

CENTRO DE INVESTIGACIÓN EN CIENCIAS DE INFORMACIÓN GEOESPACIAL  
 SECRETARIA GENERAL  
 PLANEACIÓN ESAPCIAL  
 REPORTE No 1

Dr. Rodrigo Frías López  
Responsable de Servicio Social  
Licenciatura en Tecnología  
Centro de Física Aplicada y Tecnología Avanzada

P r e s e n t e.

Me permito informarle que el alumno **ARIEL CERÓN GONZÁLEZ** con número de cuenta **110001007**, de la Licenciatura en Tecnología que se imparte en el Campus UNAM-Juriquilla del Centro de Física Aplicada y Tecnología Avanzada, que realiza su Servicio Social en esta Dependencia en el Programa: **Software para sistematizar validación de modelos de reconstrucción de Series de Tiempo con aplicación al estudio de las concentraciones de contaminantes PM 2.5,** con clave **202-61/17-3022**, en el periodo comprendido del 17 – febrero al 17 – agosto del 2020, ha desarrollado la siguiente actividad en el primer bimestre:

**Informe de la Actividad:**

*Se realizó una investigación teórica de los siguientes puntos:*

* *Series de tiempo*
* *Reconstrucción*
* *Modelos de reconstrucción*
* *Validación*
* *Validación cruzada*
* *Error*

*Además se plantea una propuesta para implementar un algoritmo de reconstrucció.*

Sin otro particular por el momento, reciba un cordial saludo.

**A T E N T A M E N T E**

**CAMPUS UNAM-Juriquilla, Qro a XX de XXXX de 2020**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Responsable del Servicio Social**

**COMPLEMENTO AL REPORTE**

**ABSTRACT.**

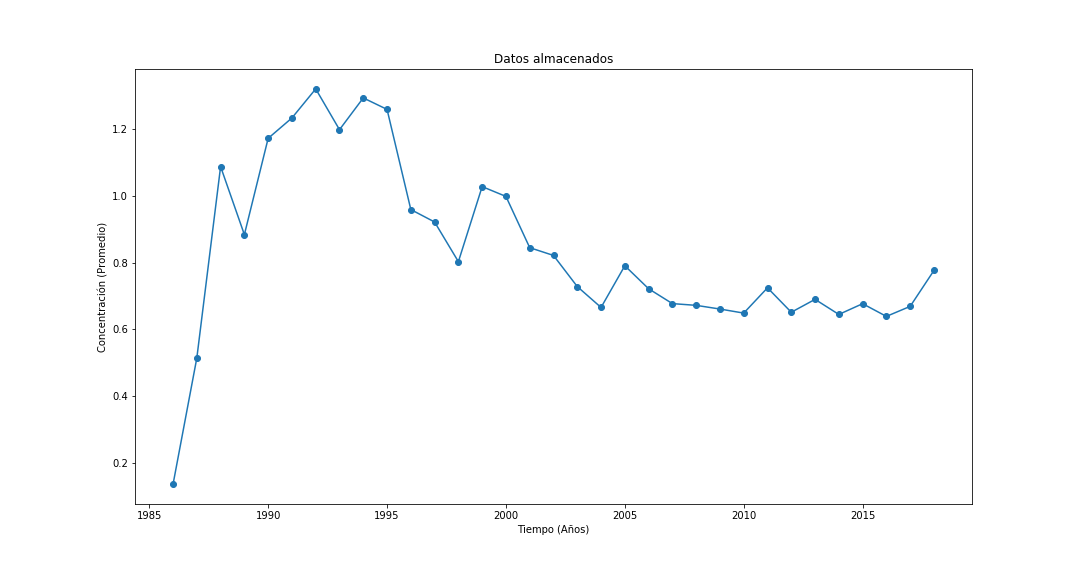
El estudio de series de tiempo es un área interdisciplinar donde interactúan diferentes áreas (pero no se limita a estas) como estadística, ciencias de la computación, sistemas dinámicos y teoría de la información. El estudio de series de tiempo tiene aplicación diversa en áreas como economía, ciencias de la tierra, química atmosférica, optimización de aprovechamiento de recursos naturales renovables y no renovables como agua y viento. La principal utilidad del estudio de modelos de series de tiempo es la posibilidad de predecir con cierta precisión a la vez que se minimiza la incertidumbre de los valores de ciertas variables. Sin embargo, para poder estimar estos modelos de predicción, en especial los modelos estadísticos de la familia Box-Jenkins y Exponential Smoothing, se parte del supuesto que la serie de tiempo no contiene datos faltantes, datos atípicos, o erróneos. Este reporte revisa modelos de validación cruzada para modelos de predicción, y basado en estos, la propuesta de validación cruzada para medir la capacidad de reconstrucción de algún modelo.

**INTRODUCCIÓN.**

**Series de Tiempo.**

Una serie de tiempo es una secuencia de tiempo orientada o una secuencia cronológica de observaciones sobre una variable de interés (Montgomery, Jennings, & Kulahci, 2015)

Figura 1. Series de tiempo de las concentraciones de PM 2.5 en la estación MERCEDES desde 1985 hasta 2018.



Muchas instituciones públicas y privadas tienen almacenados series de tiempo que reflejan la evolución de alguna variable en específico, esto hace de las series de tiempo una de las principales fuentes de consulta para la toma de decisiones o para la planeación de diferentes actividades.

Las series de tiempo pueden reflejar datos a corto, mediano y largo plazo; cada una con un conjunto de problemas propios de su escala. Para las series de tiempo a largo plazo se presentan problemas en la calidad de los datos, la desaparición de información, la perdida de formato y en ocasiones la nula existencia de información por errores en la forma de adquisición de la información. Lo anterior hace que la información disponible para la predicción no sea siempre de la mejor calidad y sea necesario una limpieza previa para mejorar los resultados.

Aunque la mayor parte del tiempo y la mayoría de la gente presupone que un pronóstico es sólo un número, la realidad es que se pueden obtener un conjunto de herramientas numéricas más poderosas que un solo número, como puede ser un intervalo de predicción, un punto estimado, errores de estimación, entre otras herramientas estadísticas.

Para obtener esas herramientas es necesario seguir un proceso de predicción que toma los siguientes pasos:

1. Definición del problema
2. Colección de datos
3. Análisis de datos
4. Modelo de selección y ajuste
5. Modelo de validación
6. Implementación del modelo desarrollado
7. Monitoreo del modelo de predicción.

La recolección de datos, como se ha mencionado, es importante para un ben análisis. Recolectar datos significa obtener datos históricos relevantes de la variable que será reconstruida. Sin embargo la colección y almacenamiento de la información, así como los métodos y sistemas cambian en el tiempo y no toda la información histórica es útil para el el problema actual. Además los sistemas y métodos presentan errores de funcionamiento o, en ocasiones, no se tienen los elementos mínimos necesarios para hacer funcionar los sistemas. Esto hace necesario trabajar con series de tiempo incompletas.

La base del trabajo se encuentra en la serie de datos que se puedan obtener para análisis. Los datos pueden tener diferentes dimensiones de calidad, las cinco más importantes son: exactitud, completitud, consistencia y representititud.

La exactitud refleja qué tan cercano se encuentra un valor a comparación de su valor real.

Completitud se refiere a que los datos son completos, que no hay datos perdidos o valores atípicos

La consistencia se refiere a qué tan cerca en formato se encuentra los datos almacenados hablando en contenido, significado y estrucutra.

El contar con buenos datos es importante para el proceso de predicción, pues la falta de datos, una mal registro o la nula contingencia de los datos puede hacer imposible la predicción o reflejar resultados fuera de la realidad.

La limpieza de los datos es el proceso mediante el cual se detectan errores potenciales, datos perdidos, valores inusuales u otras inconsistencias y corregirlos. Datos limpios pueden mejorar en gran medida el proceso de pronóstico.

Cuando se tienen un conjunto de datos completos, la necesidad de obtener nuevos valores en función de los que ya se tienen, es lo que encamina a los diferentes profesionales del área a desarrollar métodos de predicción que nos adelante en el futuro. Como se ha venido diciendo, sin embargo, si se tiene en cuenta que la serie de tiempo puede tener partes incompletas, el modelo de predicción puede ser adaptado para ser un modelo de reconstrucción y con ello poder obtener una serie de datos completa que permita realizar nuevos análisis.

Existen al manos dos enfoques ampliamente usados los estadísticos y los computacionales:

Los modelos estadísticos hacen uso de variables estadísticas y modelos de regresión para aproximar o estimar un valor, entre los modelos más usados se enceuntra:

- Exponential smoothing based models (modelo de suavisado exponenciañ) (Holt-Winters, 1960). La idea es usar promedios ponderados de observaciones pasadas, con pesos decrecientes exponencialmente tan pronto como las observaciones envejezcan, y también describen series de tiempo desde el punto de vista de tendencia y estacionalidad.

La idea del suavizado exponencial es suavizar la serie original de la forma en que se mueve promedio hace y usar la serie suavizada para pronosticar valores futuros de la variable de interés. Sin embargo, en el suavizado exponencial, queremos permitir los valores más recientes. de la serie para tener mayor influencia en el pronóstico de valores futuros que los más distantes observaciones [9]

el alisado es probablemente la clase de procedimientos ampliamente utilizados para alisar

series de tiempo discretas para pronosticar el futuro inmediato. Esta popularidad se puede atribuir

a su simplicidad, su eficiencia computacional, la facilidad de ajustar su capacidad de respuesta

a los cambios en el proceso que se pronostica, y su precisión razonable

- \*Autoreggresive-Moving Average Based Models (Modelos de media movil autoregresiva) (Box-Jenkins, 1974), describing the dynamics with autocorrelations in the data.\* el modelo ARMA es una herramienta para entender y, aún más, para predecir futuros valores de la serie. El modelo está formado por dos partes, una parte autorregresiva (AR) y otra de media móvil (MA)

- \*BATS (\*\*B\*\*ox-Cox transformation, \*\*A\*\*RMA errors, \*\*T\*\*rend and \*\*S\*\*easonal Components)) combines these models complementary. , published in 2011 by A. M. de Livera R. J. Hyndman and R. D. Snyder.\*

Los modelos computacionales aprovechan el poder computacional para realizar operaciones iterativas que permiten comprobar modelos experimentales y conceptos teóricos.

- Redes neuronales.

Las redes neuronales son modelos matemáticos que copian el funcionamiento de una neurona para el tratamiento de información, con ello se obtiene una neurona artificial que es una unidad procesadora con elementos funionales (entradas, funciones, activadores y salidas). Exsiten diferentes modelos que logran obtener diferentes resultados, en función de la arquitectura que se use, el número de capas, el modelo con el que se trabaje, la función de activación, etc.[10]

- Vecinos cercanos.

El algoritmo k-NN es uno de los algoritmos de clasificación más conocidos En este algoritmo usaremos un conjunto de ejemplos ya clasificados a los que llamaremos conjunto de entrenamiento o train para clasificar los nuevos ejemplos El algoritmo se llama así k-NN (k-Nearest Neighbors) porque clasifica cada nuevo ejemplo calculando la distancia de ese ejemplo con todos los del conjunto de train El principal problema del algoritmo k-NN es encontrar el valor de k con el que obtengamos un mayor rendimiento al clasificar, generalmente se utiliza una técnica conocida como cross validation. [11]

- Redes bayesianas.

Las redes bayesianas modelan un fen´omeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Adem´as, pueden dar informaci´on interesante en cuanto a cmo se relacionan las variables del dominio, las cuales pueden ser interpretadas en ocasiones como relaciones de causa–efecto [12]

**### Predicción (terminar basarse en la introducción del libro de Montgomery).**

**Un pronostico es una predicción de algún evento o eventos futuros. Pronosticar es un problema importante que se le presenta a una gran variedad de campos tanto cientificos, gubernamentales e industriales. Muchos de los problemas de predicción involucra el uso de series de tiempo. La importancia en la predicción de futuros eventos radica en su participación sobre la toma de decisiones en procesos y planeaciones.**

**Los modelos cuantitativos usados para predecir hacen uso de los patrones presentes en los datos para expresar una relación estadística entre valores anteriores y actuales de la variable y así mostrar patrones de la variable en el futuro.**

**Un proceso es una serie de actividades conectadas que transforman una o más entradas en una o más salidas. El conjunto de actividades para el proceso de prediccións e enumera a continuación**

**1. Definición del problema**

**2. Colección de datos**

**3. Analisis de datos**

**4. Selección de modelo y ajuste**

**5. Validación del modelo**

**6. Predicción**

**7. Monitoreo del proceso**

**Algunos de los objetivos principales del análisis de las series de tiempo son: describir, clasificar, explicar, predecir y controlar.**

**### Limitantes del proceso de predicción (terminar).**

**\*Mencionar que funciona siempre y cuando la calidad de los datos sea buena (no datos atípicos, no datos faltantes, datos calibrados).\***

**### Modelo de reconstrucción se relaciona con el de predicción (terminar).**

**\*Describir el proceso de reconstrucción y como se conecta con el proceso de predicción\***

**La base del trabajo se encuentra en la serie de datos que se puedan obtener para análisis. Los datos pueden tener diferentes dimensiones de calidad, las cinco más importantes son: exactitud, completitud, consistencia y representititud.**

**La exactitud refleja qué tan cercano se encuentra un valor a comparación de su valor real.**

**Completitud se refiere a que los datos son completos, que no hay datos perdidos o valores atípicos**

**La consistencia se refiere a qué tan cerca en formato se encuentra los datos almacenados hablando en contenido, significado y estrucutra.**

**El contar con buenos datos es importante para el proceso de predicción, pues la falta de datos, una mal registro o la nula contingencia de los datos puede hacer imposible la predicción o reflejar resultados fuera de la realidad.**

**La limpieza de los datos es el proceso mediante el cual se detectan errores potenciales, datos perdidos, valores inusuales u otras inconsistencias y corregirlos. Datos limpios pueden mejorar en gran medida el proceso de pronóstico.**

**Cuando se tienen un conjunto de datos completos, la necesidad de obtener nuevos valores en función de los que ya se tienen, es lo que encamina a los diferentes profesionales del área a desarrollar métodos de predicción que nos adelante en el futuro. Como se ha venido diciendo, sin embargo, si se tiene en cuenta que la serie de tiempo puede tener partes incompletas, el modelo de predicción puede ser adaptado para ser un modelo de reconstrucción y con ello poder obtener una serie de datos completa que permita realizar nuevos análisis.**

**### Validación (terminar).**

**\*Porque es importante validar el modelo?\***

**\*Como los datos se organizan en diferentes conjuntos: entrenamiento validacion y prueba, y para que?\* (te pongo un ejemplo en esta sección)**

**[\*ejemplo: Los modelos se entrenan con un conjunto de entrenamiento, y se mide su desempeño o capacidad de generalizacion con el conjunto de validación.\***

**\*El conjunto de prueba se utiliza para que un tercero pruebe el modelo propuesto.\***

**\*Si la muestra de los datos son consitentes, entonces el error del modelo de prueba y validación deben ser parecidos.\*]**

**La validación del modelo consiste en una evaluación del modelo de pronóstico para determinar cómo es probable que funcione en la aplicación prevista.**

**Esto puede proporcionar guía útil sobre cómo funcionará el modelo de pronóstico cuando esté expuesto**

**a los nuevos datos y puede ser un enfoque valioso para discriminar entre modelos de pronóstico competitivos.**

**Para poder validar nuestro modelo es importante contar con datos de entrenamiento, sin embargo muchas veces no es posible obtener más datos de los que ya se tienen o es ncesario trabajar unicamente con los datos que se tienen en el momento, por ello es importante contar con un modelo que pueda generalizar a partir de los datos disponibles y para ello es necesario dividir los datos en diferentes conjuntos. Normalmente se dividen los datos en dos conjuntos, de entrenamiento y de prueba.**

**Usualmente se toma un setenta porciento de datos para entrenar al modelo y el treinta porciento restante es el conjunto de prueba.**

**#### Validación de modelos para datos independientes de contexto. (atemporales y no espaciales)(terminar)**

**\*\*Validación cruzada \\*$k$-fold\\*\*\***

**[\*(Modificar el ejemplo de validación de clasificación para convertirlo en un problema de predicción)\*, usar referencia de wikipedia [https://es.wikipedia.org/wiki/Validaci%C3%B3n\_cruzada](https://es.wikipedia.org/wiki/Validación\_cruzada)]**

**La validación cruzada es una herramienta usada para determinar la valides de un modelo.**

**es una [técnica](https://es.wikipedia.org/wiki/Técnica) utilizada para [evaluar](https://es.wikipedia.org/wiki/Evaluar) los resultados de un [análisis estadístico](https://es.wikipedia.org/wiki/Análisis\_estadístico) y garantizar que son [independientes](https://es.wikipedia.org/wiki/Análisis\_discriminante) de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la [media aritmética](https://es.wikipedia.org/wiki/Media\_aritmética) obtenida de las medidas de [evaluación](https://es.wikipedia.org/wiki/Evaluación) sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la [predicción](https://es.wikipedia.org/wiki/Predicción) y se quiere [estimar](https://es.wikipedia.org/wiki/Estimación\_estadística) la precisión de un modelo que se llevará a cabo a la práctica.[1](https://es.wikipedia.org/wiki/Validación\_cruzada#cite\_note-Devijver82-1) Es una técnica muy utilizada en proyectos de [inteligencia artificial](https://es.wikipedia.org/wiki/Inteligencia\_artificial) para validar modelos generados.**

**#### Validacion de modelos para predicción de datos temporales.(terminar)**

**\*Leer la referncia https://robjhyndman.com/hyndsight/tscv/ y basado en eso escribir esta sección\***

**\*Retomar la Propuesta de Hyndman. (Re-hacer la gráfica para no infringir derechos de autor, puede ser este ejemplo, y ademas simplificado, e.g., menos datos y unos cuantos renglones )\***

**![cv1-1](https://bytebucket.org/arielcg/centromet/raw/d6f32624c75e8cf5c036421cecbfe6845932867d/Reportes/imagenes/cv1-1.png)**

**#### Validación de modelos para reconstrucción de datos temporales (terminar).**

**\*Descripción de la propuesta con palabras\***

**\*Esta figura no existe, hay que elaborarla y es importante porque es lo que proponemos\***

**![propuesta\_validacion\_cruzada\_reconstr](https://bytebucket.org/arielcg/centromet/raw/d6f32624c75e8cf5c036421cecbfe6845932867d/Reportes/imagenes/propuesta\_validacion\_cruzada\_reconstr.png)**

**### Medidas del error (yo lo completo)**

**\*Mean Squared Error: MSE\***

**\*Mean Absolute Errors: MAE\***

**\*Otras medidas del error:\***

**### Propuesta del algoritmo general de reconstrucción (terminar).**

**\*por ejemplo:\***

**1. \*Organizar serie de tiempo en conjuntos de entrenamiento, validación y prueba.\***

**2. \*generar fallas aleatorias a la serie de tiempo\***

**3. \*Reconstruir datos faltantes con un modelo de reconstrucción.\***

**4. \*Medir el error de reconstrucción\***

**\*Describir de manera general como queremos diseñar el software, puede ser textual, diagrama de bloques, UML, las funciones o clases\***

**### Conclusiones**

**### Referencias**

**1. TISSBERT: Reconstruccion de series de tiempo https://polipapers.upv.es/index.php/raet/article/view/9749/10323**

**2. What to Do about Missing Values in Time-Series Cross-Section Data. (James Honaker The Pennsylvania State University Gary King Harvard University) 2010**

**3. E. Castaño: Time Series Data Reconstruction: An Application to the Hourly Demand of Electricity. https://www.researchgate.net/publication/262722662\_Time\_Series\_Data\_Reconstruction\_An\_Application\_to\_the\_Hourly\_Demand\_of\_Electricity**

**4. R. Hyndman. Cross-validation for time series https://robjhyndman.com/hyndsight/tscv/**

**5. A. M. de Livera R. J. Hyndman and R. D. Snyder. BATS**

**6. D. Montomery. et al. Time Series Analysis and Forecasting. Second Edition.**

**7. [https://es.wikipedia.org/wiki/Validaci%C3%B3n\_cruzada](https://es.wikipedia.org/wiki/Validación\_cruzada)**

**8. Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). \*Introduction to time series analysis and forecasting\*. John Wiley & Sons.**

**9. Ostertagová, E., & Ostertag, O. (2011, September). The simple exponential smoothing model. In \*The 4th International Conference on Modelling of Mechanical and Mechatronic Systems, Technical University of Košice, Slovak Republic, Proceedings of conference\* (pp. 380-384).**

**10. Rosano, F. L. (1996). Fundamentos de redes neuronales artificiales.**

**11. Berástegui Arbeloa, G. (2018). Implementación y comparación del algoritmo de los k vecinos más cercanos (k-NN) con valores locales en k.**

**12. Sucar, L. E. (2006). Redes bayesianas. \*BS Araujo, Aprendizaje Automático: conceptos básicos y avanzados\*, 77-100.**