ella. Según el presidente de Metro S.A., Clemente Pérez, las alternativas podrían ir a lo largo de Avenida Matta (similar a la Línea 3) o Avenida Santa María [**36**].

3.2. Obtención de afluencia de pasajeros en la actualidad

En la actualidad, en Metro la afluencia de pasajeros es obtenida por un sistema de contadores diseñado especialmente para la empresa, los cuales están instalados en cada uno de los torniquetes de paso hacia los andenes de las distintas estaciones. Los datos obtenidos a través de este sistema son almacenados por minuto en la base de datos de Metro.

Cabe recalcar que el concepto 'afluencia de pasajeros' no es lo mismo que 'demanda de pasajeros'. La diferenciación establecida por la empresa para estos términos es la siguiente: la primera es la cantidad de personas que pasa por torniquete, es decir, introduciendo su boleto o pasando su tarjeta bip!, en cambio, la demanda de pasajeros es la cantidad de personas que se encuentran en el andén de Metro, es decir, las personas que pasan por torniquete, más las personas que realizan trasbordos, infiltrados, etc.

La presente investigación está destinada a pronosticar en corto plazo la afluencia de pasajeros, variable que corresponde a la parte cuantificable de la demanda de Metro, ya que los infiltrados y la cantidad de personas que efectúan transbordos hoy no pueden ser medidos, en este sentido, sólo existen aproximaciones realizadas por la empresa con la ayuda de datos como el peso de los trenes, realizado con un sistema llamado *Railweight Automatic Passenger Counting System For Metro,* y encuestas.

3.3. Datos históricos

Las principales fuentes de datos utilizadas para realizar este estudio son:

- Base de datos Metro de Santiago.
- Memorias Anuales de Metro de Santiago.
- Personal de la Subgerencia de Ingeniería de Operaciones.

La principal fuente de información otorgada por la empresa es la afluencia de pasajeros; los datos están divididos por día y cada 15 minutos a contar del año 2002 en cuatro estaciones de la Red, las cuales -como se mencionó en la sección de Alcances- fueron escogidas como representativas para realizar el análisis e identificar los patrones de comportamiento de la demanda en cada una de ellas, así como las proyecciones a corto plazo.

En la tabla 2, expuesta a continuación, se muestra un resumen del total de pasajeros que acudieron mensualmente al servicio de Metro, desde el año 2002 hasta Junio de 2009, en la estación San Pablo:

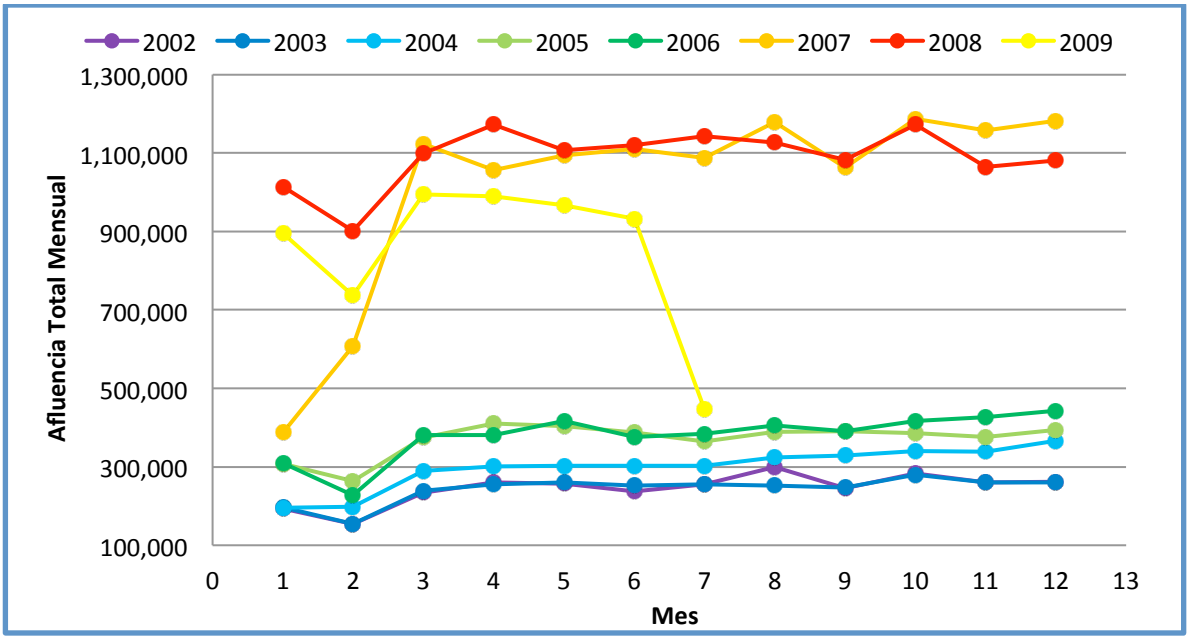
Tabla 2: Afluencia total mensual Estación San Pablo.

AFLUENCIA TOTAL MENSUAL SAN PABLO								
	2002	2003	2004	2005	2006	2007	2008	2009
ENERO	192.655	195.900	195.614	307.745	309.315	387.533	1.014.184	894.863
FEBRERO	154.196	154.532	197.570	263.637	227.076	607.941	900.733	736.350
MARZO	233.621	239.082	290.229	374.409	381.329	1.122.584	1.098.210	995.156
ABRIL	260.445	255.995	300.888	410.452	381.291	1.056.553	1.172.782	989.664
MAYO	257.647	260.251	303.439	404.251	416.252	1.094.595	1.106.551	966.888
JUNIO	237.829	251.726	303.427	387.547	376.714	1.110.577	1.120.732	932.916
JULIO	255.103	255.174	303.203	363.961	384.980	1.087.349	1.142.641	447.549
AGOSTO	298.621	252.663	324.015	389.045	405.374	1.178.083	1.126.713	
SEPTIEMBRE	245.960	247.180	329.269	391.274	390.507	1.062.899	1.082.064	
OCTUBRE	282.832	280.359	340.395	385.170	416.766	1.186.887	1.172.227	
NOVIEMBRE	260.414	259.735	338.608	376.707	426.063	1.156.880	1.064.986	
DICIEMBRE	262.277	259.998	366.088	393.855	442.193	1.181.912	1.080.545	
Total Anual	**2.941.600**	**2.912.595**	**3.592.745**	**4.448.053**	**4.557.860**	**12.233.793**	**13.082.368**	**5.963.386**

Fuente: Metro S.A.

En la figura 8 se puede observar los cambios en la afluencia de pasajeros en el tiempo y la implementación del Transantiago en la estación San Pablo.

Figura 8: Afluencia de pasajeros mensual Estación San Pablo.



Fuente: Elaboración propia.

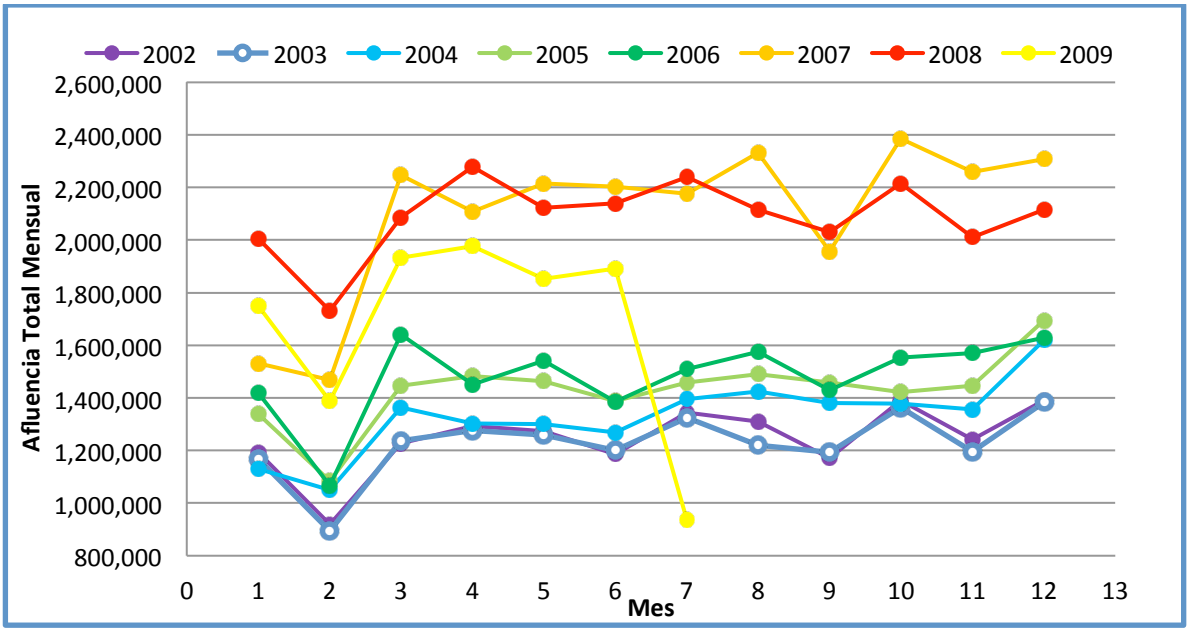
En la tabla 3 y figura 9 expuestas a continuación, se muestra un resumen del total de pasajeros en la estación Universidad de Chile:

Tabla 3: Afluencia total mensual Estación Universidad de Chile.

AFLUENCIA TOTAL MENSUAL UNIVERSIDAD DE CHILE								
	2002	2003	2004	2005	2006	2007	2008	2009
ENERO	1.192.535	1.167.970	1.129.512	1.338.395	1.419.323	1.530.667	2.005.269	1.750.740
FEBRERO	919.333	894.262	1.049.888	1.083.989	1.065.901	1.467.013	1.731.031	1.387.489
MARZO	1.228.351	1.235.873	1.362.665	1.447.199	1.641.045	2.247.461	2.085.982	1.933.451
ABRIL	1.293.321	1.274.734	1.302.310	1.482.110	1.451.105	2.107.312	2.277.005	1.977.316
MAYO	1.273.868	1.258.805	1.301.187	1.462.995	1.541.422	2.215.342	2.121.746	1.853.095
JUNIO	1.186.753	1.199.610	1.269.463	1.386.975	1.385.929	2.201.587	2.140.250	1.890.171
JULIO	1.343.028	1.325.433	1.394.625	1.459.407	1.508.786	2.176.366	2.239.592	937.455
AGOSTO	1.309.881	1.219.724	1.423.297	1.489.784	1.574.482	2.331.075	2.114.938	
SEPTIEMBRE	1.171.013	1.193.073	1.381.647	1.459.505	1.429.344	1.957.648	2.029.270	
OCTUBRE	1.390.511	1.363.489	1.378.211	1.421.013	1.552.978	2.384.932	2.214.809	
NOVIEMBRE	1.240.846	1.196.319	1.356.689	1.444.999	1.569.826	2.259.079	2.010.178	
DICIEMBRE	1.390.241	1.382.544	1.618.992	1.693.253	1.629.696	2.307.242	2.114.297	
Total Anual	**14.939.681**	**14.711.836**	**15.968.486**	**17.169.624**	**17.769.837**	**25.185.724**	**25.084.367**	**11.729.717**

Fuente: Metro S.A.

Figura 9: Afluencia de pasajeros mensual Estación Universidad de Chile.



Fuente: Elaboración propia.

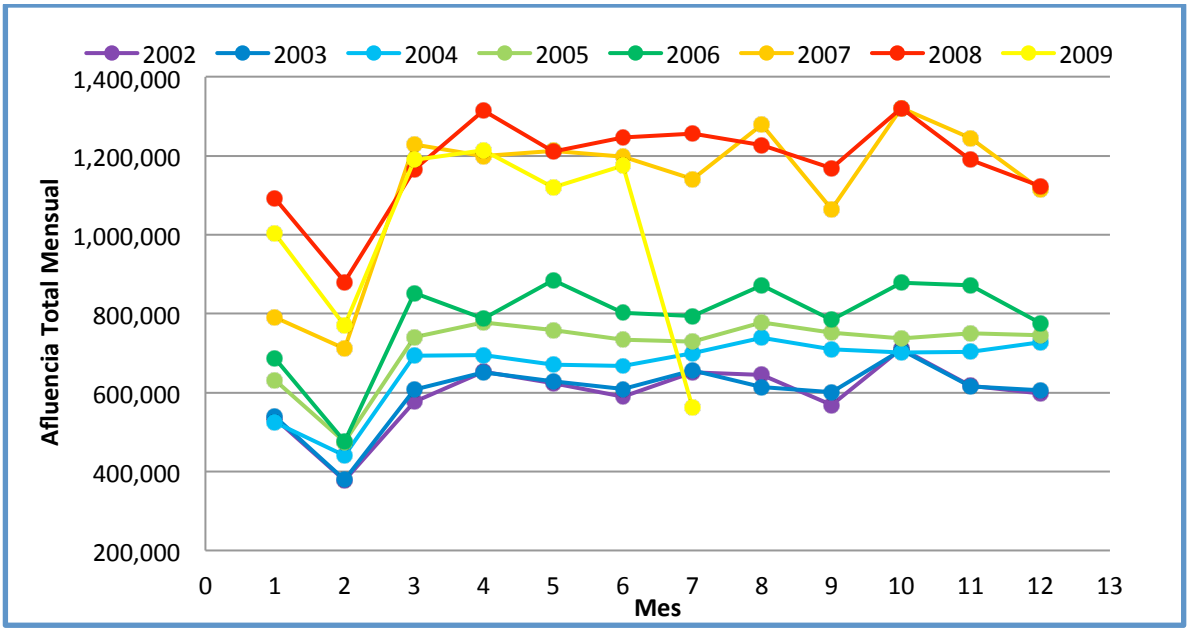
En la tabla 4 y figura 7, expuestas a continuación, se muestra un resumen del total de pasajeros, en la estación Pedro de Valdivia:

Tabla 4: Afluencia total mensual Estación Pedro de Valdivia.

AFLUENCIA TOTAL MENSUAL PEDRO DE VALDIVIA								
	2002	2003	2004	2005	2006	2007	2008	2009
ENERO	534.867	539.204	524.032	631.319	687.431	791.408	1.091.868	1.003.005
FEBRERO	377.572	379.087	440.511	475.174	475.465	710.789	880.721	769.191
MARZO	577.350	607.409	693.276	740.627	852.879	1.228.453	1.165.010	1.190.295
ABRIL	653.321	651.706	695.861	778.239	787.963	1.199.574	1.314.499	1.213.571
MAYO	623.556	628.254	671.214	758.254	884.235	1.212.238	1.209.995	1.119.845
JUNIO	590.058	609.099	667.640	733.506	802.600	1.198.026	1.246.444	1.174.429
JULIO	652.029	656.582	699.608	729.314	794.517	1.141.389	1.256.379	562.690
AGOSTO	645.482	614.004	738.270	778.445	871.188	1.278.982	1.226.551	
SEPTIEMBRE	568.868	601.468	709.664	751.218	784.826	1.064.276	1.169.130	
OCTUBRE	710.613	707.773	702.188	736.939	878.522	1.321.369	1.320.413	
NOVIEMBRE	617.395	616.218	703.048	750.749	872.347	1.244.951	1.190.161	
DICIEMBRE	597.472	605.973	727.609	745.117	776.533	1.115.338	1.123.033	
Total Anual	**7.148.583**	**7.216.777**	**7.972.921**	**8.608.901**	**9.468.506**	**13.506.793**	**14.194.204**	**7.033.026**

Fuente: Metro S.A.

Figura 10: Afluencia de pasajeros mensual Estación Pedro de Valdivia.



Fuente: Elaboración propia.

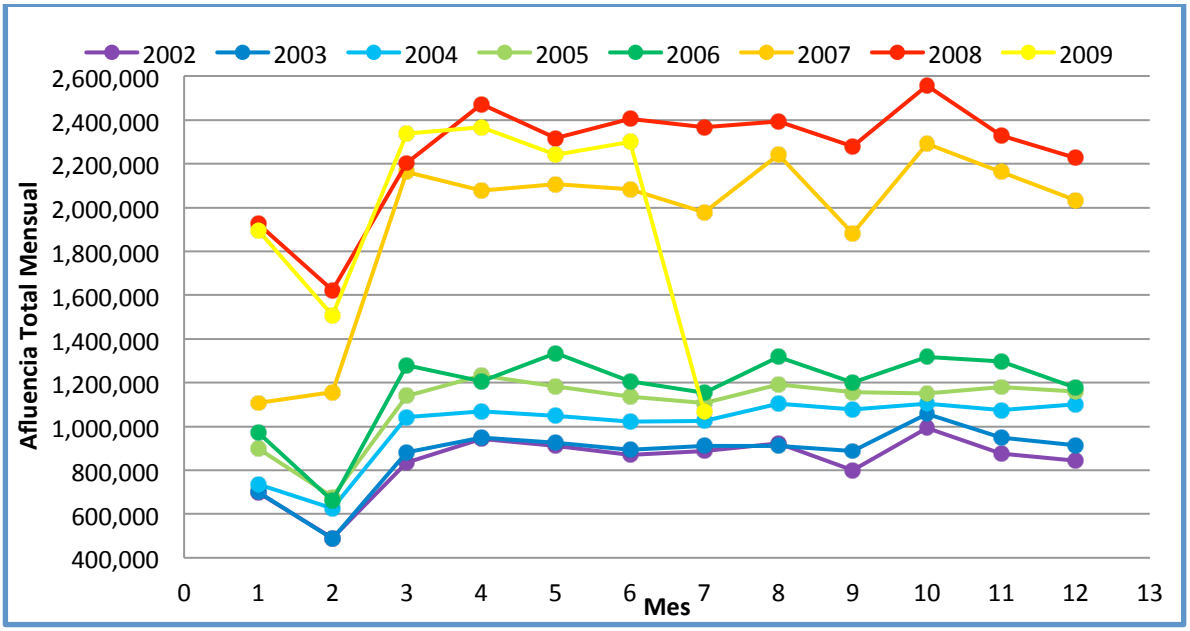
En la tabla 5 y figura 11 expuestas a continuación, se muestra un resumen del total de pasajeros en la estación Escuela Militar:

Tabla 5: Afluencia total mensual Estación Escuela Militar.

	AFLUENCIA TOTAL MENSUAL PEDRO DE VALDIVIA							
	2002	2003	2004	2005	2006	2007	2008	2009
ENERO	698.690	701.444	734.544	897.912	971.994	1.107.702	1.924.647	1.897.292
FEBRERO	489.147	487.232	625.768	678.502	660.675	1.156.407	1.622.254	1.508.864
MARZO	834.808	881.308	1.041.985	1.140.168	1.278.599	2.163.791	2.204.048	2.339.045
ABRIL	942.698	948.976	1.068.335	1.231.354	1.206.508	2.076.773	2.471.907	2.367.262
MAYO	911.163	926.013	1.048.528	1.182.184	1.336.278	2.108.133	2.316.855	2.240.277
JUNIO	871.451	894.163	1.022.737	1.135.623	1.205.768	2.084.729	2.405.272	2.299.822
JULIO	888.611	912.314	1.025.807	1.105.872	1.152.969	1.978.796	2.366.748	1.066.177
AGOSTO	922.529	911.947	1.105.009	1.192.025	1.316.073	2.242.088	2.394.172	
SEPTIEMBRE	799.476	888.573	1.077.305	1.156.804	1.200.578	1.882.053	2.279.897	
OCTUBRE	992.077	1.056.121	1.104.079	1.151.351	1.316.996	2.292.059	2.555.693	
NOVIEMBRE	877.268	947.710	1.074.210	1.178.733	1.297.449	2.161.974	2.328.289	
DICIEMBRE	843.023	915.323	1.101.974	1.160.008	1.178.186	2.034.597	2.226.545	
Total Anual	**10.070.941**	**10.471.124**	**12.030.281**	**13.210.536**	**14.122.073**	**23.289.102**	**27.096.327**	**13.718.739**

Fuente: Metro S.A.

Figura 11: Afluencia de pasajeros mensual Estación Escuela Militar.



Fuente: Elaboración propia.

Por otro lado, la empresa hace una división de la demanda por tipo de día y horario, es decir, divide los días del mes en laborales, sábados y domingos o feriados; y los horarios de días laborales en horario punta (de 07.00 a 10.00 y de 19.00 a 21.00) y horario valle, el resto del día.

Capítulo 4: Aplicación y discusión de la utilización de pronósticos

Este apartado revisa en detalle la aplicación de los diversos modelos de pronósticos en series de tiempo revisados en el Capítulo 2, buscando encontrar el mejor modelo para pronosticar la afluencia de pasajeros de Metro en las cuatro estaciones representativas seleccionadas para el estudio: Escuela Militar, Pedro de Valdivia, Universidad de Chile y San Pablo.

Se analizó la afluencia agregada mensual para las cuatro estaciones de metro seleccionadas. En primera instancia, se utilizó modelos autorregresivos, series de promedios móviles y Modelos de Suavización Exponencial, herramientas con las cuales es posible establecer el mejor modelo para comprender la tendencia en los datos globales de afluencia por estación. En dicho estudio fueron empleados nueve modelos distintos: Regresión Lineal, Regresión Exponencial, Regresión Logarítmica, *Power Regression*, Regresión Polinomial (de 2^{do} y 6^{to} órdenes) y Promedios Móviles[5], Suavización Exponencial, *Support Vector Regression* y Redes Neuronales.

En el estudio se utilizaron todos los datos históricos, vale decir, enero 2002 a julio 2009, es decir, 91 observaciones de la afluencia mensual. Además, la bondad del modelo se puede medir usando la medida del error promedio, o sea, el valor R^2 y MAPE (*Mean Absolute Percentual Error*) (ver Capítulo 2).

A continuación, se expone el análisis para la estación Universidad de Chile, en el caso de las otras estaciones se entregarán sólo los mejores modelos obtenidos, sin embargo, el detalle para cada una de ellas se encuentra en el Anexo A.

4.1. Estudio Inicial para la Aplicación de Modelos Regresivos

Una vez obtenidos los datos, es necesario efectuar un análisis previo para generar los modelos finales. Es así como se realizó un estudio inicial con el fin de entender la factibilidad de la aplicación de los diversos modelos explicados *in extenso* en el Capítulo 2.

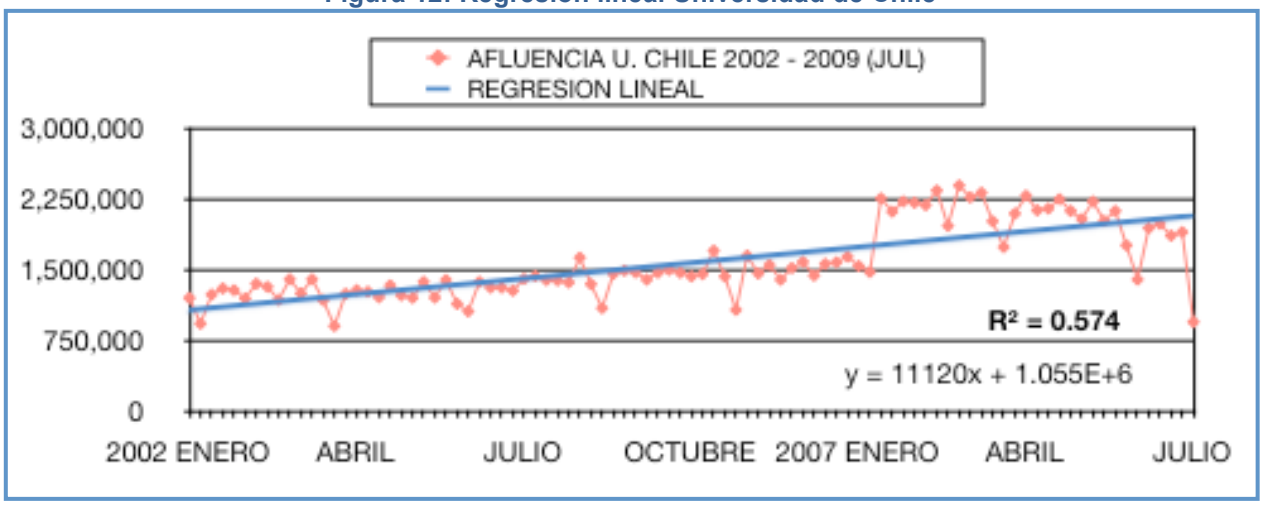
Estos modelos fueron aplicados sobre los datos de afluencias mensuales sin ningún tipo de corrección. Una vez obtenido el modelo se graficó el comportamiento real de la afluencia (en rojo) y simultáneamente se diseñó el modelo de regresión en particular (de la Figura 12 a la Figura 17).

Además, se calculó el factor de correlación R^2 para establecer la calidad del modelo obtenido. De manera simple, si el valor de R^2 se acerca a 1, entonces el modelo se ajusta de manera casi perfecta a las observaciones; por el contrario, si el factor es cercano a 0, entonces el modelo no se ajusta de buena manera a los datos.

Un factor R^2 igual o superior a 0.8 da cuenta de un modelo con un grado de calidad adecuado para los fines planteados. Sin embargo, como puede apreciarse, en todos los modelos obtenidos el factor de correlación es inferior a esta cifra.

[5]Con ventanas de 2, 3, 4, 5, 6, 7 y 8 meses de periodo móvil, aunque sólo se mostrarán los gráficos correspondientes a ventanas de 2, 4, 6 y 8. Todos los gráficos se encuentran en el Anexo A de este trabajo.

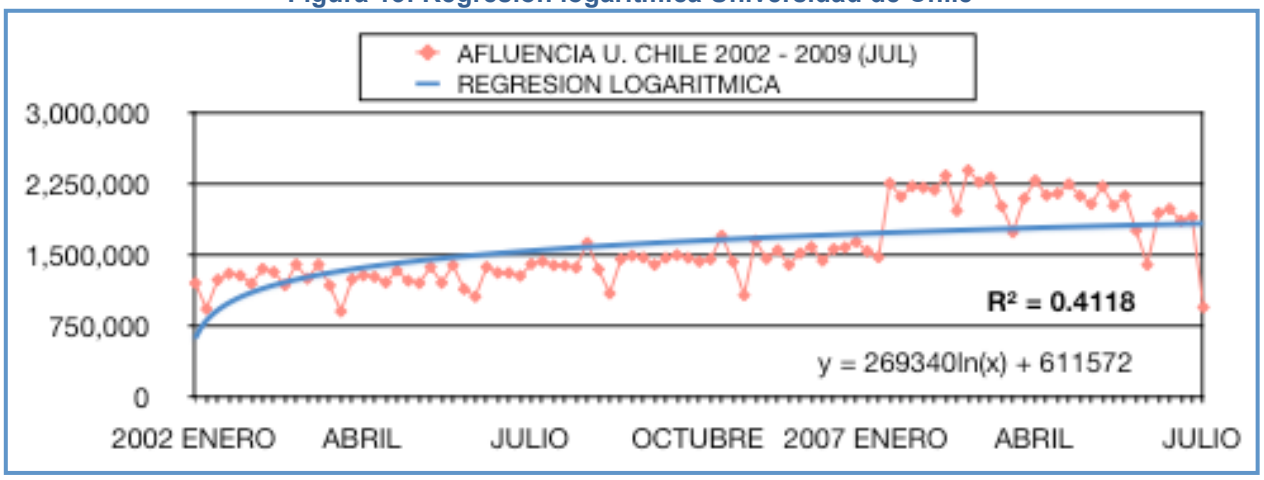
Figura 12: Regresión lineal Universidad de Chile



Fuente: Elaboración propia.

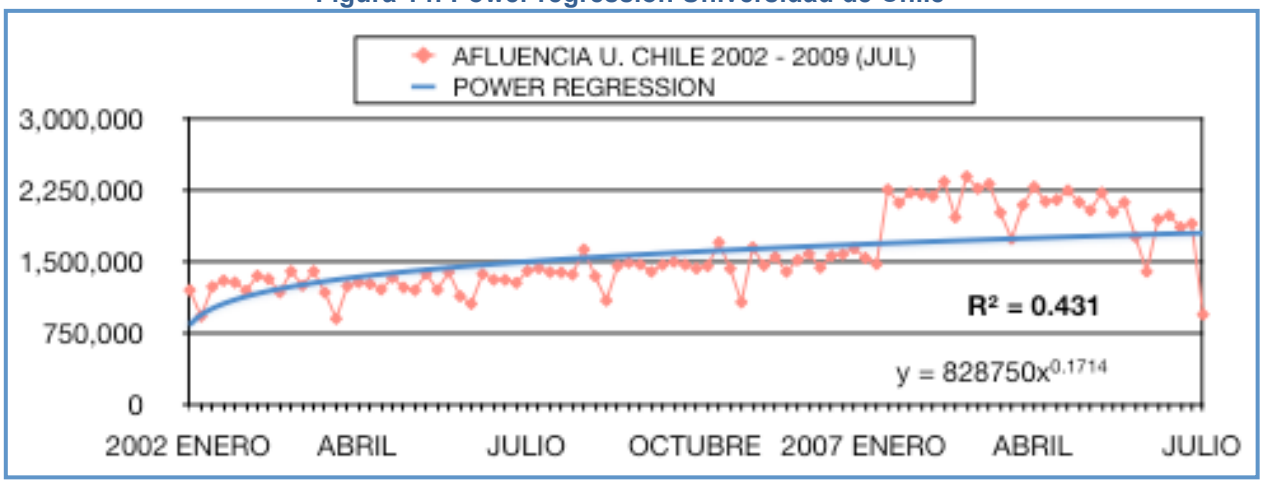
El mejor modelo regresivo obtenido es el de Regresión Polinomial de 6to orden (ver Figura 17), el cual obtuvo un factor de correlación R^2=0.78. Sin embargo, este valor de ajuste no es suficiente para afirmar que el modelo es el adecuado para predecir la afluencia de pasajeros.

Figura 13: Regresión logarítmica Universidad de Chile



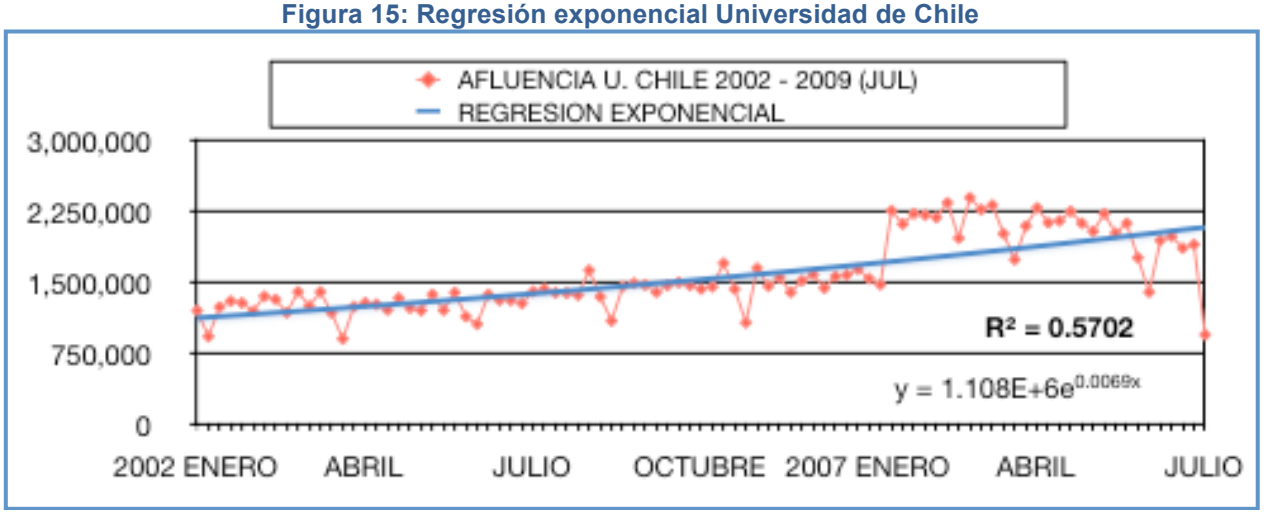
Fuente: Elaboración propia.

Figura 14: Power regression Universidad de Chile



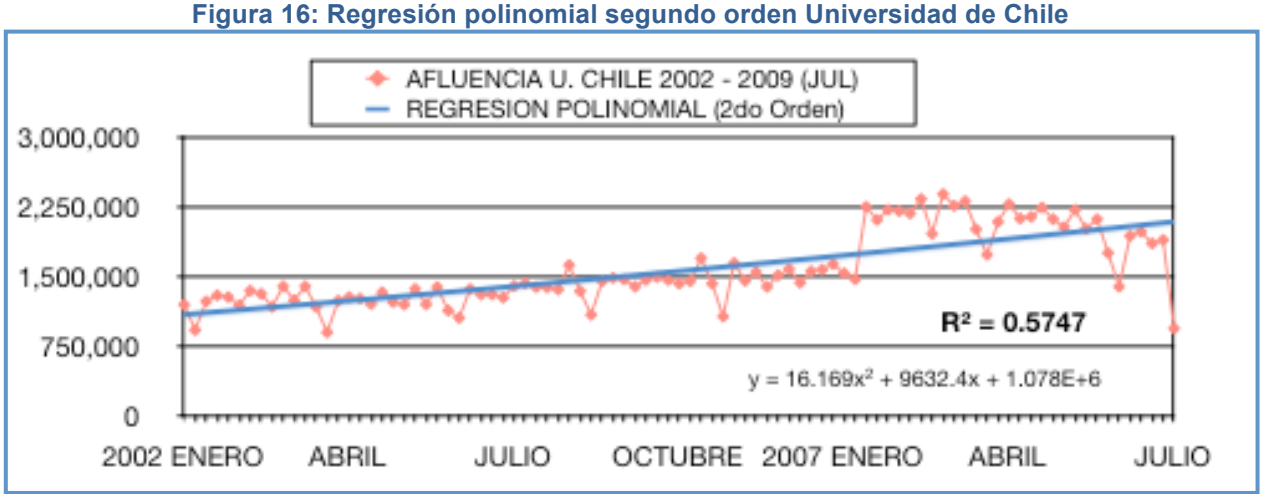
Fuente: Elaboración propia.

Figura 15: Regresión exponencial Universidad de Chile



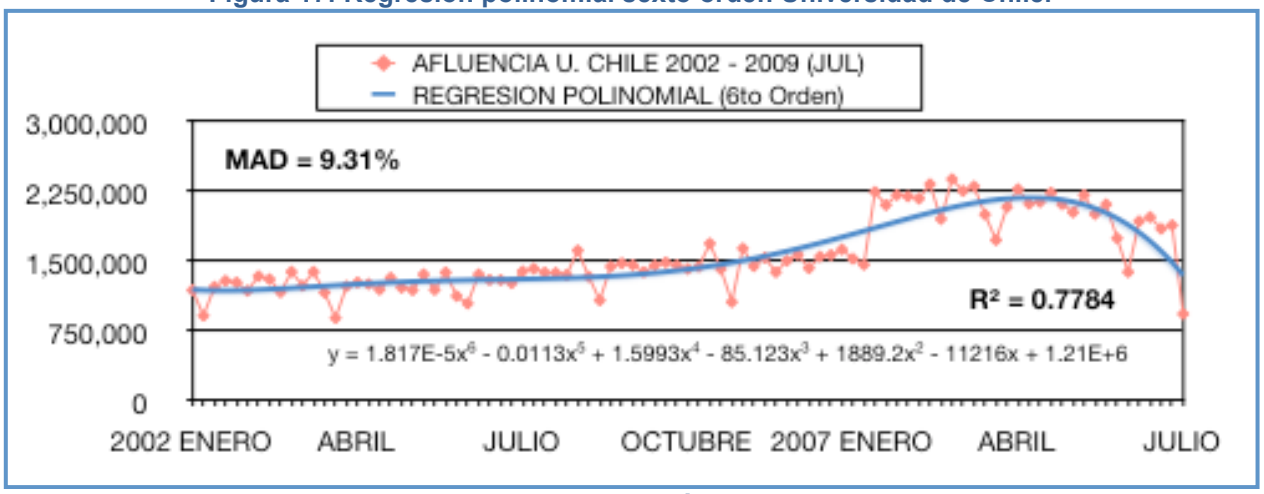
Fuente: Elaboración propia.

Figura 16: Regresión polinomial segundo orden Universidad de Chile



Fuente: Elaboración propia.

Figura 17: Regresión polinomial sexto orden Universidad de Chile.



Fuente: Elaboración propia.

Por otra parte, sólo con el factor de correlación no es posible afirmar con certeza cuál es el poder predictivo de los modelos regresivos calculados. Para esto es necesario calcular el error medio al realizar un pronóstico sobre un horizonte dado de tiempo. Es decir, se debe calcular el MAPE para un periodo T que ha establecido un horizonte de 6 meses para los modelos de regresión. En otras palabras, se utilizó 85 observaciones (de un universo de 91 observaciones) para calcular un modelo, y luego éste fue aplicado para obtener el valor de la afluencia en el periodo 86. Con este valor se estimó un error porcentual del pronóstico en dicho periodo. Se realizó lo mismo para los períodos 87, 88, 89, 90 y 91. Finalmente, el promedio de los errores de estos pronósticos dan origen al MAPE para cada método (ver Tabla 6).

Tabla 6: MAPE Algortimos de Regresión

	Lineal	Logarítmica	Power	Exponencial	Polinomial (2do)	Polinomial (6to)
San Pablo	32.62%	26.48%	28.49%	41.60%	76.31%	63.91%
U. de Chile	32.00%	24.15%	24.32%	32.14%	33.19%	23.07%
P. de Valdivia	28.41%	27.83%	27.96%	29.32%	41.82%	29.73%
Escuela Militar	28.86%	29.54%	29.70%	28.89%	57.70%	56.90%

Fuente: Elaboración propia.

De la Tabla 6 se infiere que no existe un modelo universal que permita predecir en todos los casos la afluencia de pasajeros. Es decir, para cada estación existe uno que funciona mejor que otros. Por ejemplo, para la Estación Universidad de Chile el mejor modelo es el Polinomial de 6to grado, pues el que presenta menor error medio al predecir (23,07%), le sigue la Regresión Logarítmica (con 24,15% de MAPE). A diferencia de la estación Escuela Militar, donde el mejor es la Regresión Lineal y el peor la Regresión Polinomial de 6to orden con un error de medio de 56,9%.

Asimismo, se calcularon pronósticos usando promedios móviles con una ventana de tiempo de 2, 3, 4, 5, 6, 7 y 8 meses. Los resultados del MAPE para cada uno de estos se pueden apreciar en forma resumida en la Tabla 7: Resumen MAPE en Promedios Móviles en pronósticos de afluencias mensuales.

Tabla 7: Resumen MAPE en Promedios Móviles en pronósticos de afluencias mensuales

	Pedro de Valdivia	Universidad de Chile	Escuela Militar	San Pablo
Moving Average (8)	11.38%	8.48%	13.13%	12.76%
Moving Average (7)	11.37%	8.44%	12.91%	12.12%
Moving Average (6)	10.79%	7.95%	12.15%	11.29%
Moving Average (5)	10.57%	7.84%	11.89%	10.72%
Moving Average (4)	9.66%	7.76%	10.87%	9.79%
Moving Average (3)	8.45%	7.18%	9.12%	8.39%
Moving Average (2)	**6.29%**	**5.74%**	**6.56%**	**5.71%**

Fuente: Elaboración propia.

Del mismo modo, posteriormente se utilizó *rapid miner* para testear otros modelos más avanzados, como las Redes Neuronales y las SVRs. Los resultados de la aplicación de estos modelos se aprecian en la Tabla 8.

Tabla 8: Resumen MAPE en Suavización Exponencial, Support Vector Regression y Redes Neuronales para pronóstico de afluencias mensuales

	Exp. Smoothing	SVR	Neural Net
Pedro de Valdivia	**12.31%**	15.29%	13.83%
Universidad de Chile	**9.93%**	15.13%	11.31%
Escuela Militar	**13.04%**	15.76%	14.75%
San Pablo	**11.40%**	15.92%	13.95%

Fuente: Elaboración propia.

El cálculo de la Suavización Exponencial se realizó utilizando Excel y luego aplicando la herramienta *Solver* para encontrar el alfa que permite minimizar el MAPE. Estos resultados también se pueden ver en la Tabla 8.

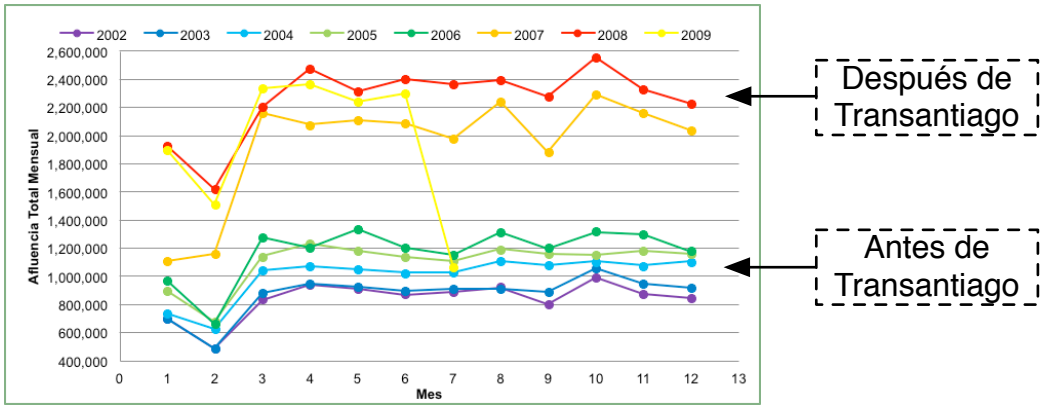
4.2. Pre procesamiento y corrección de datos

De la sección anterior se infiere que los Modelos de Promedios Móviles presentan buenos resultados y, por lo tanto, es totalmente factible su aplicación. Sin embargo, los Modelos Regresivos, las Redes Neuronales y la SVR no entregan tan buenos resultados.

Normalmente una serie de tiempo puede presentar tendencias, temporalidades o ciclos, y otros efectos que deben ser removidos antes de realizar una aplicación exitosa de éstos. En la sección anterior no se aplicó ningún ajuste, ni se eliminó de estos elementos de la serie. Por lo tanto, los resultados deberían mejorar al removerlos.

En particular, se puede ver que antes la puesta en marcha de Transantiago la afluencia promedio por año era alrededor de 2.2 veces menor que luego de iniciado el nuevo sistema de transporte.

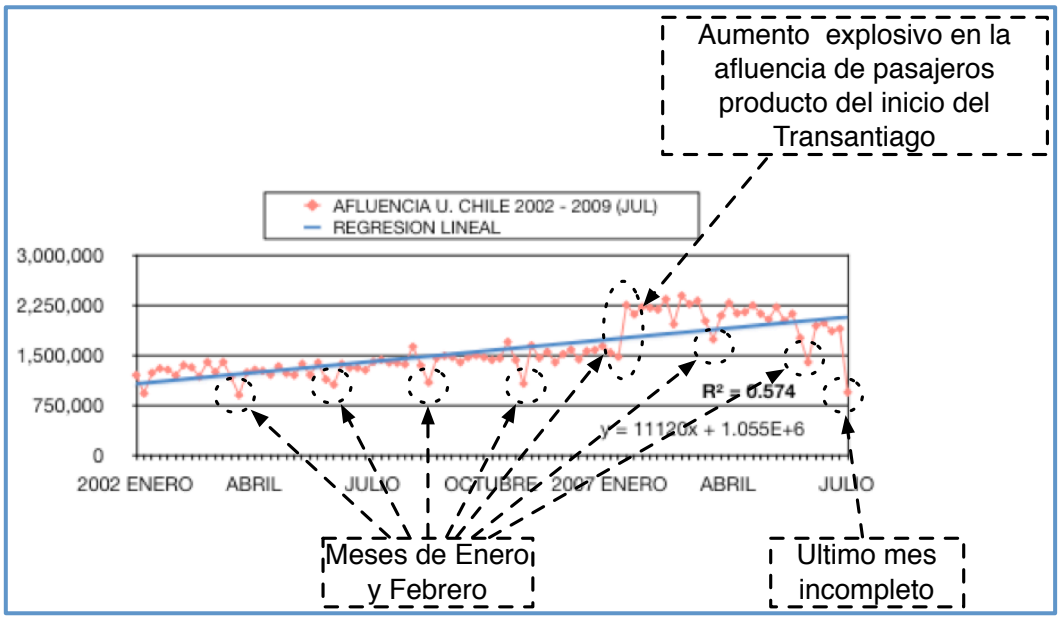
Figura 18: Afluencia promedio, antes y después de transantiago



Fuente: Elaboración propia.

Esto se expone claramente en los gráficos presentados en el Capítulo 3 (Figura 8 a 11) donde se aprecia que, a pesar del aumento explosivo en la afluencia, el comportamiento de las curvas antes y después de Transantiago es similar.

Figura 19: Serie de tiempo original y elementos a corregir



Fuente: Elaboración propia.

De manera semejante, si se representan los datos como una serie de tiempo, en ellos se refleja el inicio de Transantiago en Marzo de 2007 (ver Figura 19): se produce un brusco aumento, alrededor del doble, en la afluencia promedio con respecto a años anteriores. Por esta razón se eliminó este efecto en el análisis mediante un escalamiento de la serie de tiempo existente antes del evento. Para ello, se calculó un factor de corrección para la serie de tiempo anterior a Transantiago, donde dicho factor se calcula como el porcentaje del promedio anual que tiene la diferencia entre la afluencia luego de Transantiago con la afluencia antes de él. Este factor permitió

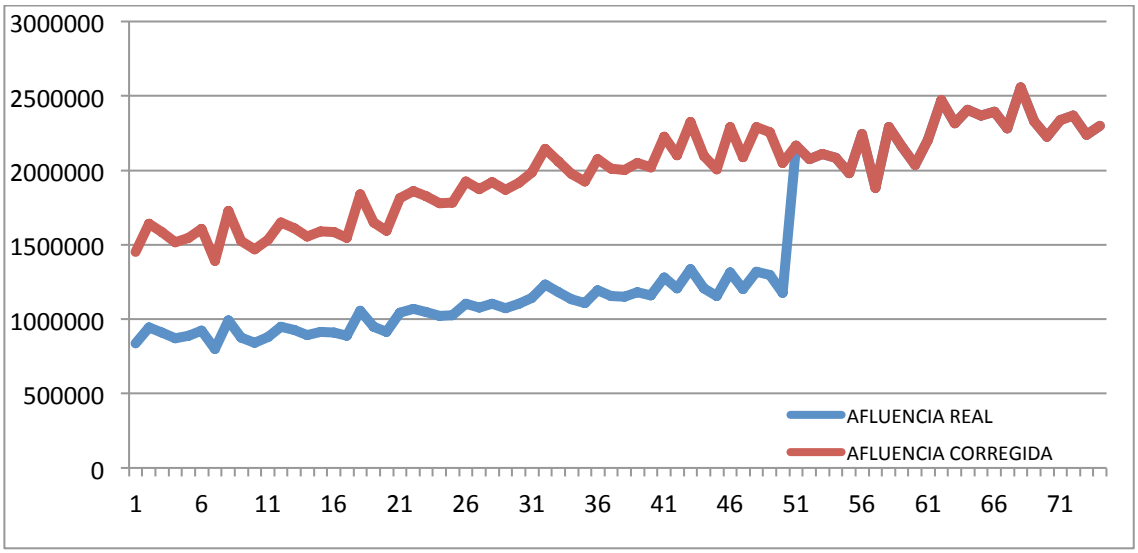
'amplificar' la serie anterior para eliminar el efecto de la puesta en marcha de Transantiago.

Asimismo, se eliminó la afluencia de la última observación, ya que no se cuenta con el dato de afluencia final para el mes de Julio de 2009, por tanto, no es un antecedente válido para realizar los modelos.

Finalmente, se eliminó de la serie los meses de enero y febrero, debido a que reflejan un fenómeno estacional producto del verano en Chile. En este periodo la afluencia de pasajeros disminuye notablemente debido a que los santiaguinos normalmente eligen estos meses para vacacionar. En consecuencia, es evidente optar por eliminar la estacionalidad del modelo y generar modelos alternativos para los meses de verano.

En conclusión, luego de aplicada la corrección de los datos y el pre-procesamiento, se construyó una serie modificada para la afluencia de pasajeros de Metro en cada una de las cuatro estaciones estudiadas. Un ejemplo se aprecia en la Figura 20, donde aparece la serie original sin los meses de enero y febrero (en rojo) versus la serie corregida final (en azul); la serie resultante (en azul) es más regular sin el efecto de Transantiago ni de las estacionalidades producto del verano. Con esto, se espera generar modelos de mejor calidad (i.e. MAPE más cercanos a cero).

Figura 20: Serie de tiempo original vs corregida para estación Escuela Militar.

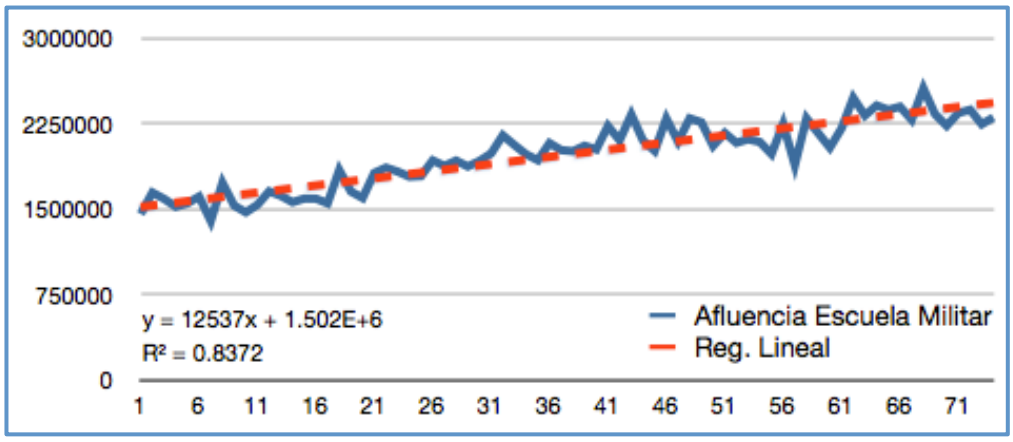


Fuente: Elaboración propia.

4.3. Aplicación de Modelos Regresivos

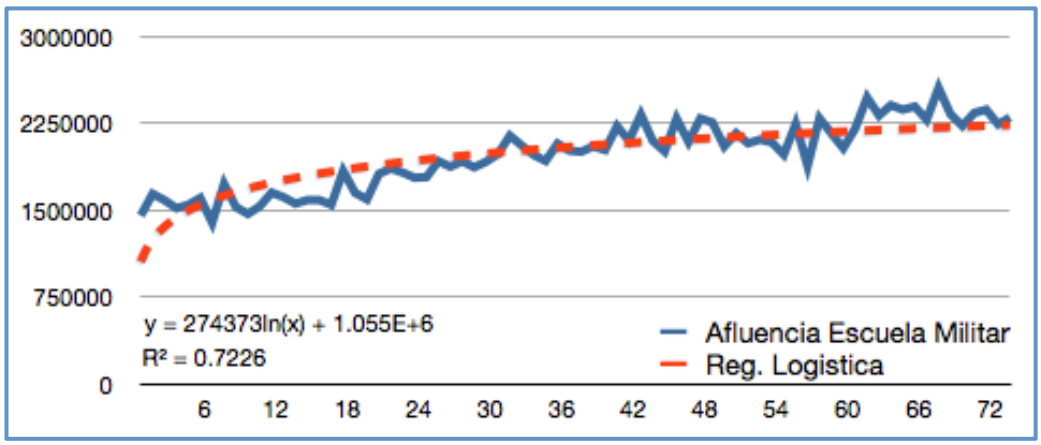
Nuevamente se aplicarán todos los modelos antes mencionados a las series corregidas, comenzando con los Modelos Regresivos. En las Figura 21 a Figura 24 se puede apreciar las series corregidas versus el pronóstico para la estación Escuela Militar. Las gráficas para el resto de las estaciones se encuentra en el Anexo A.

Figura 21: Regresión Lineal estación Escuela Militar, usando datos corregidos



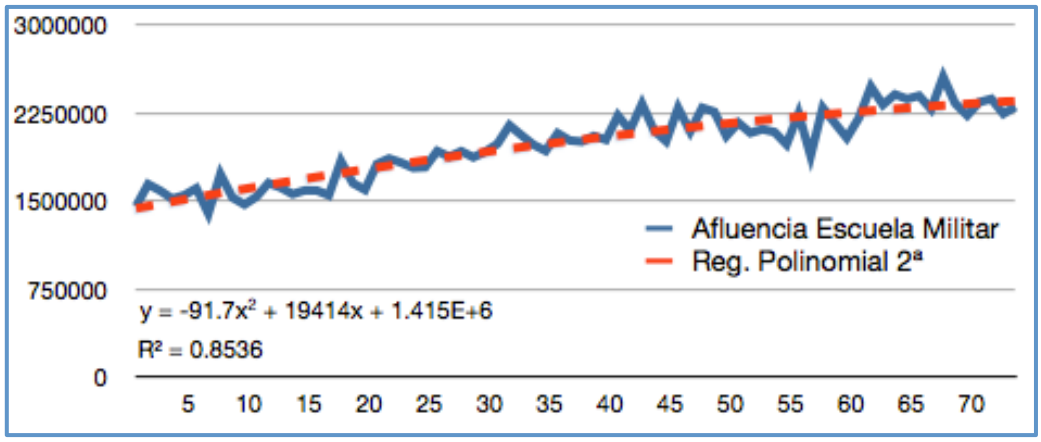
Fuente: Elaboración propia.

Figura 22: Regresión Logística Escuela Militar



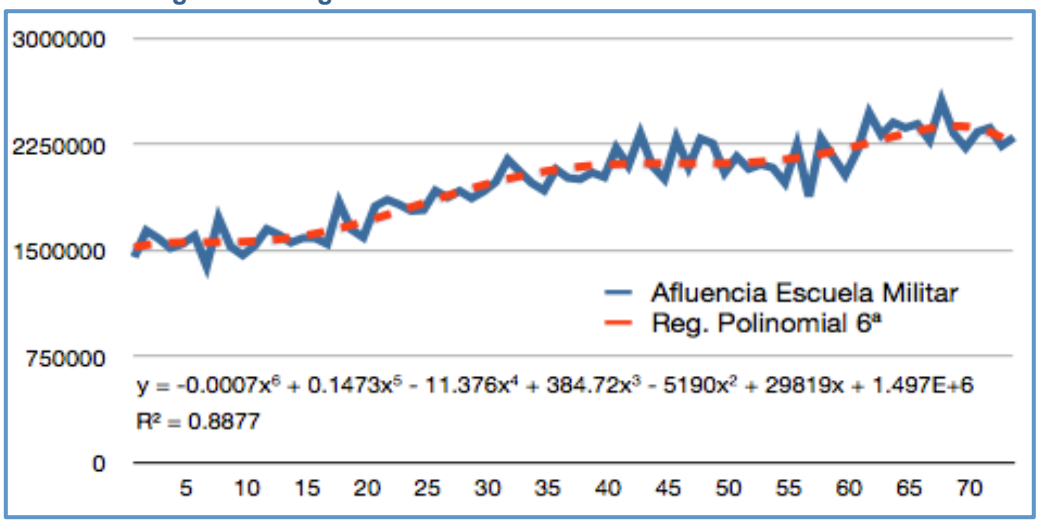
Fuente: Elaboración propia.

Figura 23: Regresión Polinomial de 2ª Orden Escuela Militar



Fuente: Elaboración propia.

Figura 24: Regresión Polinomial de 6ª Orden Escuela Militar



Fuente: Elaboración propia.

Los modelos resultantes de la aplicación de cada uno de los algoritmos a las series, se resumen desde la Tabla 9 a la Tabla 12.

Tabla 9: Modelos Finales para Estación Escuela Militar.

Regresión	Moldeo
Lineal	$y = 12537x + 1.502\text{E}6$
Logística	$y = 274373 \times ln(x) + 1.055\text{E}6$
Potencia	$y = 1.197\text{E}6\ x^{0.146}$
Exponencial	$y = 1.526\text{E}6e^{0.0065x}$
Polinomial 2ª	$y = -91.7x^2 + 19414x + 1.415\text{E}6$
Polinomial 3ª	$y = -1.5751x^3 + 85.501x^2 + 14062x + 1.45E6$
Polinomial 4ª	$y = 0.217x^4 - 34.131x^3 + 1662.3x^2 - 12634x + 1.556\text{E}6$
Polinomial 5ª	$y = -0.0097x^5 + 2.0406x^4 - 156.12x^3 + 5128x^2 - 50795x + 1.66\text{E}6$
Polinomial 6ª	$y = -0.0007x^6 + 0.1473x^5 - 11.376x^4 + 384.72x^3 - 5190x^2 + 29819x + 1.497\text{E}6$

Fuente: Elaboración propia.

Tabla 10: Modelos Finales para Estación Universidad de Chile

Regresión	Modelo
Lineal	$y = 5635.2x + 1.81E6$
Logística	$y = 136589 \times ln(x) + 1.565E6$
Potencia	$y = 1.597E6\ x^{0.0692}$
Exponencial	$y = 1.809E6e^{0.0028x}$
Polinomial 2ª	$y = -186.4x^2 + 19615x + 1.633E6$
Polinomial 3ª	$y = -8.7399x^3 + 796.84x^2 - 10081x + 1.82E6$
Polinomial 4ª	$y = -0.0174x^4 - 6.1341x^3 + 670.63x^2 - 7943.8x + 1.816E6$
Polinomial 5ª	$y = 0.029x^5 + 0.5232x^4 - 42.295x^3 + 169822x^2 - 19255x + 1.847E6$
Polinomial 6ª	$y = -0.0004x^6 + 0.080x^5 - 6.5953x^4 + 244.9x^3 - 3776.6x^2 + 23515x + 1.761E6$

Fuente: Elaboración propia.

Tabla 11: Modelos Finales para Estación Pedro de Valdivia.

Regresión	Modelo
Lineal	$y = 5257x + 895975$
Logística	$y = 117841 \times ln(x) + 698866$
Potencia	$y = 745738\ x^{0.060.112}$
Exponencial	$y = 901314e^{0.0049x}$
Polinomial 2ª	$y = -76.324x^2 + 10981x + 823467$
Polinomial 3ª	$y = -3.7764x^3 + 348.52x^2 - 1849.7x + 906339$
Polinomial 4ª	$y = 0.0493x^4 - 11.177x^3 + 706.97x^2 - 7918.5x + 930470$
Polinomial 5ª	$y = 0.0008x^5 - 0.0991x^4 - 1.2492x^3 + 424.85x^2 - 4812.9x + 921975$
Polinomial 6ª	$y = -0.0002x^6 + 0.0459x^5 - 3.9548x^4 + 154.18x^3 - 2540.6x^2 + 18354x + 875027$

Fuente: Elaboración propia.

Tabla 12: Modelos Finales para Estación Pedro de Valdivia.

Regresión	Modelo
Lineal	$y = 7496.6x + 649655$
Logística	$y = 170469 \times ln(x) + 360458$
Potencia	$y = 469641x^{0.1982}$
Exponencial	$y = 660761e^{0.0086x}$
Polinomial 2ª	$y = -135.34x^2 + 17647x + 521078$
Polinomial 3ª	$y = -6.387x^3 + 583.19x^2 - 4053.6x + 661240$
Polinomial 4ª	$y = 0.0392x^4 - 12.263x^3 + 867.77x^2 - 8871.8x + 680397$
Polinomial 5ª	$y = -0.0052x^5 + 1.0077x^4 - 77.052x^3 + 2708.9x^2 - 29139x + 735831$
Polinomial 6ª	$y = -0.00043x^6 + 0.0909x^5 - 7.2036x^4 + 253.95x^3 - 3606.4x^2 + 20197x + 635848$

Fuente: Elaboración propia.

Para evaluar la calidad de los pronósticos encontrados se utilizó el MAPE (tal como se indicó en el Capítulo 2). Los resultados se encuentran en la Tabla 13.

Tabla 13: Calidad de los Modelos usando MAPE

	Escuela Militar	U. de Chile	P. de Valdivia	San Pablo
Reg. Lineal	4.86%	5.62%	5.10%	7.76%
Reg. Logística	6.59%	5.79%	6.07%	9.69%
Reg. Polinomial 2ª	9.54%	4.97%	4.67%	6.70%
Reg. Polinomial 3ª	4.44%	3.98%	4.18%	4.28%
Reg. Polinomial 4ª	4.27%	3.97%	4.13%	4.21%
Reg. Polinomial 5ª	4.19%	3.98%	4.19%	4.56%
Reg. Polinomial 6ª	4.67%	7.09%	4.45%	9.57%
Reg. Exp.	5.20%	5.70%	5.28%	8.37%
Reg. Potencia	6.02%	5.70%	5.77%	8.87%

Fuente: Elaboración propia.

Es posible ver de la Tabla 13 que el mejor modelo para la estación Escuela Militar es un modelo Polinomial de 5º orden, pues presenta un MAPE de 4.19%. Pero en el caso de las estaciones Universidad de Chile, Pedro de Valdivia y San Pablo el mejor modelo es una Regresión Polinomial de 3º orden. En el caso de la Universidad de Chile, el modelo de 5º orden logra el mismo MAPE que el modelo de 3º orden.

En general, se puede concluir que los modelos Polinomiales permiten pronosticar mejor la afluencia de pasajeros en las estaciones de Metro.

4.4. Aplicación de Modelos de Promedios Móviles

Aunque los modelos regresivos utilizados presentaron pronósticos de buena a muy buena calidad[6], se decidió aplicar otra gama de modelos de predicción, comenzando por los más simples: los modelos de Promedios Móviles.

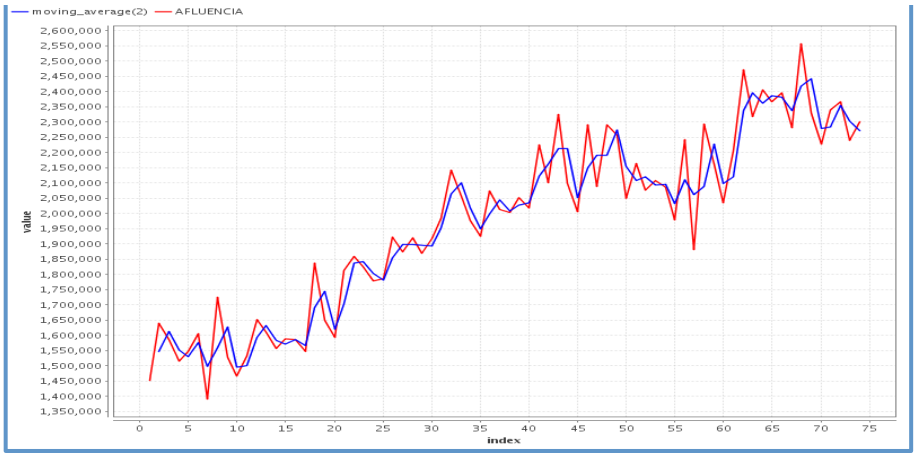
En este estudio se aplicaron ventanas de tiempo de 2, 3, 4, 5, 6, 7 y 8 meses. Sin embargo, sólo se han graficado los modelos con ventanas de tiempo de 2, 4, 6 y 8 meses para facilitar la comprensión de los resultados obtenidos y no agregar información redundante.

Los resultados de estos modelos para la estación Escuela Militar se han graficado desde la Figura 25 a la Figura 28 (para revisar todos los gráficos resultantes asociados a las estaciones en estudio ver Anexo A a Anexo D). En cada figura la línea roja representa la afluencia real corregida y la azul el pronóstico obtenido.

Es importante notar que para el cálculo de estos modelos se utilizó el software *RapidMiner*, el cual será explicado en un próximo apartado.

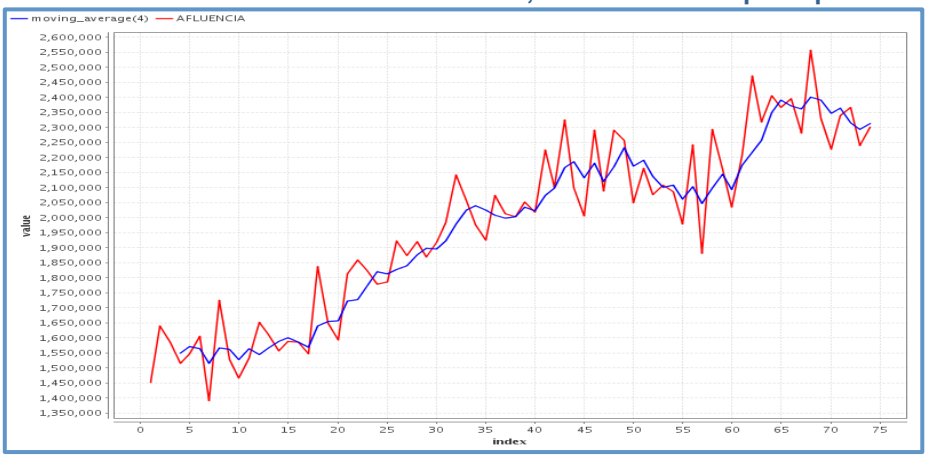
[6]Se definió 'buena calidad' considerando un MAPE de 10% y 'muy buena calidad' un MAPE inferior al 5%.

Figura 25: Promedio Moviel Escuela Militar, Ventana de Tiempo T=2 Periodos



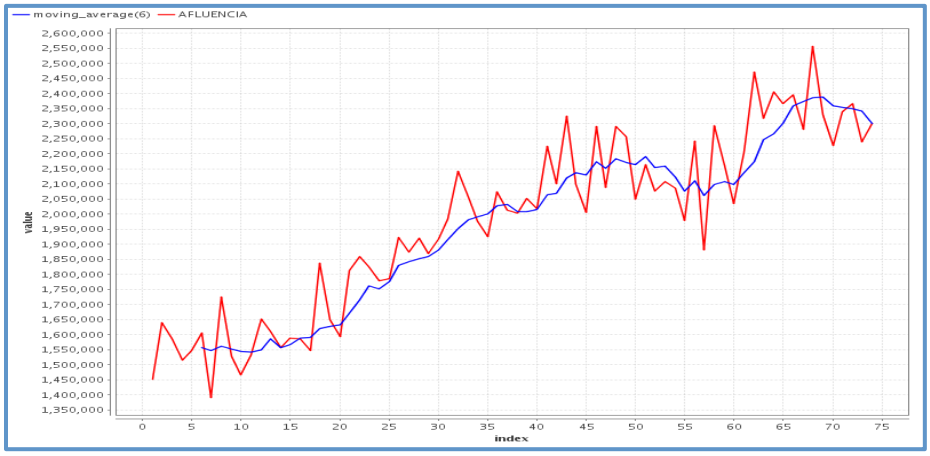
Fuente: Elaboración propia.

Figura 26: Promedio Móvil en Escuela Militar, ventana de tiempo: 4 períodos.



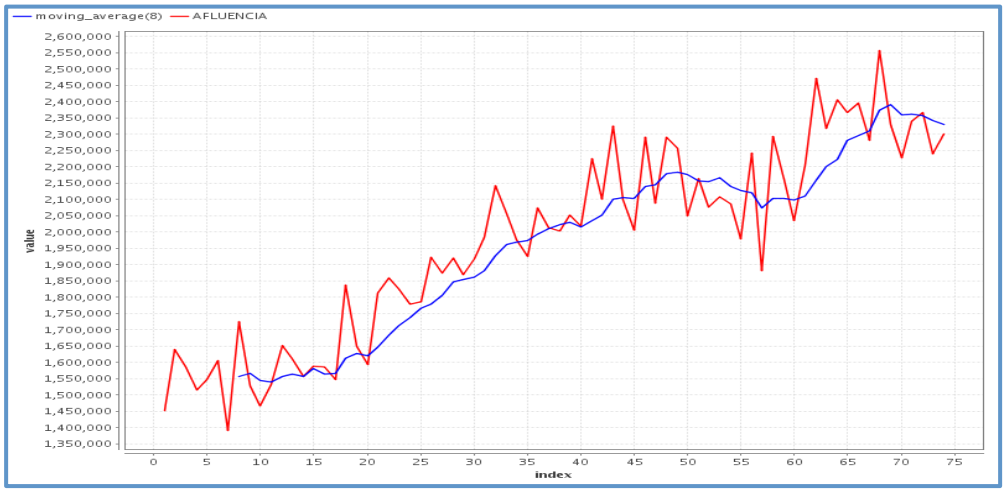
Fuente: Elaboración propia.

Figura 27: Promedio Móvil en Escuela Militar, ventana de tiempo: 6 períodos.



Fuente: Elaboración propia.

Figura 28: Promedio Móvil en Escuela Militar, ventana de tiempo: 8 períodos.



Fuente: Elaboración propia.

La calidad de los modelos de Promedios Móviles aplicados a los datos corregidos se observa en la Tabla 14. La calidad de los pronósticos de afluencia usando Promedios Móviles es muy buena ya que se logra un MAPE inferior al 5% en todos los casos experimentados.

Tabla 14: MAPE para Modelos de Promedios Moviles (ventanas de tiempo T=2 a T=8)

	Escuela Militar	U. De Chile	P. De Valdivia	San Pablo
moving_average(8)	4.15%	4.28%	3.95%	4.70%
moving_average(7)	4.10%	4.14%	3.92%	4.31%
moving_average(6)	3.82%	3.98%	3.87%	4.14%
moving_average(5)	3.56%	4.02%	3.68%	3.74%
moving_average(4)	3.38%	3.71%	3.53%	3.25%
moving_average(3)	3.24%	3.47%	3.64%	2.98%
moving_average(2)	**3.02%**	**3.23%**	**3.19%**	**2.48%**

Fuente: Elaboración propia.

El mejor modelo de Promedios Móviles se obtiene al usar una ventana de tiempo de dos períodos. La predicción incrementa su grado de error a medida que se aumenta el número de intervalos (en negrita en la Tabla 14)

Ciertamente, cuando se incorpora sólo dos meses a la predicción del siguiente periodo, el promedio responde muy bien ante variaciones bruscas en la afluencia, manteniéndose estable cuando esta última mantiene un comportamiento continuo. Este es justamente el caso de la afluencia de pasajeros en Metro. En las figuras se puede ver que el comportamiento se mantiene relativamente estable, y producto del alza explosiva en marzo del 2007, los pronósticos de ventana móvil tardan en adecuarse al nuevo nivel de afluencia de pasajeros. Por eso, el Promedio Móvil con sólo dos meses es el que mejor se ajusta a toda la serie de tiempo.

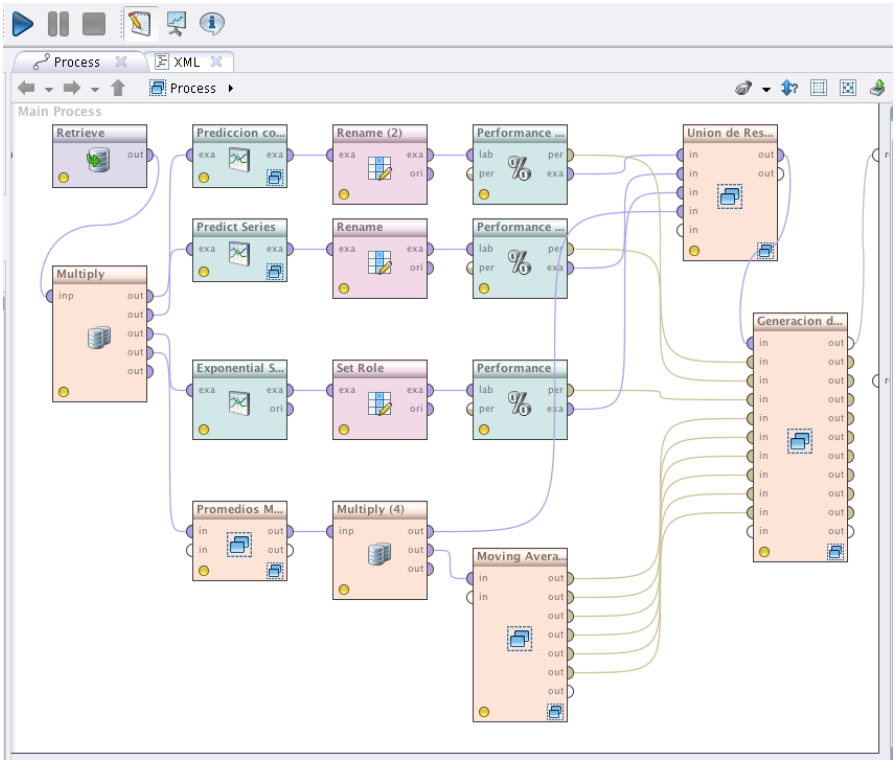
Se puede concluir que éste es un excelente método para predecir la afluencia de pasajeros en todas las estaciones del metro seleccionadas, ya que el error es inferior al 5% en todas las estaciones analizadas. En efecto, es incluso mejor que cualquiera de los modelos regresivos presentados en la sección anterior.

4.5. Aplicación de Redes Neuronales, *Support Vector Regressions* y Suavización Exponencial

Si bien, hasta este punto del análisis se ha demostrado que el modelo de Promedios Móviles es el mejor predictor para la afluencia de pasajeros, de igual manera se optó por experimentar con modelos más avanzados de Minería de Datos; con este fin se seleccionaron las Redes Neuronales y las *Support Vector Regression* (SVR). Además, se aplicará el modelo de Suavización Exponencial, uno de los más utilizados en el mundo por su simplicidad y buenos resultados.

En el caso de predicción utilizando Red Neuronal, se eligió una red muy simple con *input layer*, *hidden layer* y capa de salida, a la que se denomina red neuronal de una sola capa (del ingles *single layer neural network*).

Figura 29: Modelo usando Rapid Miner 5.0.



Fuente: Elaboración propia.

Todos los modelos antes detallados y los de esta sección fueron realizados utilizando *Rapidminer* 5.0[7], excepto los de Suavización Exponencial. Si bien se

[7] Visite http://rapid-i.com para descargar el software.

utilizaron otras herramientas al inicio (SPSS por ejemplo), finalmente, se decidió migrar todo a *Rapidminer* pues resulta simple presentar los modelos realizados a través de este software, el cual cuenta con todas las funciones necesarias para esta investigación de manera gratuita (ver Figura 29).

La versión 5.0 de *Rapidminer* posee un módulo para predicción de series de tiempo, el cual tiene la particularidad de simplificar el uso de modelos complejos, como es el caso de las Redes Neuronales o las *Support Vector Regressions,* además de presentar muchísimas opciones, como funciones para el aprendizaje, número de iteraciones, etc., que permiten el análisis exploratorio usando estos modelos con diversas combinaciones de parámetros.

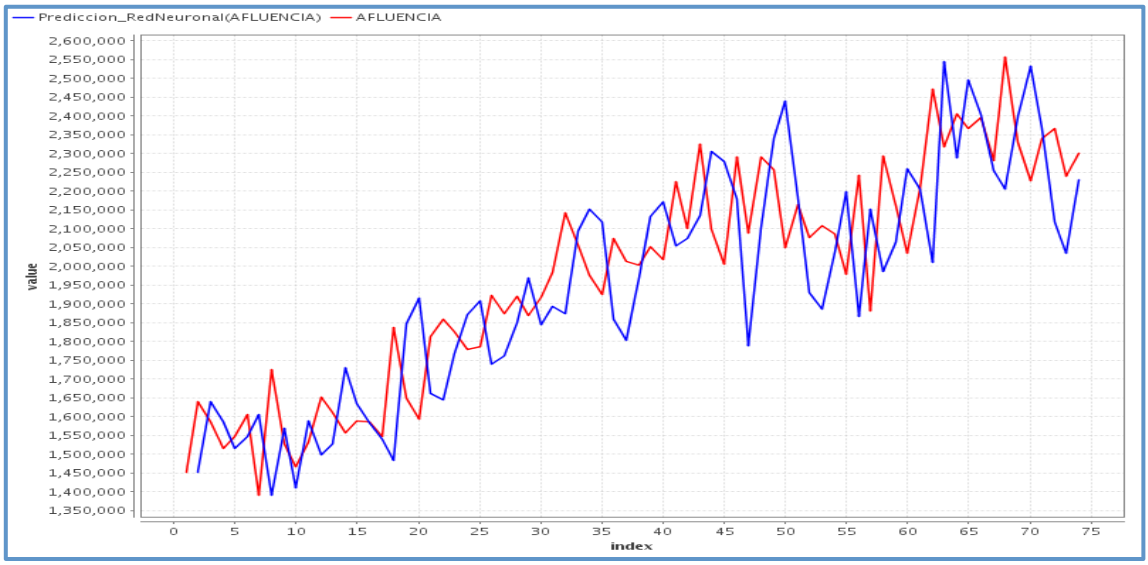
Tabla 15: Parámetro α, óptimo para el modelo de Suavización Exponencial.

	Escuela Militar	U. de Chile	P. de Valdivia	San Pablo
α^*	0.5171	0.5142	0.3193	0.6531

Fuente: Elaboración propia.

Lamentablemente, el módulo de Suavización Exponencial de *Rapidminer* tiene un error: provoca que los resultados de los pronósticos sean mucho mejores de lo que son en realidad, por esta razón y por su simplicidad, finalmente se utilizó Excel para calcularlos. Además, se empleó la herramienta gratuita de Excel *solver* para obtener el parámetro *alfa óptimo* (α^*) del modelo. Esto se resolvió minimizando el MAPE, usando como variable α, sujeto a que $0 \leq \alpha \leq 1$. Los resultados pueden encontrarse en la Tabla 15.

Figura 30: Red Neuronal para estimar afluencia de pasajeros en Estación Escuela Militar.



Fuente: Elaboración propia.

Para configurar los parámetros de cualquier algoritmo en *Rapidminer*, existen operadores que permiten encontrar el conjunto de parámetros generadores de la mejor predicción. Sin embargo, cuando se experimentó con dichos operadores, el proceso no pudo finalizar, incluso luego de varias horas de procesamiento (alrededor de 15h). Por esta razón se realizó una calibración manual de los parámetros fundamentales, como las funciones de kernel usadas en las *Support Vector Regressions*, hasta encontrar los parámetros que permitieran minimizar el MAPE de los pronósticos.

La aplicación del modelo de Red Neuronal para la predicción de afluencia en la estación Escuela Militar tiene un error del 7.71% (ver Tabla 16), cifra mucho peor que los mejores modelos regresivos (Tabla 13) y que la predicción efectuada usando Promedios Móviles (Tabla 14). En la Figura 30 se ha graficado la afluencia real versus la predicción de la afluencia usando Red Neuronal.

Asimismo, al aplicar una Red Neuronal para estimar la afluencia de pasajeros en las cuatro estaciones seleccionadas y compararla con los modelos previos, se aprecia que las Redes Neuronales en general tienen un comportamiento deficiente con respecto a las regresiones y los Promedios Móviles.

Por otra parte, al aplicar una *support vector regression* para predecir la afluencia de pasajeros en las cuatro estaciones de Metro, los resultados son mejores que al aplicar Redes Neuronales. Esto se comprueba al tomar como ejemplo nuevamente la estación Escuela Militar, donde la predicción usando una SVR da un MAPE de 4.59% versus la Red Neuronal que presenta un MAPE de7.71% (el gráfico de la estimación de afluencia versus afluencia real usando SVM se encuentra en la Figura 31).

Tabla 16: Resumen MAPE

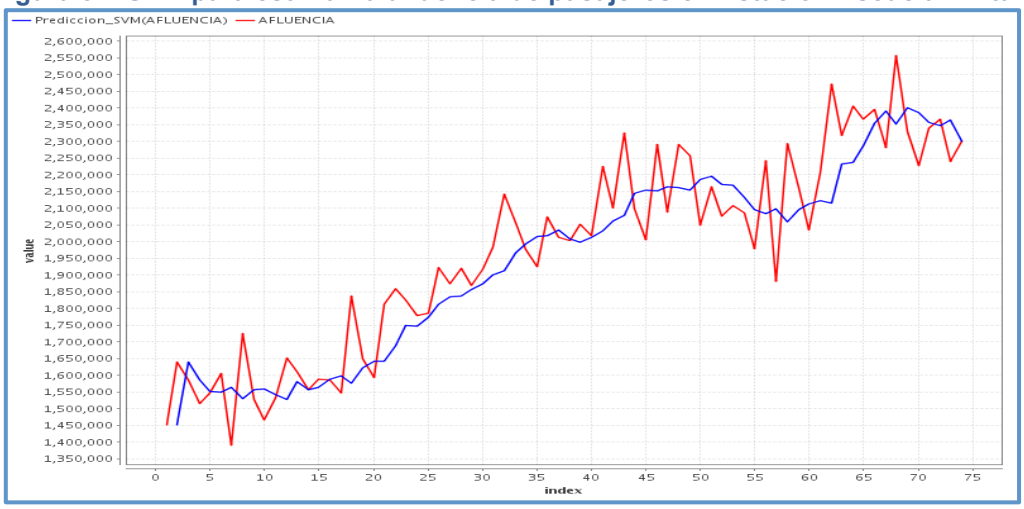
	Escuela Militar	U. de Chile	P. de Valdivia	San Pablo
Suavización Exponencial	**4.59%**	4.86%	**4.74%**	**4.57%**
Red Neuronal	7.71%	7.61%	8.92%	6.20%
SVR	**4.59%**	**4.81%**	4.77%	5.00%

Fuente: Elaboración propia.

Finalmente, se aplicó con gran éxito el modelo de Suavización Exponencial para la estimación de afluencia, puesto que las cuatro estaciones presentaron un MAPE inferior a 5% (ver Tabla 16) Además, la configuración de parámetros requirió muy poco trabajo ya que el modelo tiene sólo dos para configurar.

La Figura 32 presenta el gráfico de estimación de afluencia de pasajeros versus la afluencia real en la estación Escuela Militar, donde se refleja la gran cercanía entre el pronóstico generado y la serie real.

Figura 31: SVM para estimar la afluencia de pasajeros en Estación Escuela Militar.

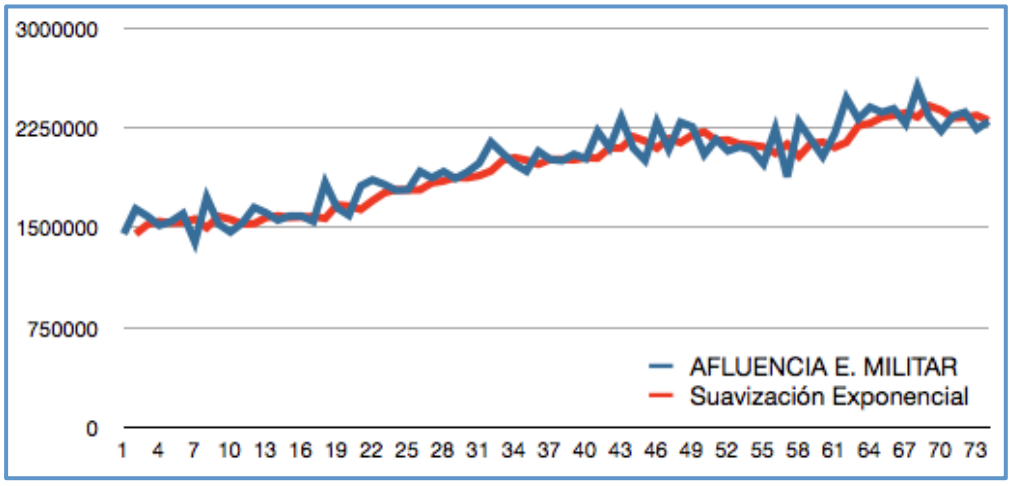


Fuente: Elaboración propia.

En conclusión, los mejores modelos para predecir la afluencia de pasajeros son el de Suavización Exponencial y las SVR, ya que el MAPE obtenido en todos los casos de estudio es inferior o muy cercano a 5%. Por el contrario, el uso de Redes Neuronales para predecir la afluencia entrega un MAPE superior a este valor.

Existen otras características que vuelven muy interesante la Suavización Exponencial, como son la simplicidad teórica del modelo, la velocidad de cálculo y la simpleza de la configuración de los parámetros. Esto lo hace muy recomendable incluso para personas no expertas en Minería de Datos o Estadística, quienes pueden aplicarlo sin perder de vista el problema en cuestión. Por otra parte, gracias a esta misma simplicidad, es posible el empleo de herramientas de uso cotidiano como Excel.

Figura 32: Modelo de Suavización Exponencial para estimar afluencia de pasajeros en Estación Escuela Militar.



Fuente: Elaboración propia.

4.6. Uso de *Rapid Miner* para modelamiento

Como ya se discutió en la sección anterior, la herramienta seleccionada para la elaboración y aplicación de los modelos seleccionados es *Rapidminer* 5.0., software con diversas características que lo vuelven muy apropiado para este análisis.

En primer lugar, *Rapidminer* es gratuito, se entrega con un *General Public Licence* (GPL). En segundo lugar es OpenSource, por lo que la empresa podría realizar modificaciones al código fuente del software si así lo requiriera, o bien, puede conectar los sistemas ya existentes a *Rapidminer* para obtener visualmente los resultados de los pronósticos de manera automática. También es posible generar procesos autoejecutables cada cierto período de tiempo (una vez al mes, diariamente) para calcular los pronósticos sin intervención humana.

Otro elemento fundamental es que permite trabajar con repositorios de datos locales o remotos: éstos no necesariamente deben ser generados específicamente para el proceso de pronósticos, pues es posible realizar una conexión a la base de datos para obtenerlos. Los motores de bases de datos que actualmente soporta *Rapidminer* son: Mysql, PostgreSql, Ingres, Microsoft SQL Server, Oracle & Sybase.

El trabajo con esta herramienta se hace por medio de la creación de procesos que pueden ser almacenados y luego aplicados a los datos de una o más estaciones, sin necesidad de mayor configuración. Obviamente, es tarea del Analista a cargo realizar cualquier ajuste que estime conveniente en los parámetros. Sin embargo, utilizando los módulos de optimización de parámetros, el modelo podría encontrarlos automáticamente. Es lamentable que dichos módulos no se terminaran de procesar en un tiempo razonable, probablemente con el hardware adecuado, esta situación se podría mejorar.

Finalmente, *Rapidminer* posee un módulo/operador de generación de reportes, el cual puede crear diversos tipos de reportes en Word, Excel, PDF, etc., y reportes gráficos sin programación de por medio. De esta forma se diseñaron varias de las figuras de esta investigación, así como el volcamiento de todos los pronósticos a tablas Excel, donde los resultados pueden ser manipulados libremente.

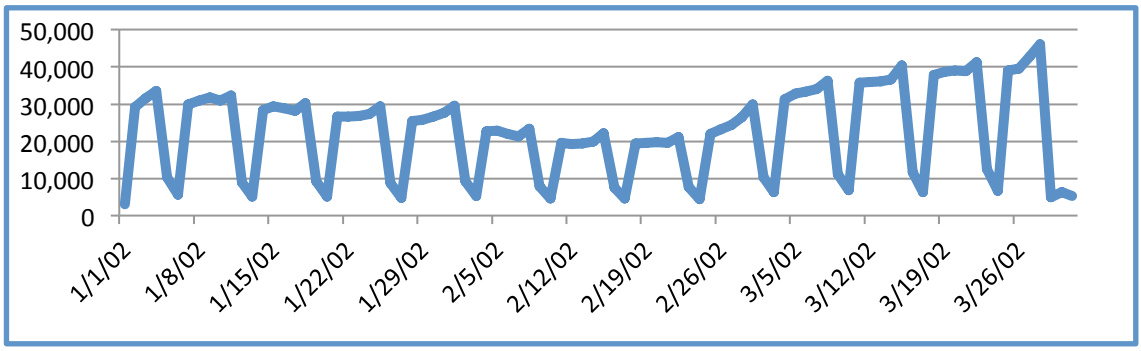
4.7. Pronósticos a muy corto plazo

Además de los promedios mensuales, se intentó obtener predicciones a muy corto plazo utilizando todos los modelos de pronósticos utilizados en el cálculo de pronósticos mensuales (Modelos Regresivos, Modelos de Promedios Móviles, Suavización Exponencial, Redes Neuronales y SVMs).

Al analizar los datos diarios de afluencia en la estación Escuela Militar se aprecia una estacionalidad muy marcada a lo largo del año, siendo las semanas menos constantes las de los meses de verano (ver Figura 33 y Figura 34), sin embargo, aunque éstas son las menos variables, su comportamiento se mantiene dentro de la tendencia y estacionalidades esperadas.

Un estudio de afluencia diaria realizado para las otras estaciones seleccionadas en esta investigación, arroja la misma estacionalidad.

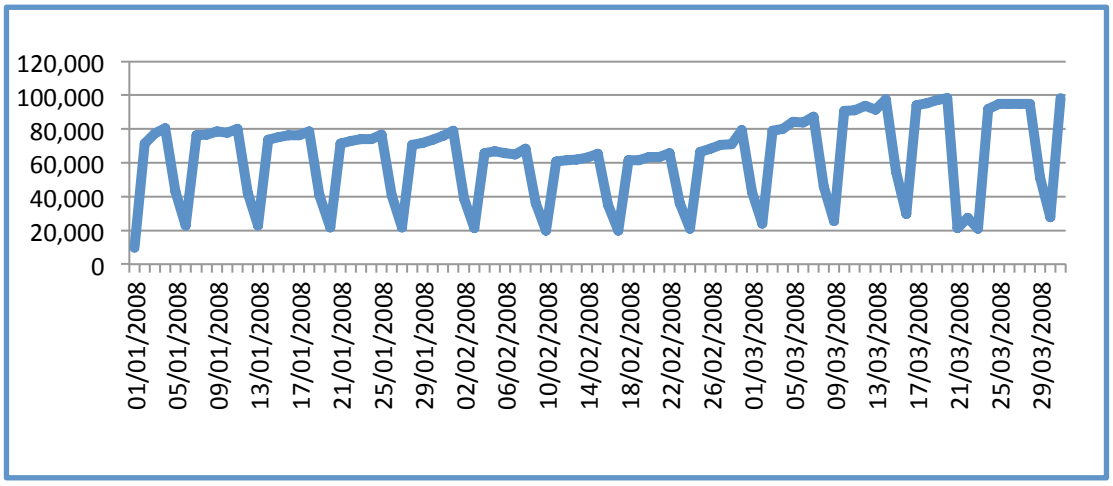
Figura 33: Afluencia diaria Estación Escuela Militar (Ene-Mar 2002)



Fuente: Elaboración propia.

Para calcular los modelos en cada estación, se tomó como base el mismo proceso utilizado en la estimación de afluencia de pasajeros mensual. Lo único necesario fue la creación de un archivo con los datos de la afluencia diaria en lugar de los datos de afluencia mensual. Con esto, el proceso creado en *Rapidminer* se pudo aplicar sin margen de error, y los reportes fueron generados correcta y automáticamente.

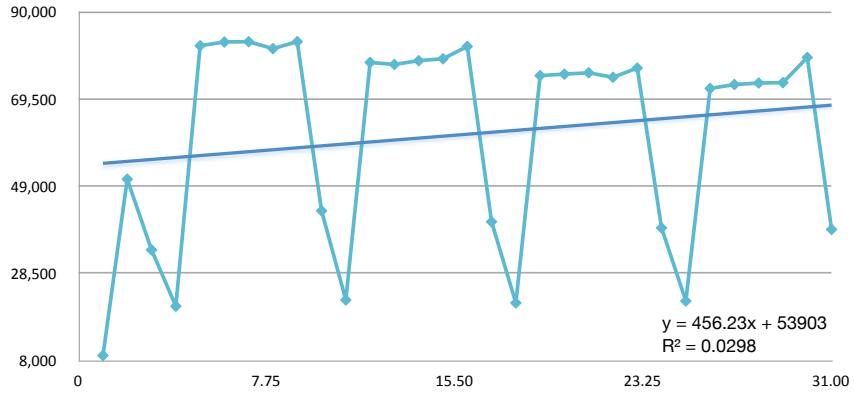
Figura 34: Afluencia Diaria Escuela Militar (Ene-Mar 2008)



Fuente: Elaboración propia.

Los resultados de la aplicación de Modelos Lineales tuvieron un factor de correlación lineal menores a 0,1 en todos los casos (ver Figura 35). Por tanto se confirmó la hipótesis de que los Modelos de Regresión no eran adecuados para este caso, dado que el comportamiento de la afluencia diaria no es lineal, por ende, se ajustaría mejor un Modelo Polinomial.

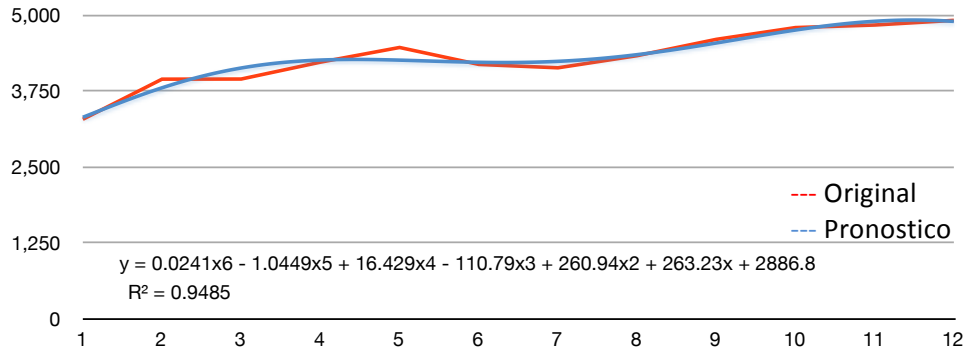
Figura 35: Afluencia Diaria de Pasajeros en Escuela Militar (Ene 2009)



Fuente: Elaboración Propia

Sin embargo, al eliminar los fines de semana y mantener solamente los días de semana, se puede construir Modelos Regresivos Lineales con factores de correlación más cercanos a 1; en la Figura 36 se puede apreciar un ejemplo.

Figura 36: Regresion Polinomial 6º, Afluencia Diaria 12 días en Escuela Militar



Fuente: Elaboración Propia

Al aplicar en este caso Promedios Móviles, Suavización Exponencial, Redes Neuronales y SVRs, todos los modelos calculados arrojaron errores inferiores al 2% (a diferencia del caso de pronóstico de afluencia mensual). Esto afirma que los modelos aplicados a la predicción de afluencia diaria fueron un éxito. Sin embargo, también se aprecia que la estacionalidad y los niveles de afluencia son tan constantes, que los modelos prácticamente no ofrecen ninguna nueva información a la hora de predecir la afluencia de pasajeros en el muy corto plazo.

En consecuencia, se concluye que la aplicación de modelos para calcular la afluencia diaria de pasajeros no tiene mayor valor para la compañía. Aunque, una recomiendación sería el uso de promedios moviles por ser la técnica más simple de

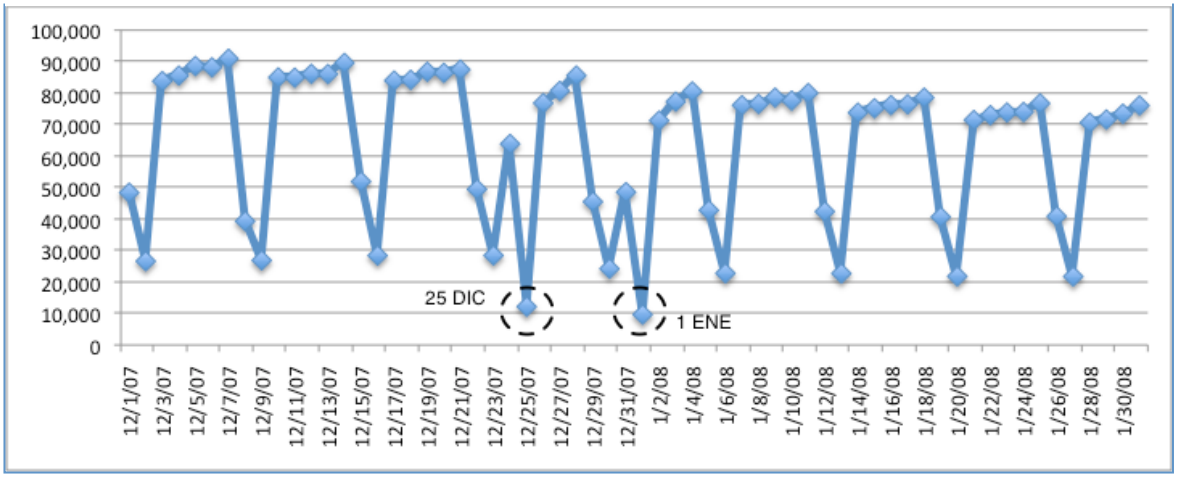
implementar y que entrega errores inferiores al 2% al predecir la afluencia en el muy corto plazo.

4.8. Patrones de afluencia anómalos en fechas especiales

Para caracterizar la afluencia de pasajeros de la mejor forma posible, realizó un análisis de aquellas fechas donde se sabe *a priori* que la afluencia de pasajeros será afectada, es decir, se producirá un aumento o disminución en la afluencia de pasajeros en una o varías estación(es) específica(s) de metro. Esto se pude deber, por ejemplo, a la realización de un partido de futbol de la selección chilena un día sábado en el estadio Nacional, habrá mayor concurrencia de personas a este lugar y aumentará la afluencia de pasajeros en las estaciones más cercanas y que permiten tomar movilización hacia allá, como las estaciones de Av. Grecia, Ñuble o Irarrázaval.

Lamentablemente, un fenómeno como el anterior no se puede explicar con los datos actuales, pues no se cuenta con la información de todas las estaciones involucradas. Realizar un estudio de tal magnitud, se encuentra fuera del alcance de esta investigación. No obstante lo anterior, se intentó buscar fechas que presentaran un comportamiento anómalo de afluencia de pasajeros y que afectara una o varias de las cuatro estaciones seleccionadas para este estudio.

Figura 37: Navidad y Año Nuevo 2007

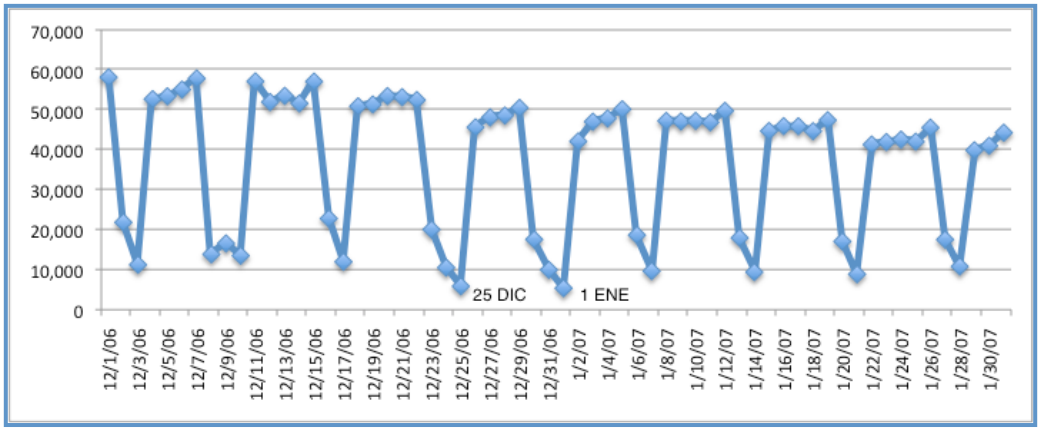


Fuente: Elaboración propia.

El principal problema al analizar los patrones de afluencia anómalos es la escasez de datos, la cual impide generar un modelo. Por ejemplo, si se aísla la afluencia de pasajeros en Navidad (ver Figura 37 y Figura 38), se puede estudiar la afluencia desde el año 2002 al año 2009 en los días 24 y 25 de diciembre, pero esto corresponde a sólo 7 puntos en cada día. Además, el año 2007, con la puesta en marcha de Transantiago, los datos históricos usados para predecir específicamente estas fechas anómalas quedaron obsoletos, pues aunque la tendencia y los valores sean escalados, los efectos de este aumento de demanda no pueden ser explicados sólo con 7 observaciones, es decir, no permiten generar un modelo (ver Tabla 17). Por otra parte,

en estos fenómenos o fechas especiales también afecta la afluencia de pasajeros el hecho de ser día laboral o fin de semana.

Figura 38: Navidad 2006 y Año Nuevo 2007.



Fuente: Elaboración propia.

En resumen, no hay datos suficientes para construir un modelo con significancia estadística ni tampoco es posible establecer un análisis para el día 24 de diciembre (víspera de Navidad) o el 1 de enero.

Tabla 17: Afluencia de pasajeros en Escuela Militar para Navidad.

Fecha	Afluencia
Dec 25, 2002	3,729
Dec 25, 2003	3,802
Dec 25, 2004	4,721
Dec 25, 2005	5,626
Dec 25, 2006	5,851
Dec 25, 2007	12,169

Fuente: Elaboración propia.

Del mismo modo, una alternativa de estudio eran los días 11 de septiembre, la semana del 18 de septiembre o el Día del Joven Combatiente (29 de Marzo), fechas donde se sabe *a priori* que la afluencia de pasajeros sufre desviaciones a su comportamiento normal. Pero por las razones antes mencionadas no fue posible generar un modelo confiable.

Por otra parte, intentar una generalización en términos de las variables que afectan la afluencia de pasajeros teniendo pocas observaciones tampoco es posible. Incluso en este caso, ya que es aún más complejo obtener un modelo creíble, pues se requiere al menos establecer una hipótesis previa indicando cuáles son los eventos externos que afectan la afluencia de pasajeros, para luego acumular datos que comprueben el modelo. Esto no ha sido efectuado por Metro y sería tarea de varios años generar los datos para hacerlo

Capítulo 5: Conclusiones

El presente trabajo fue realizado con el objeto de predecir la afluencia de pasajeros en Metro S.A. Para esto se estableció un conjunto de cuatro estaciones representativas del comportamiento de la Línea 1: Universidad de Chile, Pedro de Valdivia, Escuela Militar y San Pablo, según recomendaciones de analistas expertos.

Se analizó el comportamiento de afluencia de pasajeros desde enero del 2002 hasta julio del 2009 en estas cuatro estaciones, y se estudió la afluencia mensual en cada una de ellas. Luego se seleccionaron diversos modelos: Regresiones (lineal, polinomial, exponencial, etc.), Promedios Móviles (con ventanas de tiempo de 2 a 8 meses), Redes Neuronales, *Support Vector Regression* y Suavización Exponencial, y se implementaron en *Rapidminer* 5.0 (se pueden aplicar todos de una sola vez).

Se detectó un error en *Rapidminer* al calcular la Suavización Exponencial, por tanto fue necesario implementar este algoritmo en Excel y calcular el parámetro alfa óptimo usando la herramienta *Solver* de excel, que es gratuita.

Se realizó una primera aplicación de los modelos a las series de afluencia originales para detectar su factibilidad de aplicación, y se calculó el MAPE como medida de calidad en los pronósticos.

Gracias a lo anterior se determinó que las series sí eran pronosticables por los modelos, y se detectó varios efectos que debían ser removidos, como las estacionalidades y el efecto de Transantiago sobre la afluencia de pasajeros. Las series originales dieron origen a series corregidas, que fueron los datos utilizados finalmente para la construción de los modelos y el calculo de los pronósticos.

Los resultados de la aplicación de estas técnicas demostraron que el mejor modelo para pronosticar la afluencia de pasajeros en el corto plazo (mensual) es el de Promedios Móviles, con un MAPE bajo 3.3% al predecir la afluencia de pasajeros en todas las estaciones seleccionadas para los experimentos. Por lo tanto, se puede afirmar que es posible predecir la afluencia de pasajeros mensual de manera muy precisa.

Por otra parte, se intentó pronosticar la afluencia en el muy corto plazo. Para esto se analizó la afluencia diaria en las cuatro estaciones seleccionadas y se aplicaron los modelos antes descritos. El resultado mostró que el comportamiento de la afluencia es muy periódico y constante, por ello los efectos de todos los pronósticos eran muy buenos (menores a 2% de error). Lo cual puede ser interpretado como un éxito. Sin embargo, para efectos de entregar nueva información a los tomadores de decisión, no lo es tanto. Porque incluso con una inspección simple de los datos, el experto puede entregar un pronóstico de bajo error usando un primedio simple de los datos. Por lo tanto, la aplicación de técnicas de Minería de Datos sólo ayuda a entregar mayor objetividad a las decisiones del día a día.

Finalmente, se intentó caracterizar ciertas fechas o períodos en los cuales se sabe de antemano que la afluencia de pasajeros será afectada. Fechas como Navidad, Año

Nuevo, víspera de partidos o manifestaciones, son caracterizadas como fenómenos de afluencia anómala. Sin embargo, el grave problema al que se enfrentó la investigación al aplicar un modelo o técnica estadística, fue el no contar con suficientes observaciones para estas fechas. La entrada en vigencia de Transantiago en marzo del 2007 fue una de las razones de esta complicación, pues sólo existen datos representativos desde esta fecha. Por otra parte, debido a que se trata de fenómenos anómalos (sólo hay una oportunidad en el año para estudiarlas), existen cuatro observaciones en el mejor de los casos y dos en el peor de los casos (por ejemplo, Navidad sólo cuenta con la afluencia del 2008 y 2009). Además, es muy importante contar con al menos dos semanas, porque estas series de tiempo son fuertemente afectadas por el día que corresponde, ya que los fines de semana hay muchísimo menos afluencia que en un día laboral. En consecuencia, se requieren al menos 14 d para generar un modelo con significancia estadística.

5.1 Trabajo Futuro

Como trabajo a futuro es muy recomendable aplicar los modelos analizados para estudiar las fechas con afluencia anómala. Sin embargo, para realizarlo se deben recopilar datos que hoy en día no existen, pues con ellos además se podría estudiar cómo y en qué magnitud ciertas variables afectan la afluencia de pasajeros.

Otro factor interesante de poder estudiar es la afluencia de pasajeros por zona horaria. Por ejemplo, en horarios punta ó en horarios valle, dicha información puede ser muy util para poder establecer la frecuencia de los trenes.

Bibliografía

1] Metro S.A., "Memoria Anual Metro S.A.," Metro S.A., Santiago, 2006.

2] Metro, "Memoria Anual Metro S.A.," Santiago, 2007.

3] U. Fayyad, G. Piatesky-Shapiro, and P. Smith, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, pp. 17 (3): 37-54., 1996.

4] J. Han, M. Kamber, and J. Pei, *"Data Mining: Concepts and Techniques"*, 2nd ed.: Morgan Kaufmann, 2005.

5] H. Bozdogan, Ed., *Statistical Data Mining and Knowledge Discovery*, 1st ed., 2003.

6] E. Frank I. H. Witten, *"Data Mining: Practical Machine Learning Tools and Techniques"*, 2nd ed.: Morgan Kaufmann, 2005.

8] V. Devedzic, "Knowledge discovery and data mining in databases," Technical report, School of Business Administration, University of Belgrade, Yugoslavia, 2002.

7] M. J. Berry G. S. Linoff, *"Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management"*, 2nd ed.: Wiley Computer Publishing, 2004.

9] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *J Data Warehousing*, no. 5, p. 13—22, 2000.

10] S. Sumathi and S.N. Sivanandam, *Introduction to Data Mining and its Applications*, 1st ed.: Springer, 2006.

11] SAS Publishing, *"Data Mining Using SAS Enterprise Miner: A Case Study Approach"*, 2nd ed.: SAS Publishing, 2009.

12] A. Azevedo and M. F. Santos, "KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW," , 2008.

13] G. Maddala, *Introducción a la Econometría*.: Prentice-Hall Hispanoamerica, 1996.

14] J. Vial, "An Econometric Study of the World Copper Market," in *Ph.D. Dissertation.*, University of Pennsylvania. CIEPLAN., 1988, p. Notas Técnicas Nº112.

15] D. Larose, *Discovering Knowledge in Data*. New Jersey, EEUU: Wiley, 2005.

16] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, 3rd ed. Potomac, U.S.A.: Two Crows, 1999.

17] S. Vallejos, "Minería de Datos," Facultad de Ciencias Exactas y Agrimensura, Universidad Nacional del Nordeste, Corrientes, Tesis 2006.

18] R. Yaffee and M. McGee, *An Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS*.: Academic Press, 2000.

19] D. Popovic A. K. Palit, *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*.: Springer, 2005.

20] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed.: Springer, 2002.

21] J. Guajardo, R. Weber, and J. Miranda, "A model updating strategy for predicting time series with seasonal patterns," *Appl. Soft Comput.*, no. 1, pp. 276-283, 2010.

22] M. Ávila, E. Gómez, J. Vilasis, O. Mulet & F. Mazzanti A. Gonzáles, "Redes Neuronales Para Identificación y Predicción de Series de Tiempo," *Centro de Investigación de la Universidad de La Salle*, pp. 45-65, 2000.

23] K. Gurney, *An Introduction to Neural Networks*.: CRC Press, 1997.

24] L. V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms And Applications*, 1st ed.: Prentice Hall, 1993.

25] Ch. M. Bishop, *Neural Networks for Pattern Recognition*, 1st Edition, Ed. USA: Oxford University Press, 1996.

26] J Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *J. Neurocomputing*, vol. 74, no. 1-3, 2010.

27] M L. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry, Expanded Edition*.: The MIT Press, 1987.

28] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1st ed.: Cambridge University Press, 2000.

29] S. Salcedo-Sanz, A. M. Pérez-Bellido, J. A. Portilla-Figueras E. G. Ortiz-García, "Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions," *Neurocomputing*, vol. 16-18, no. 3683-3691, Oct. 2009.

30] K. Huang, I. King, M. R. Lyu H. Yang, "Localized support vector regression for time series prediction," *Neurocomput.*, vol. 10-12, no. 2659-2669., June 2009.

31] A. J. Smola and B. Schölkopf., "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2004.

32] Wikipedia.org. (2007) Wikipedia. [Online]. http://es.wikipedia.org/wiki/Transantiago

33] Metro S.A. (1999) Sitio Web Metro S.A. [Online]. www.metro.cl

34] Mercurio. (2008) Historia del Metro. [Online]. www.emol.cl

35] Wikipedia. (2009) [Online]. http://es.wikipedia.org/wiki/Metro_de_Santiago

36] Plataforma Urbana. (2008, Septiembre) El futuro del metro de Santiago: ¿Vitacura, Irarrázaval o Cerrillos? [Online]. http://www.plataformaurbana.cl/archive/2009/09/08/el-futuro-del-metro-de-santiago-%c2%bfvitacura-irarrazaval-o-cerrillos/

37] U. Fayyad, G. Piatesky-Shapiro, and P. Smith, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, pp. 17 (3): 37-54., 1996.

38] R. Pindyck and D. Rubinfeld, *Econometria: Modelos y pronósticos*. Mexico D.F.: Mc Graw Hill, 2001.

39] P. Isasi and I. Galvan, *Redes Neuronales Artificiales. Un enfoque práctico*. Madrid: Pearson Educación S.A., 2004.

40] J. Hanke and A. Reitsch, *Estadística para negocios*.: Irwin Professional Publishing, 1995.

41] Christián Foix, *Proyección del precio del cobre: ¿Herramientas de inteligencia computacional o series de tiempo?* Santiago, Chile: Universidad de Chile, 2007.

42] W Enders, *Applied Econometric Time Series*.: John Wiley & Sons, Inc., 2004.

43] S. Vishal and D. Srinivasan, ""Evolutionary Computation and Economic Time Series Forecasting"," in *2007 IEEE Congress on Evolutionary Computation*, New York, Brasil, 2005, pp. 188-195.

44] D. Gujarati, *Econometria*, 4th ed. Mexico: Mc. Graw Hill, 2004.

45] D. Pyle, *Business Modeling and Data Mining*.: Morgan Kaufmann, 2003.

46] A. Novales, *Estadística y Econometría*. Madrid, España: Mc Graw Hill, 1997.

47] T. Khabaza. (2002) Hard Hats for Data Miners: Myths and Pitfalls of Data Mining, Business intelligence, data warehousing and analytics editorial. [Online]. http://www.spss.ch

48] R. Kimball and J. Castera, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*, 1st ed.: Wiley, 2004.

49] R. Nau. (2006) What's the bottom line? How to compare models. [Online]. www.duke.edu/~rnau/compare.htm

50] L. Aburto and R. Weber, "Improved supply chain management based on hybrid demand forecast," *Applied Soft Computing*, vol. 7, no. 1, pp. 136-144, 2007.

51] M Clements and D Hendry, *Forecasting Non-Stationary Economic Time Series*.: The MIT Press, 2001.

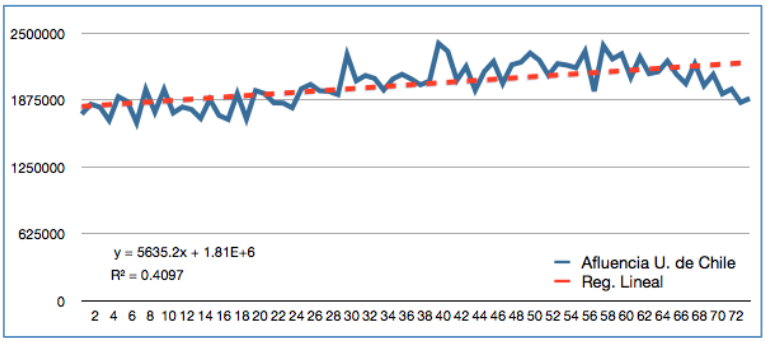
52] R. Ghazali, M. N. Salle N. M. Nawi, "The development of improved back-propagation neural networks algorithm for predicting patients with heart disease," in *First international conference on Information computing and applications (ICICA'10)*, 2010, pp. 317-324.

Anexos

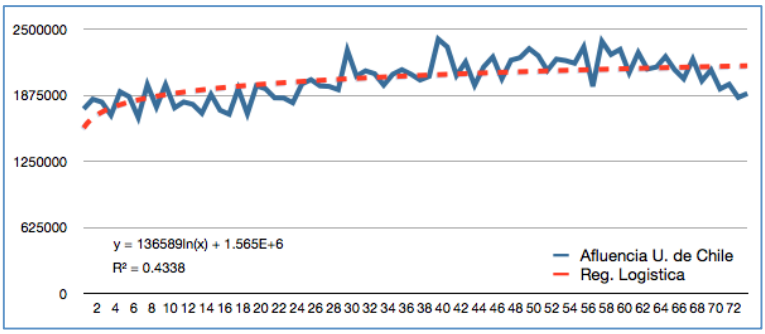
ANEXO A – Estudio Universidad de Chile

A. 1 Modelos Regresivos

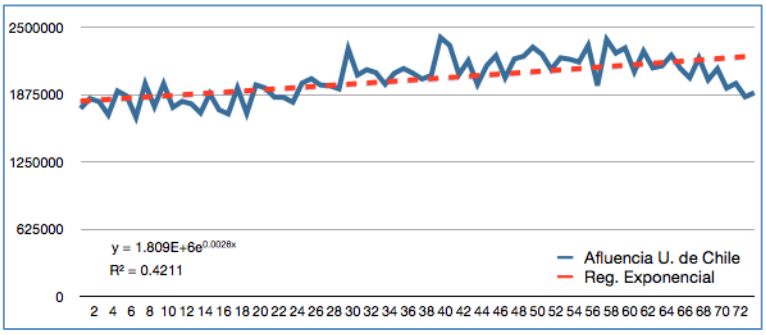
1. Regresión lineal Universidad de Chile.



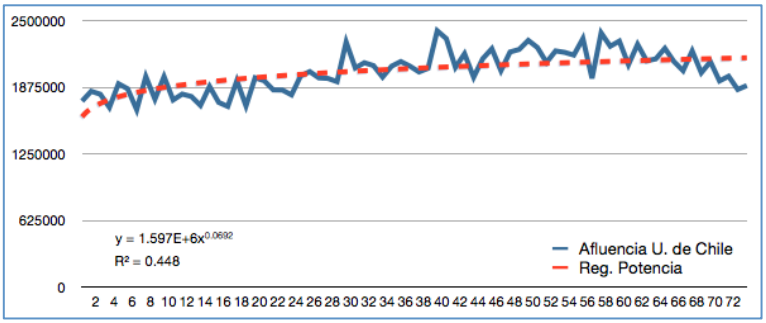
2. Regresión Logística Universidad de Chile.



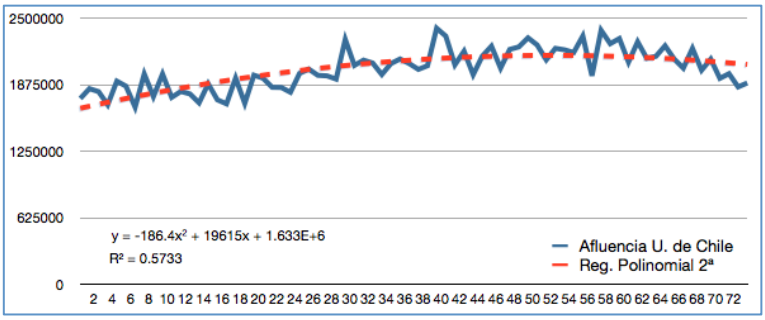
3. Regresión Exponencial Universidad de Chile.



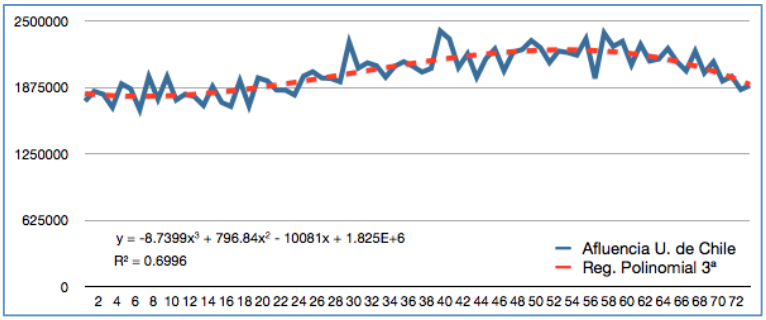
4. Regresión Potencia Universidad de Chile.



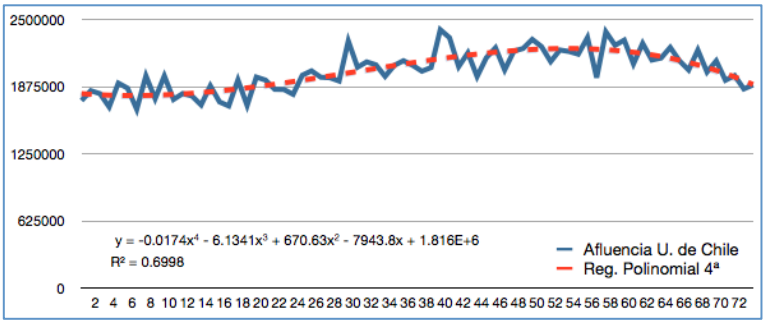
5. Regresión Polinomial segundo grado Universidad de Chile.



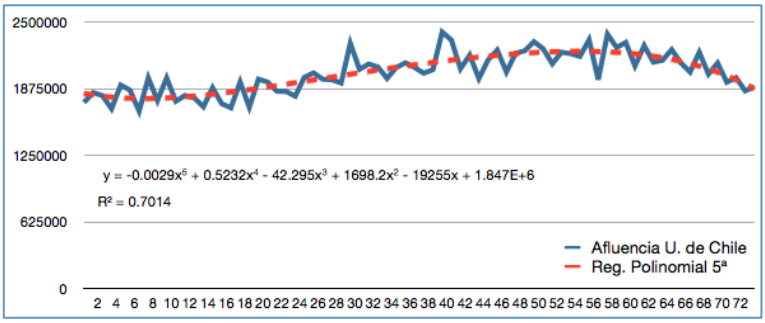
6. Regresión Polinomial tercer grado Universidad de Chile.



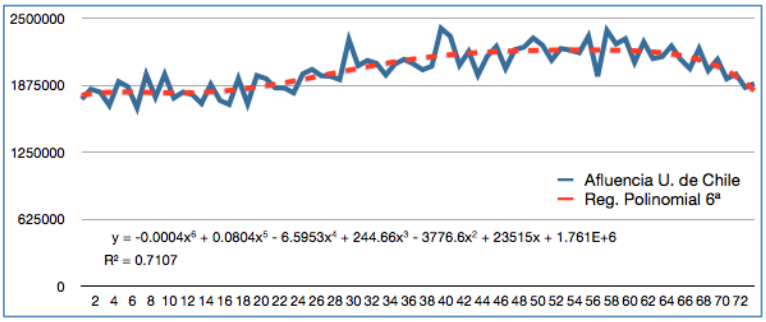
7. Regresión Polinomial cuarto grado Universidad de Chile.



8. Regresión Polinomial quinto grado Universidad de Chile.

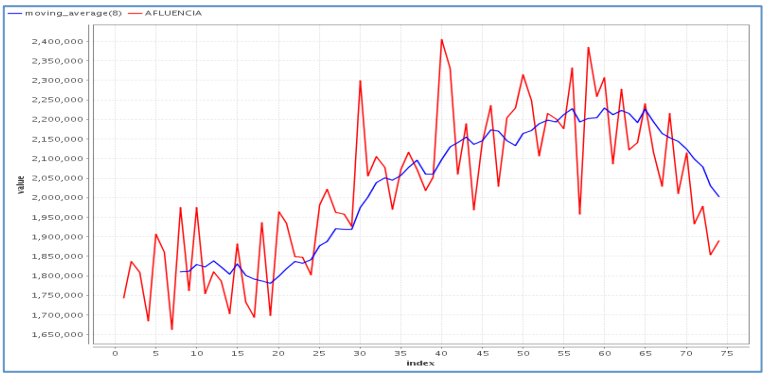


9. Regresión Polinomial sexto grado Universidad de Chile.

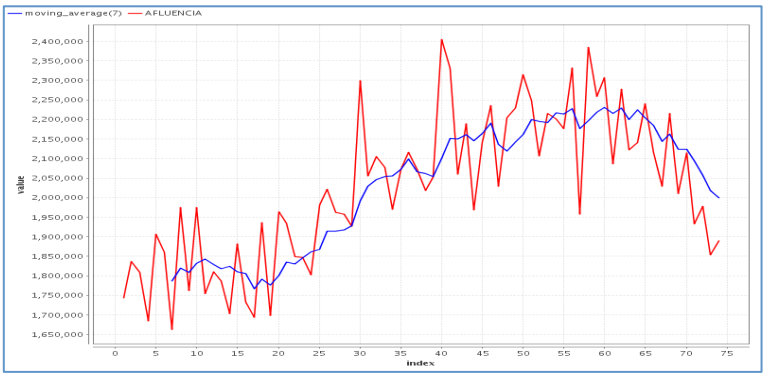


A.2 Modelos de Promedios Móviles

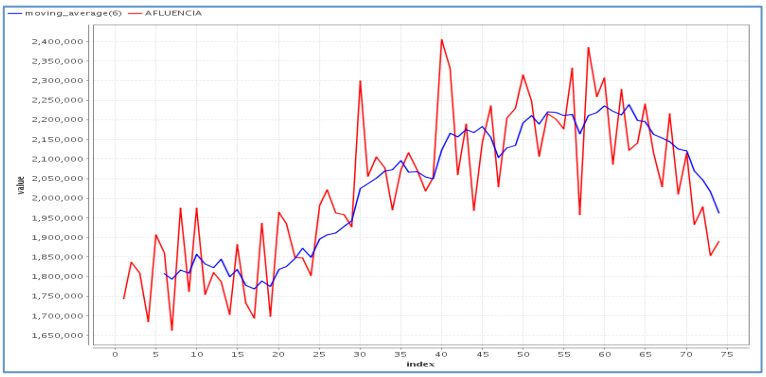
1. Promedio Móvil con ventana de tiempo T=8, Universidad de Chile.



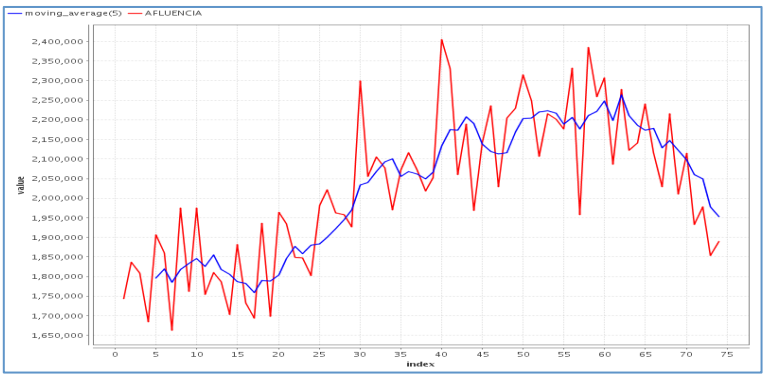
2. Promedio Móvil con ventana de tiempo T=7, Universidad de Chile.



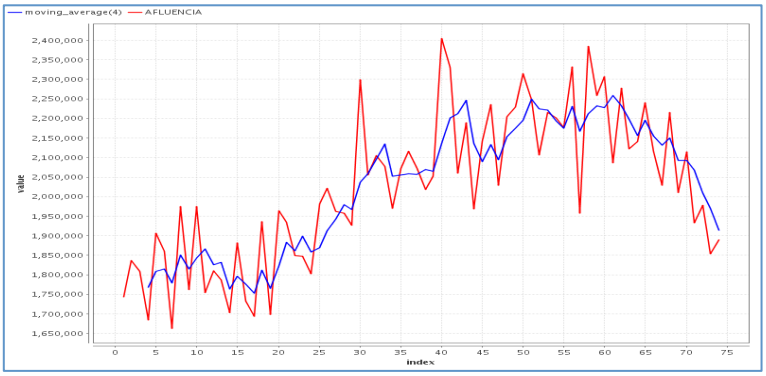
3. Promedio Móvil con ventana de tiempo T=6, Universidad de Chile.



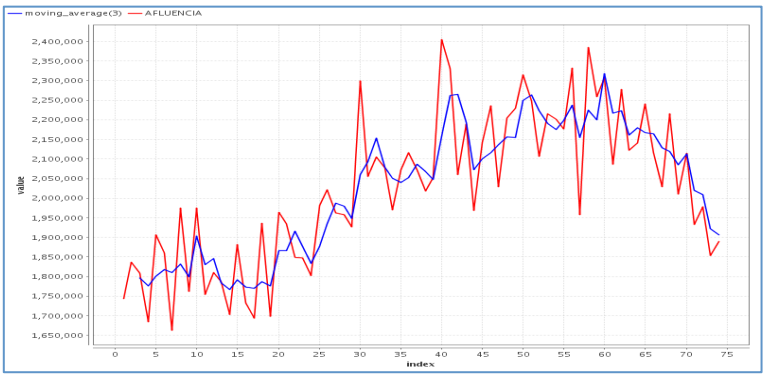
4. Promedio Móvil con ventana de tiempo T=5, Universidad de Chile.



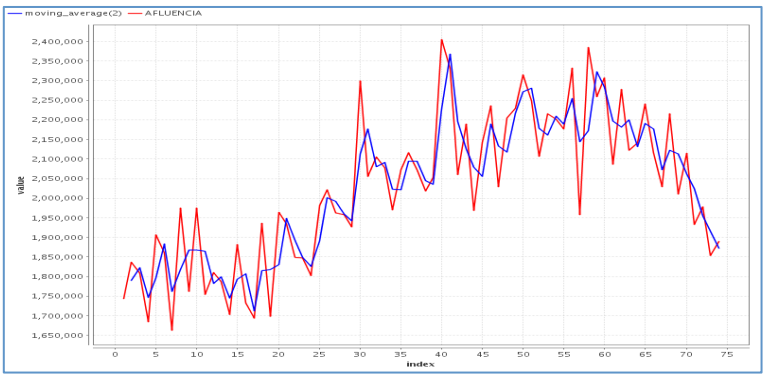
5. Promedio Móvil con ventana de tiempo T=4, Universidad de Chile.



6. Promedio Móvil con ventana de tiempo T=3, Universidad de Chile.

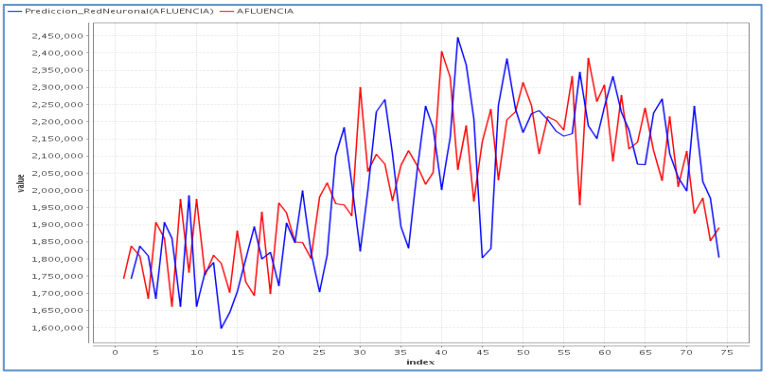


7. Promedio Móvil con ventana de tiempo T=2, Universidad de Chile.

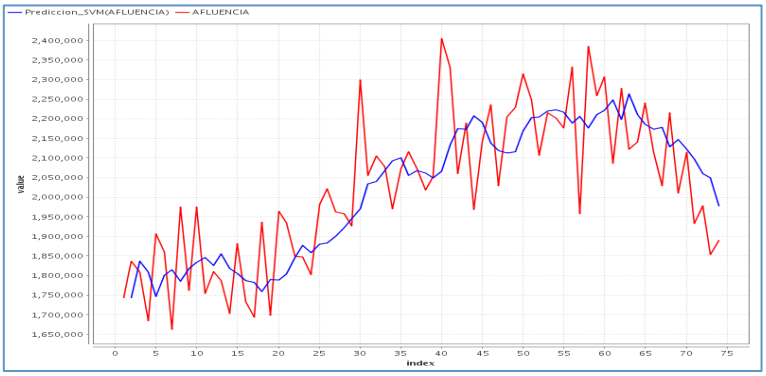


A.3 Redes Neuronales, SVR y Suavización Exponencial

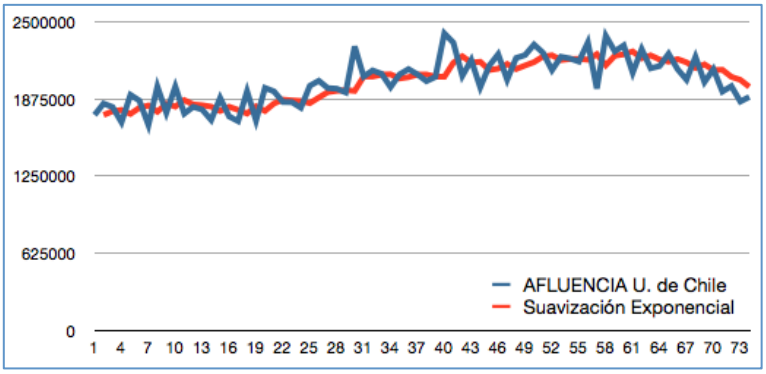
1. Redes Neuronales, Universidad de Chile.



2. SVM, Universidad de Chile.



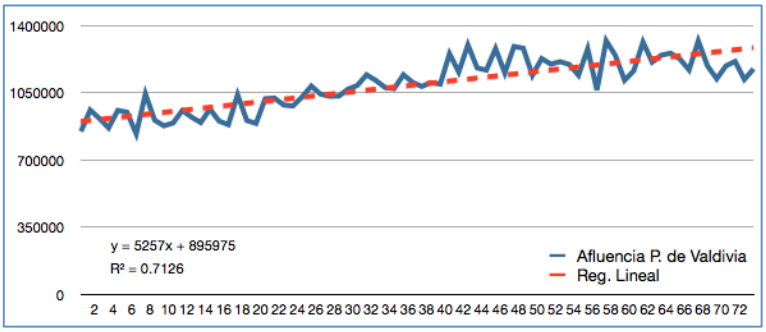
3. Suavización Exponencial, Universidad de Chile.



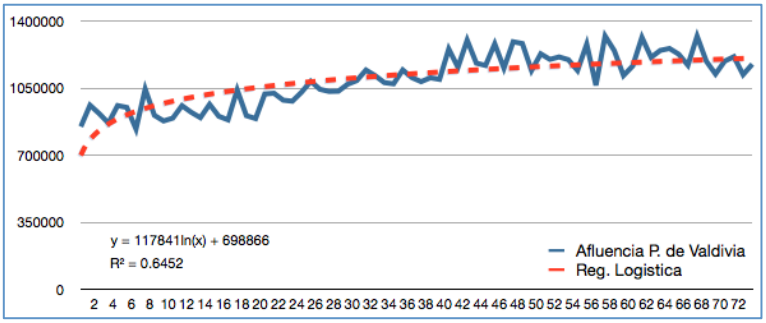
ANEXO B – Estudio Pedro de Valdivia

B.1 Modelos Regresivos

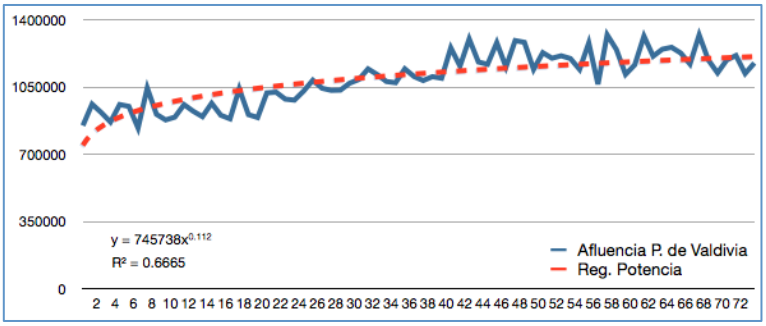
1. Regresión lineal Pedro de Valdivia.



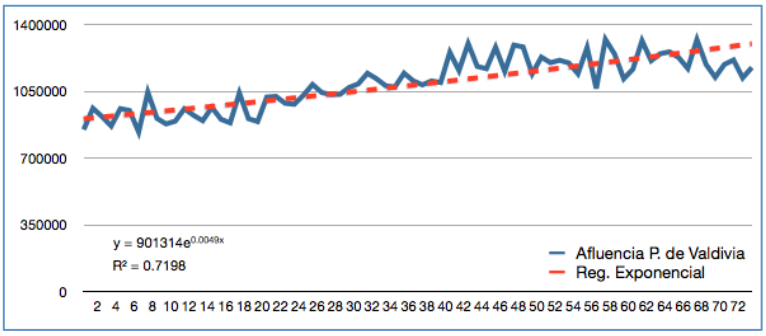
2. Regresión Logística Pedro de Valdivia.



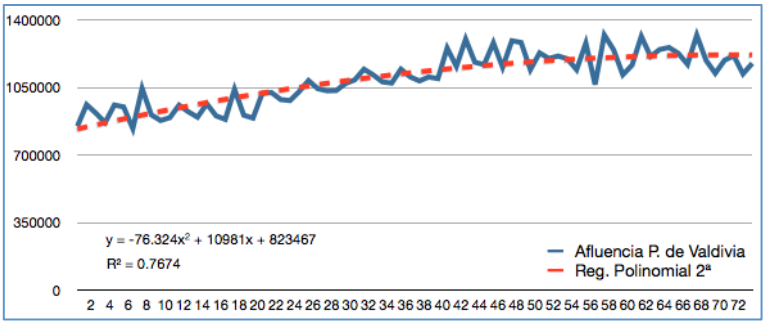
3. Regresión Exponencial Pedro de Valdivia.



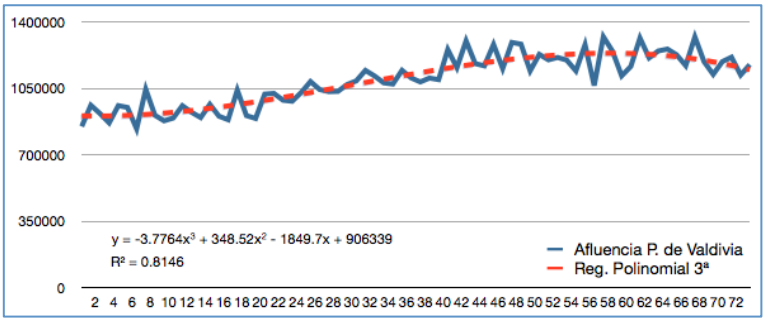
4. Regresión Potencia Pedro de Valdivia.



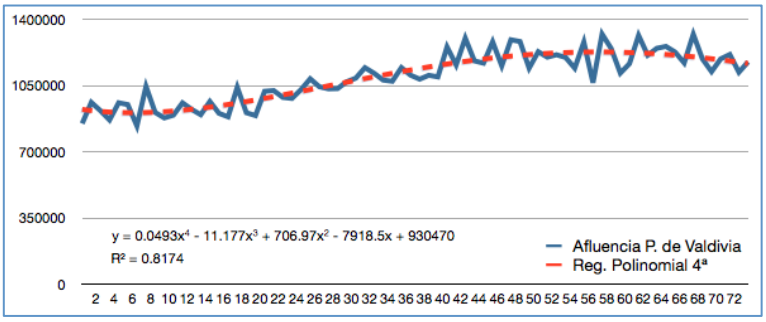
5. Regresión Polinomial segundo grado Pedro de Valdivia.



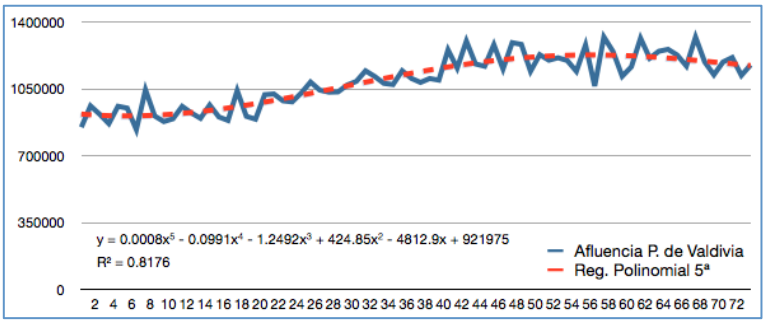
6. Regresión Polinomial tercer grado Pedro de Valdivia.



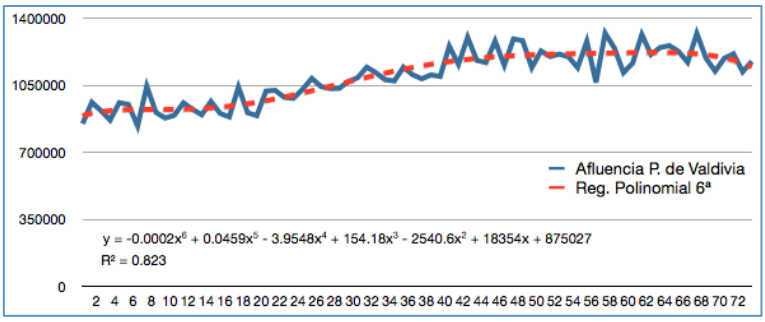
7. Regresión Polinomial cuarto grado Pedro de Valdivia.



8. Regresión Polinomial quinto grado Pedro de Valdivia.

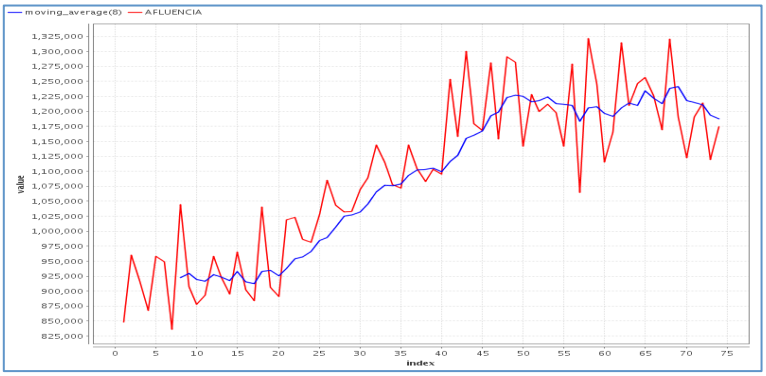


9. Regresión Polinomial sexto grado Pedro de Valdivia.

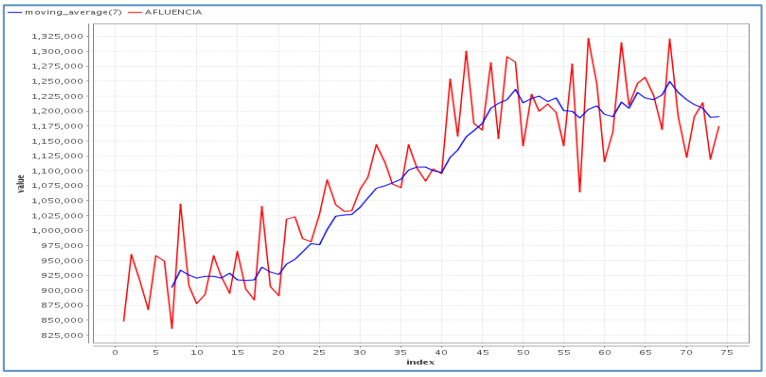


B.2. Modelos de Promedios Móviles

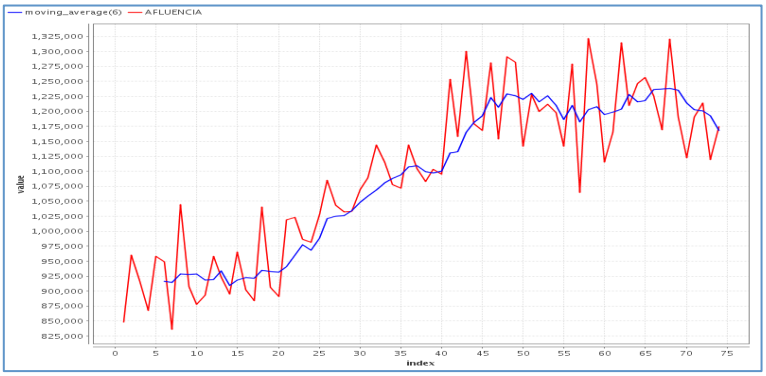
1. Promedio Móvil con ventana de tiempo T=8, Pedro de Valdivia.



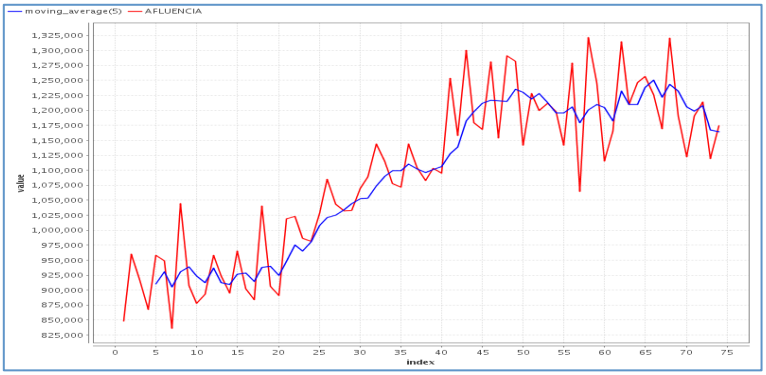
2. Promedio Móvil con ventana de tiempo T=7, Pedro de Valdivia.



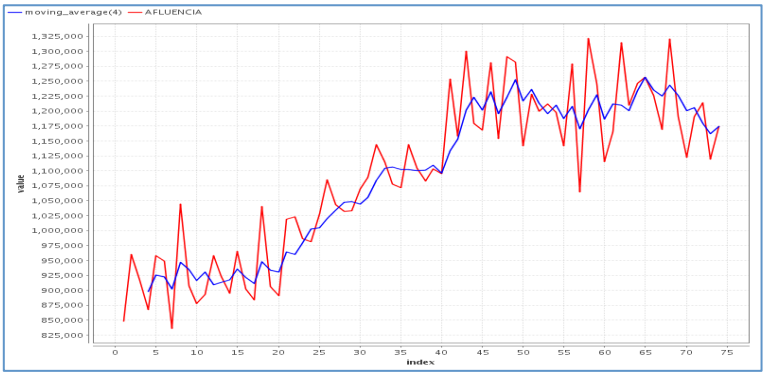
3. Promedio Móvil con ventana de tiempo T=6, Pedro de Valdivia.



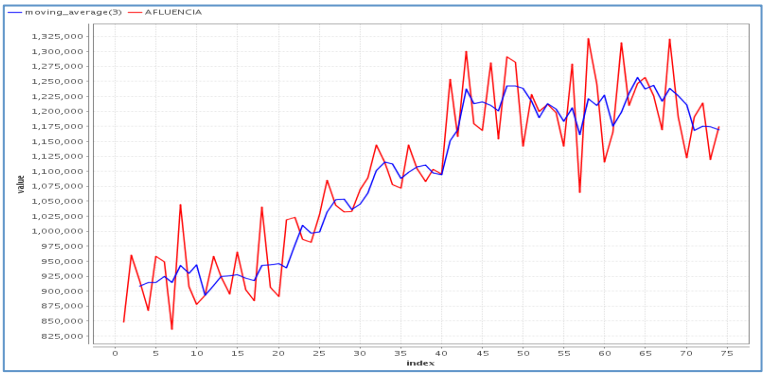
4. Promedio Móvil con ventana de tiempo T=5, Pedro de Valdivia.



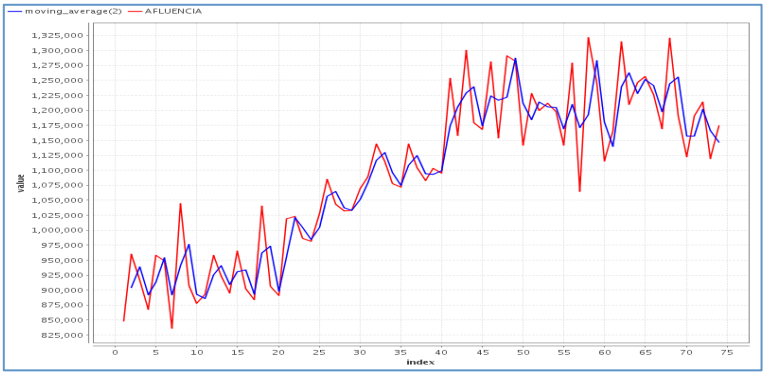
5. Promedio Móvil con ventana de tiempo T=4, Pedro de Valdivia.



6. Promedio Móvil con ventana de tiempo T=3, Pedro de Valdivia.

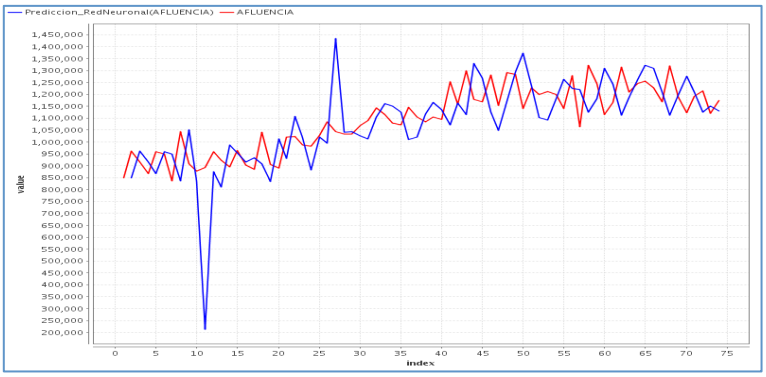


7. Promedio Móvil con ventana de tiempo T=2, Pedro de Valdivia.

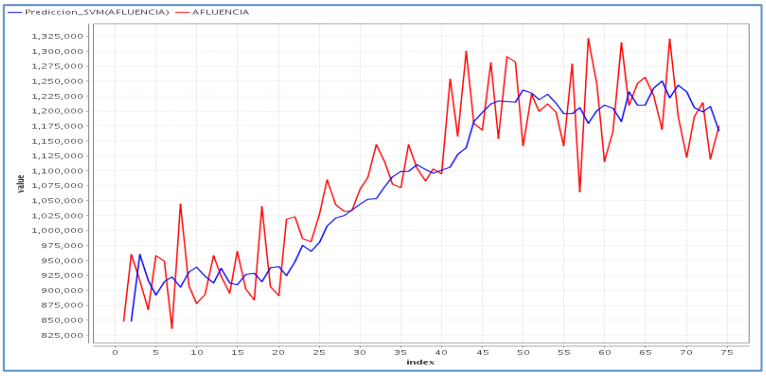


B.3 Redes Neuronales, SVM y Suavización Exponencial

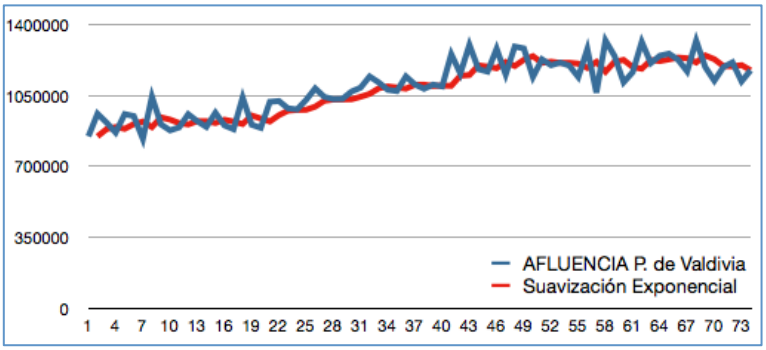
4. Redes Neuronales, Pedro de Valdivia.



5. SVM, Pedro de Valdivia.



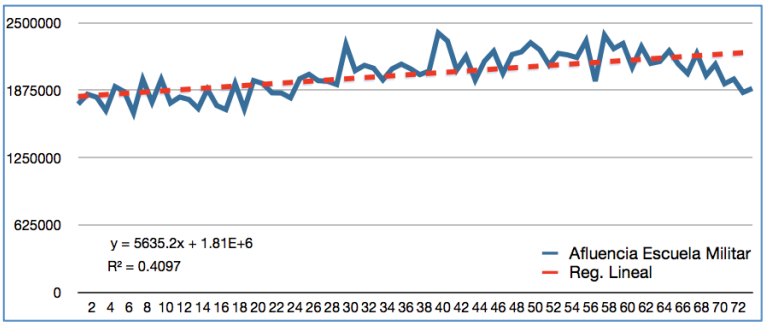
6. Suavización Exponencial, Pedro de Valdivia.



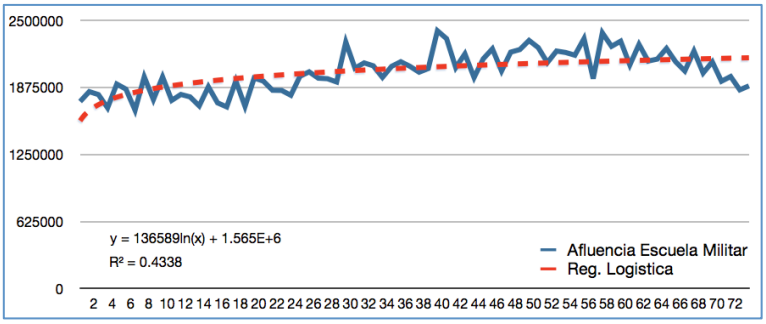
ANEXO C – Estudio Escuela Militar

C.1 Modelos Regresivos

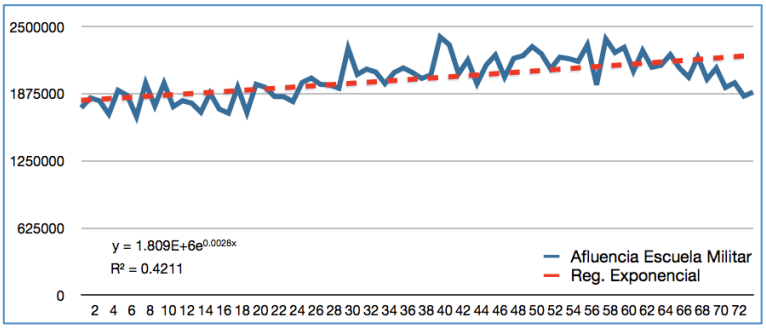
1. Regresión lineal Escuela Militar.



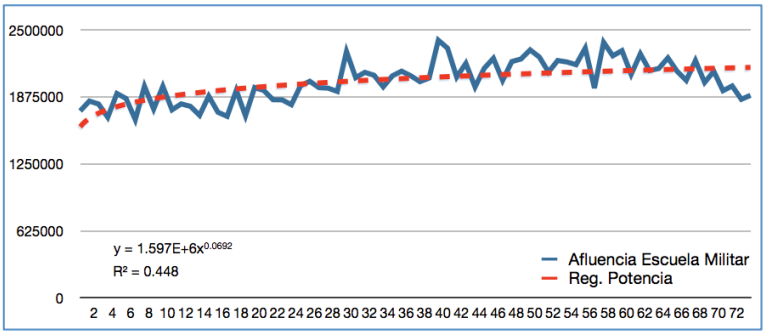
2. Regresión Logística Escuela Militar.



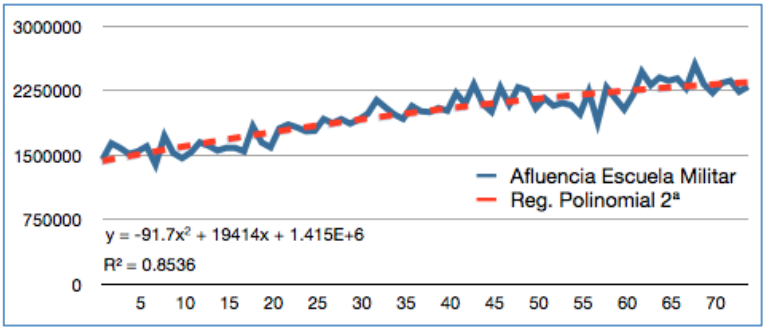
3. Regresión Exponencial Escuela Militar.



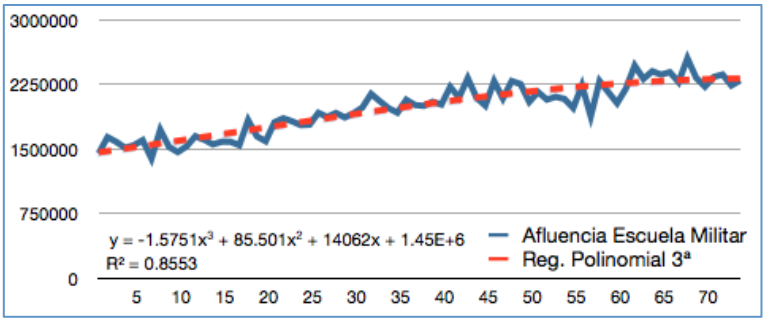
4. Regresión Potencia Escuela Militar.



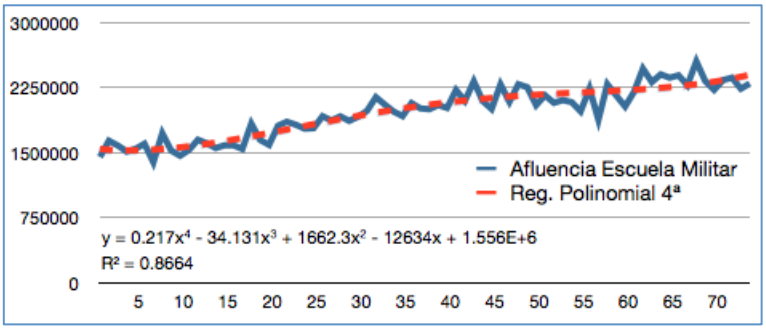
5. Regresión Polinomial segundo grado Escuela Militar.



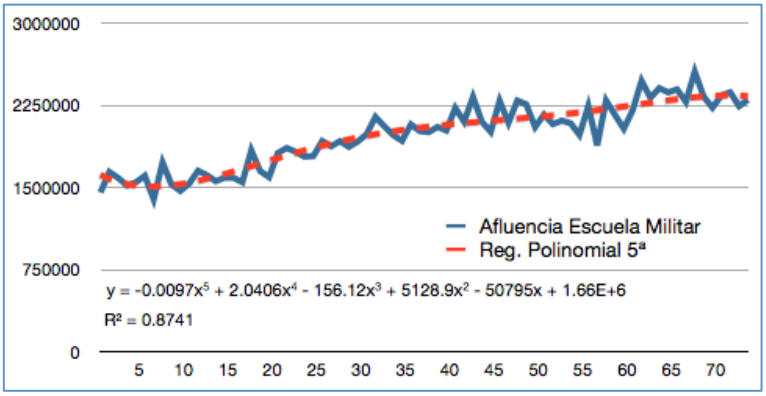
6. Regresión Polinomial tercer grado Escuela Militar.



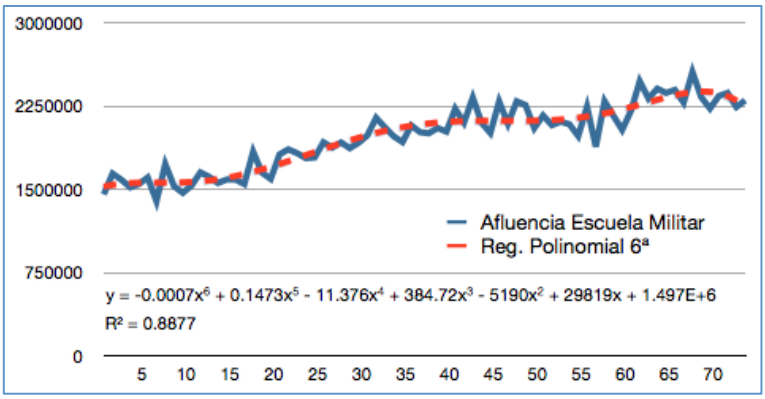
7. Regresión Polinomial cuarto grado Escuela Militar.



8. Regresión Polinomial quinto grado Escuela Militar.

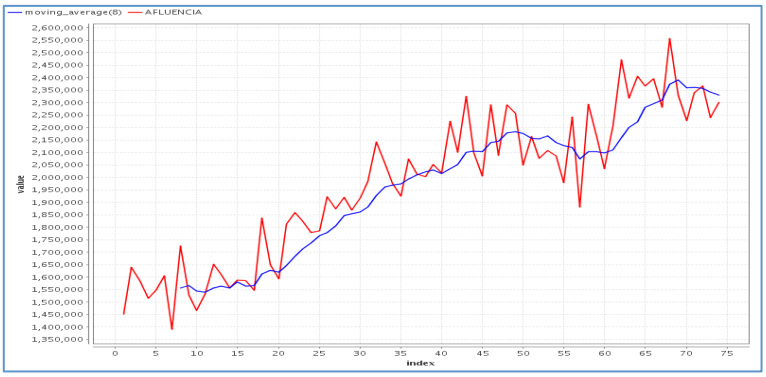


9. Regresión Polinomial sexto grado Escuela Militar.

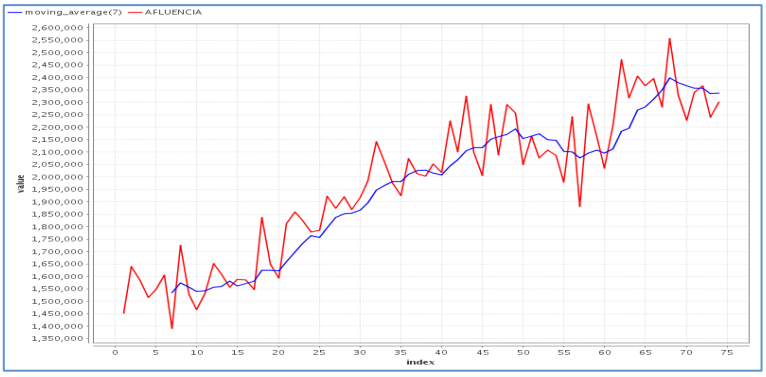


C.2 Modelos de Promedios Móviles

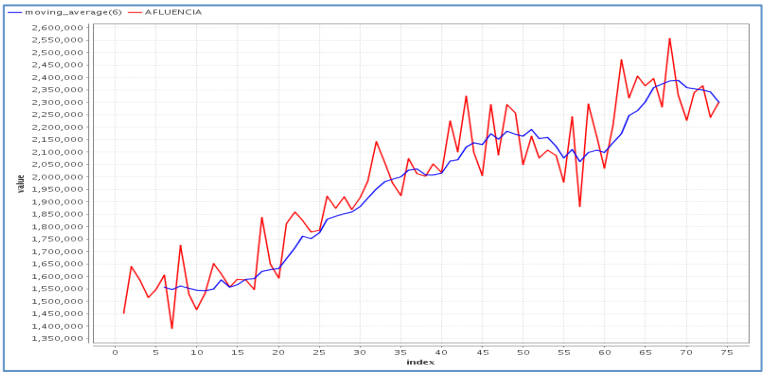
1. Promedio Móvil con ventana de tiempo T=8, Escuela Militar.



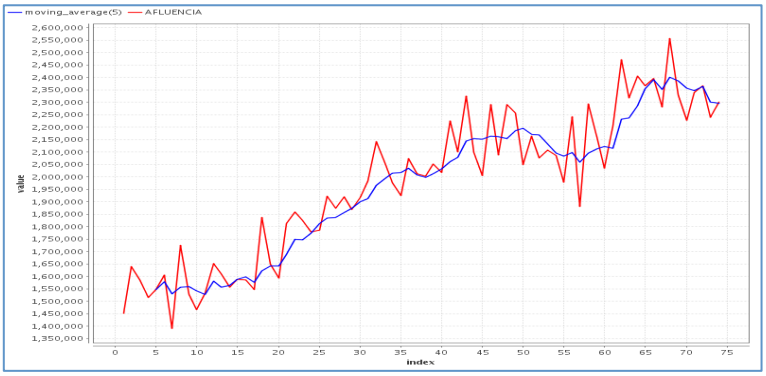
2. Promedio Móvil con ventana de tiempo T=7, Escuela Militar.



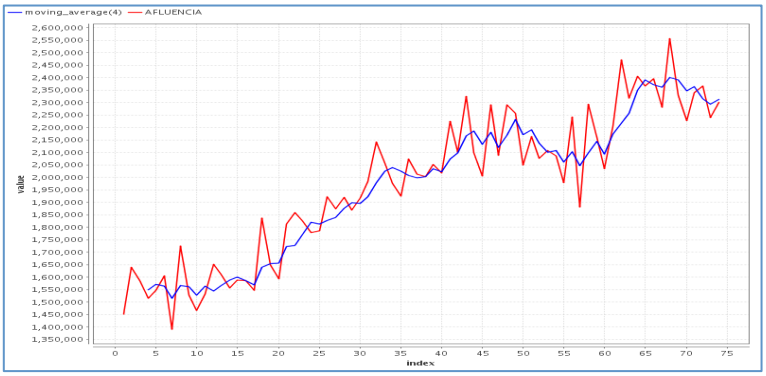
3. Promedio Móvil con ventana de tiempo T=6, Escuela Militar.



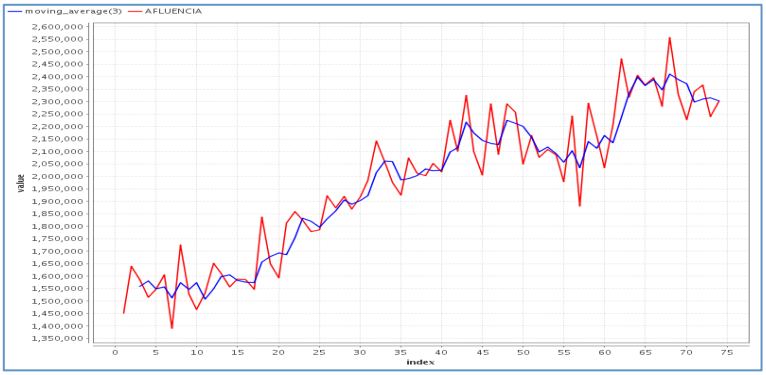
4. Promedio Móvil con ventana de tiempo T=5, Escuela Militar.



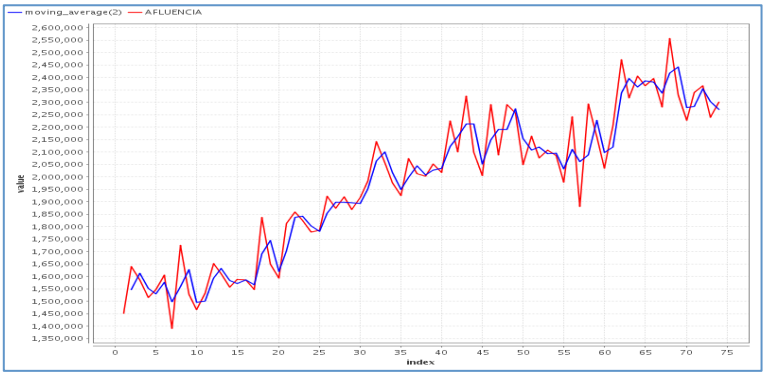
5. Promedio Móvil con ventana de tiempo T=4, Escuela Militar.



6. Promedio Móvil con ventana de tiempo T=3, Escuela Militar.

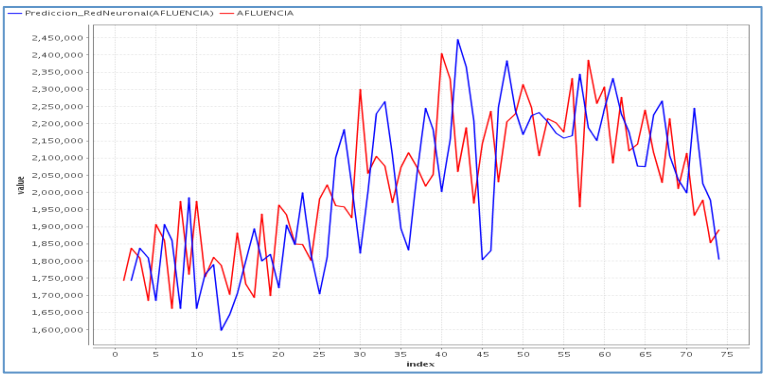


7. Promedio Móvil con ventana de tiempo T=2, Escuela Militar.

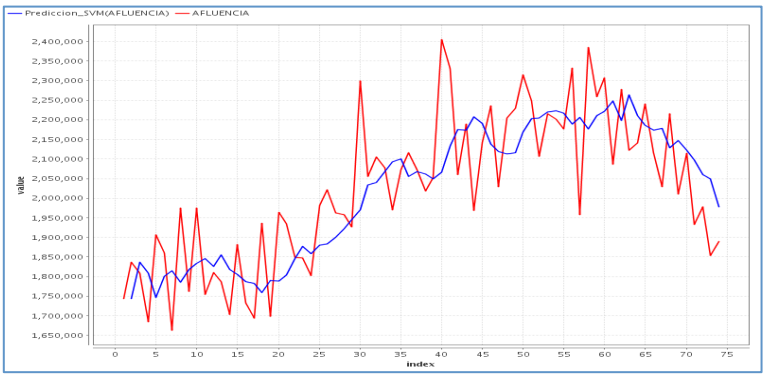


C.3 Redes Neuronales, SVM, Suavización Exponencial

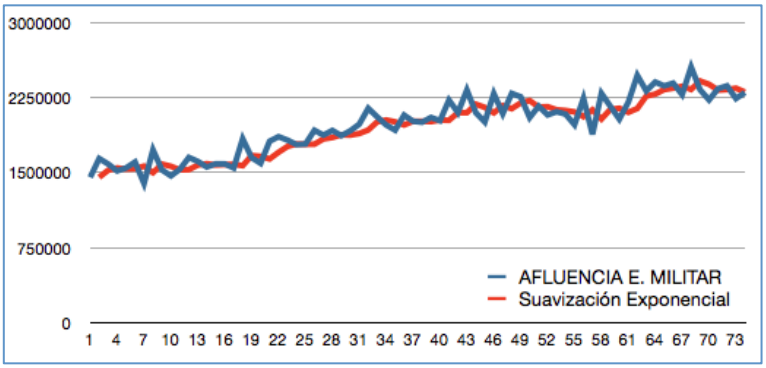
7. Redes Neuronales, Escuela Militar.



8. SVM, Escuela Militar.



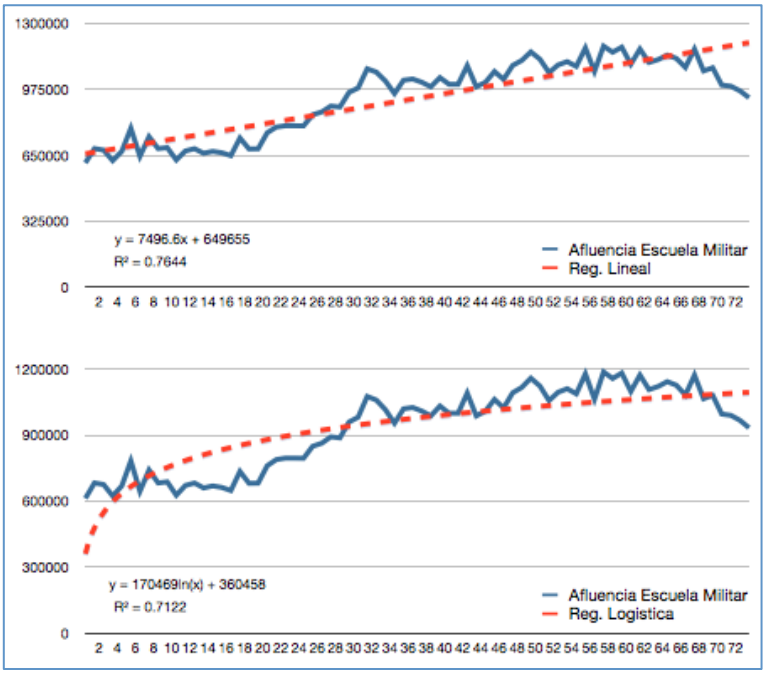
9. Suavización Exponencial, Escuela Militar.



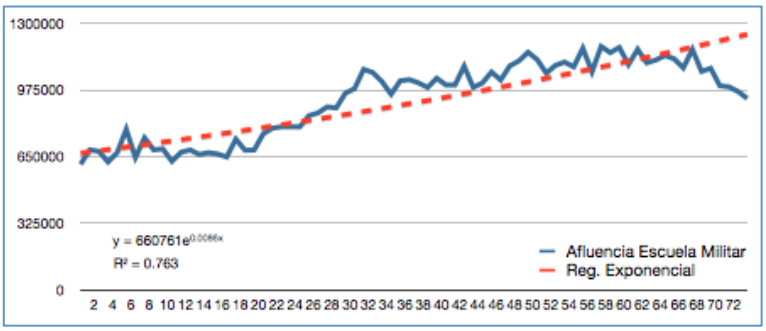
ANEXO D – Estudio San Pablo

D.1 Modelos Regresivos

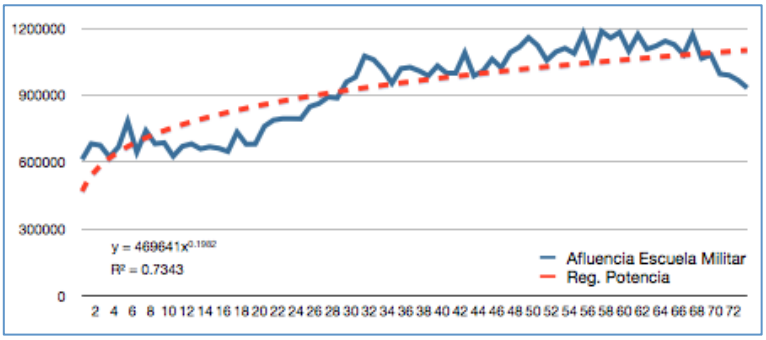
1. Regresión Lineal San Pablo.
2. Regresión Logística San Pablo.



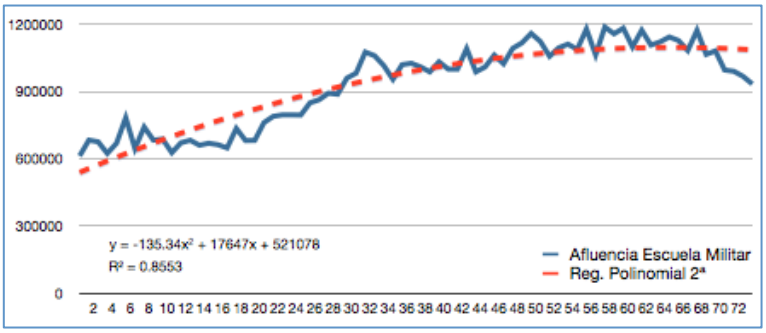
3. Regresión Exponencial San Pablo.



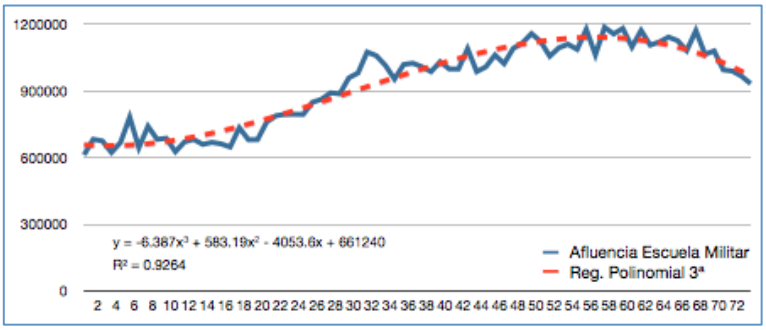
4. Regresión Potencia San Pablo.



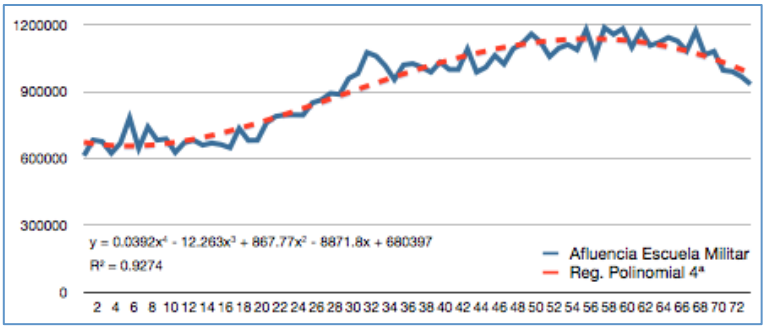
5. Regresión Polinomial segundo grado San Pablo.



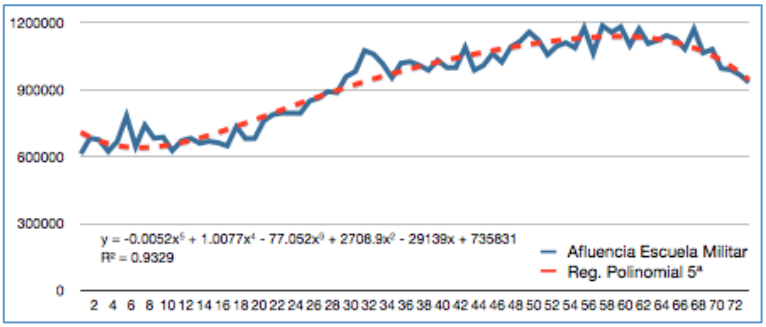
6. Regresión Polinomial tercer grado San Pablo.



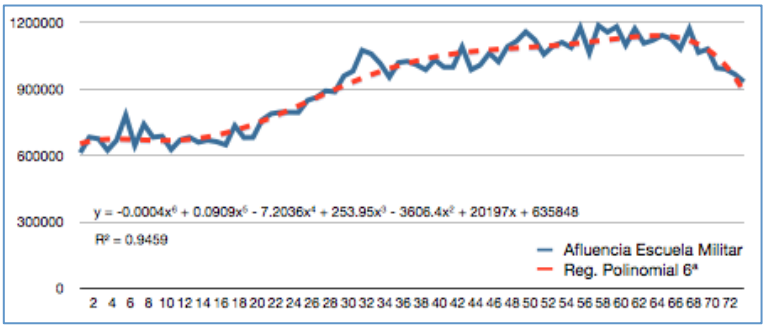
7. Regresión Polinomial cuarto grado San Pablo.



8. Regresión Polinomial quinto grado San Pablo.

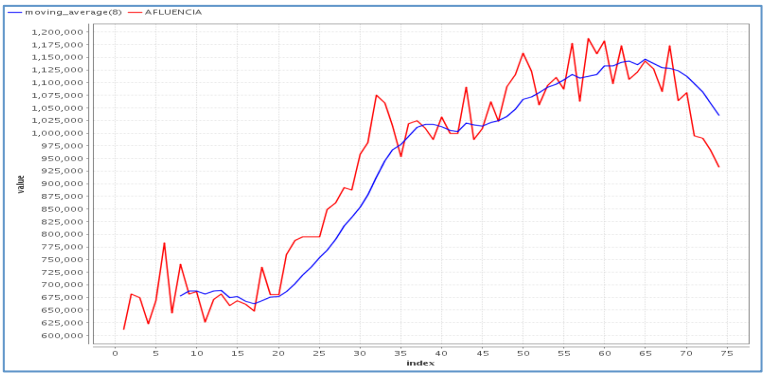


9. Regresión Polinomial sexto grado San Pablo.

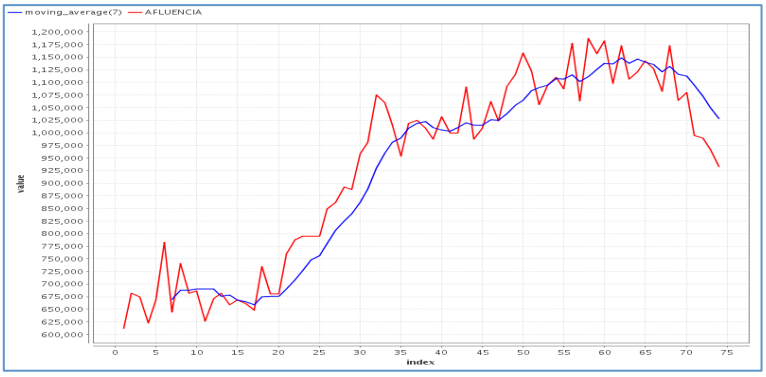


D.2 Modelos de Promedios Móviles

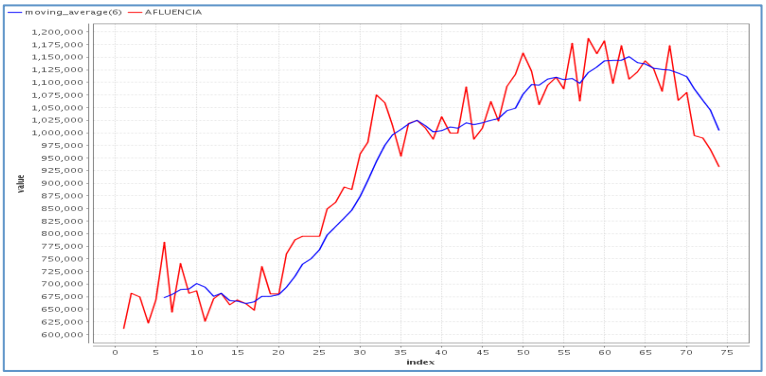
1. Promedio Móvil con ventana de tiempo T=8, San Pablo.



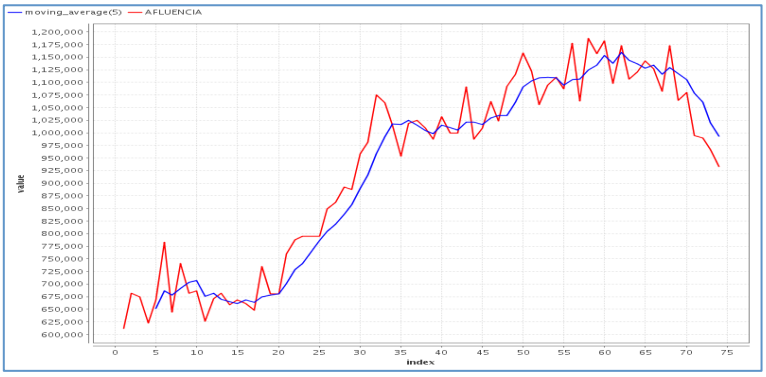
2. Promedio Móvil con ventana de tiempo T=7, San Pablo.



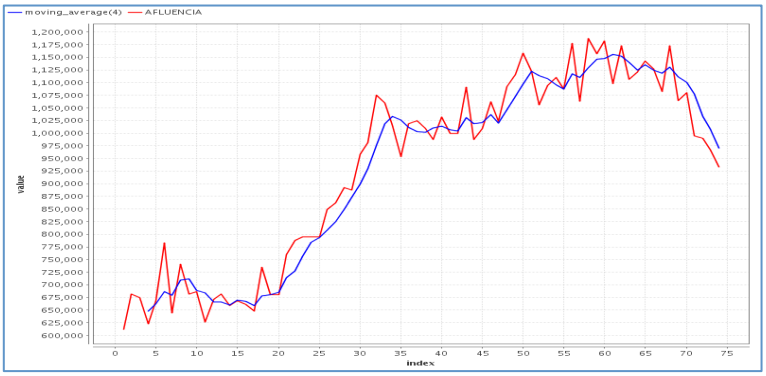
3. Promedio Móvil con ventana de tiempo T=6, San Pablo.



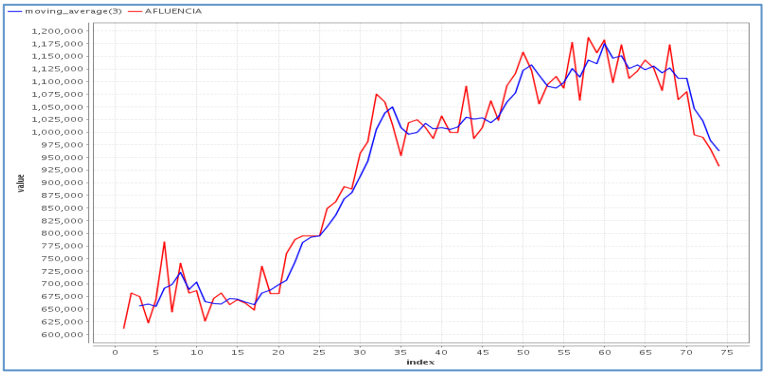
4. Promedio Móvil con ventana de tiempo T=5, San Pablo.



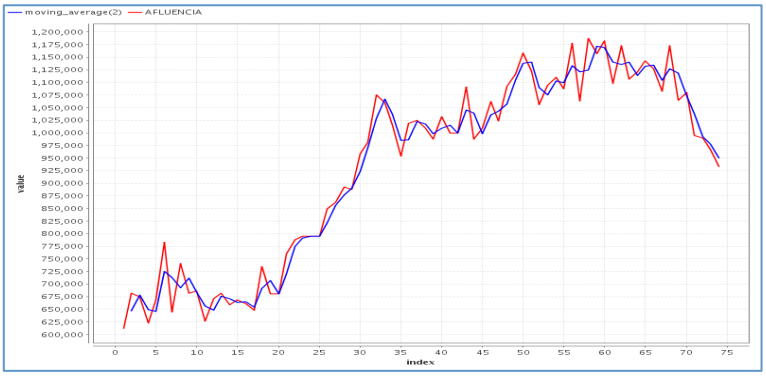
5. Promedio Móvil con ventana de tiempo T=4, San Pablo.



6. Promedio Móvil con ventana de tiempo T=3, San Pablo.

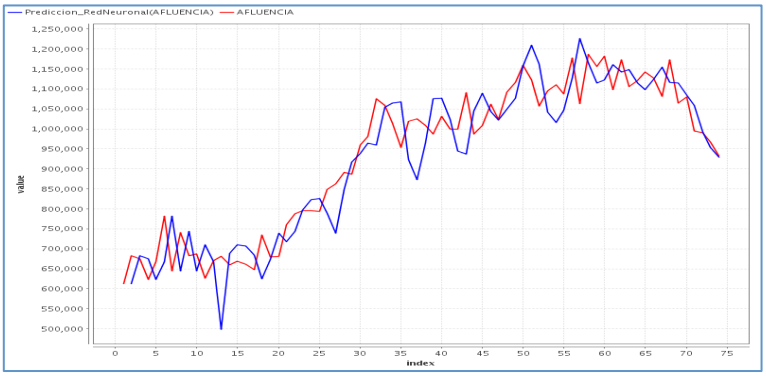


7. Promedio Móvil con ventana de tiempo T=2, San Pablo.

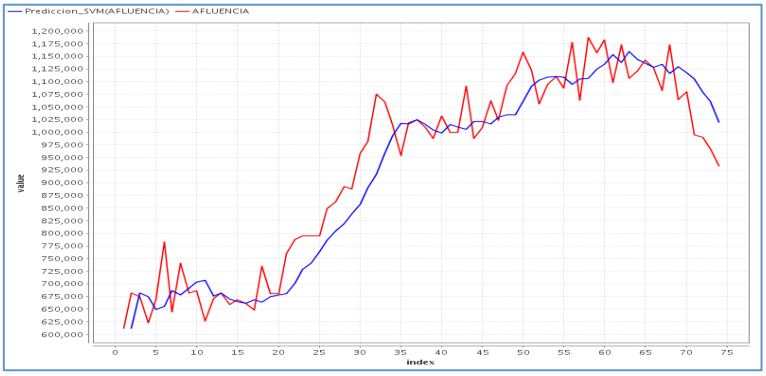


D.3 Redes Neuronales, SVM, Suavización Exponencial

10. Redes Neuronales, San Pablo.



11. SVM, San Pablo.



12. Suavización Exponencial, San Pablo.