UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS Y COMPUTACIÓN

Proyecto de Investigación previo a la obtención del título de Ingeniero en Sistemas y Computación

TRABAJO DE TITULACIÓN

ANÁLISIS DE LAS TÉCNICAS DE SUAVIZADO PARA SERIES TEMPORALES APLICADAS A LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH

AUTOR:
Alexis Fabricio Mata Hidalgo

TUTOR:
Ing. Lida Barba, Ph.D.

Riobamba - Ecuador:

Año 2019

Los miembros del Tribunal de Graduación del proyecto de investigación de título:

"Análisis de las técnicas de suavizado para series temporales aplicadas a la base de datos del sistema académico de la UNACH", presentado por el Sr. Alexis Fabricio Mata Hidalgo y dirigida por la: Ing. Lida Barba, Ph.D.

Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación con fines de graduación escrito en el cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la Facultad de Ingeniera de la UNACH.

Para constancia de lo expuesto firman:

Ing. Lida Barba, PhD
Directora del Proyecto

Mgs. Ana Congacha
Miembro del Tribunal

Mgs. Lady Espinoza
Miembro del Tribunal

AUTORÍA DE LA INVESTIGACIÓN

"La responsabilidad del contenido de este proyecto de graduación, corresponde exclusivamente a Alexis Fabricio Mata Hidalgo, autor del proyecto de investigación, bajo la dirección de la Ing. Lida Mercedes Barba Maggi y el patrimonio intelectual de la misma Universidad Nacional de Chimborazo"

Alexis Fabricio Mata Hidalgo

060377923-2

DEDICATORIA

Dedico esta investigación a toda mi familia quienes siempre estuvieron a mi lado apoyándome en los buenos y malos momentos de mi vida, a mis padres Rodrigo y Geovanna que han sido mi apoyo incondicional para poder culminar mis estudios, a mis hermanos Tatiana y Frederick, abuelitos, tíos, primos que siempre me brindaron su apoyo, consejos para poder culminar esta etapa importante de mi vida.

AGRADECIMIENTO

Agradezco infinitamente a Dios por protegerme y guiar mi camino en todos estos años de vida, por brindarme salud y darme las fuerzas necesarias para lograr culminar mi carrera universitaria.

Agradezco a mis padres, hermanos, abuelitos, tíos, primos por el apoyo incondicional, por sus consejos, por siempre estar presentes en cada paso de mi vida.

Mi gratitud entera para la Universidad Nacional de Chimborazo por abrirme las puertas y darme la oportunidad de ser un profesional, a la carrera de Ingeniería en Sistemas y Computación, docentes, compañeros y amigos de clases Jackeline, Mónica, Erika, Alex, Juan, Julio, Kleber, Andrés por compartir sus conocimientos y experiencias. En especial a la Ing. Lida Barba, Ph.D tutora de tesis quien me brindó su apoyo incondicional, de igual forma a mis tutores colaboradores, MsC. Lady Espinoza y MsC. Ana Congacha.

ÍNDICE GENERAL

Contenido **Pág.**

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El pronóstico basado en series de tiempo genera conocimiento útil en la toma de decisiones, sin embargo, lograr la exactitud en los modelos es el mayor reto para el investigador. En diferentes trabajos se han aplicado técnicas y métodos buscando crear modelos más competitivos. En la presente investigación se aplican tres técnicas de suavizado de series temporales, Media Móvil, Suavizado Exponencial y Descomposición de valores singulares de Hankel en un modelo Autoregresivo Lineal. Los resultados son comparados para identificar la técnica que contribuye de mejor manera en la exactitud del modelo. Los datos utilizados corresponden al número de publicaciones científicas de los docentes de la Universidad Nacional de Chimborazo entre los años 2014 al 2018 por medio del sistema de base de datos Sicoa, además se implementa un portal interactivo por medio del cual se realiza el proceso de suavizado para cualquier serie de tiempo que ingrese el usuario.

Palabras claves: Media Móvil, Suavizado Exponencial y Descomposición de valores singulares de Hankel, Universidad Nacional de Chimborazo, sistema de base de datos SICOA, proceso de suavizado, registros de usuarios digitales.

Abstract

The forecast based on time series generates useful knowledge in making decision; however, achieving accuracy in models is a great challenge for a researcher. In previous investigations, some techniques and methods have been applied to create more competitive models. On this research, three time series have been employed about smoothing techniques such as Moving Average, Exponential Smoothing and Decomposition of Hankel singular values based on a Linear Autoregressive model. The results allowed identifying the best technique that contributes to the accuracy of the model. The data used correspond to the number of scientific publications by professors from the National University of Chimborazo from 2014 to 2018. They employed a database system called SICOA. It is an interactive portal which employs the smoothing process while the digital user logs in at any time series.

Key words: Moving Average, Exponential Smoothing and Decomposition of Hankel singular values, the National University of Chimborazo, database system called SICOA, smoothing process, digital user logs.

Reviewed and corrected by: Lic. Armijos Jacqueline, MsC.

INTRODUCCIÓN

En la actualidad las instituciones educativas buscan mejorar sus procesos para lograr una competitividad con el resto de instituciones, el constante cambio de tecnologías ha exigido a las instituciones realizar estudios en el ámbito educativo y administrativo. Es por eso que últimamente existe un gran interés por aplicar técnicas de minería de datos en ambientes de educación superior. La minería de datos conocida también como Descubrimiento de Conocimiento en Base de datos, es el campo que permite descubrir información nueva y potencial de grandes cantidades de datos (Galindo & Garcia, 2010). También se la puede describir como un área multidisciplinaria en la cual influyen varios paradigmas computacionales, además trabaja con varios métodos, los cuales analizan, exploran y visualizan la información de sistemas computacionales de aprendizaje (Huapaya, Lizarralde, Arona, & Massa, 2012).

Es de suma importancia el pronóstico y estimación para apoyar a la planificación y a la toma de decisiones de la Universidad Nacional de Chimborazo (UNACH). El pronóstico basado en series temporales permite extender valores del pasado para poder predecir valores futuros, para realizar dichas predicciones primero se tiene que encontrar reglas o modelos (Olmedo, Valderas, Mateos, & Gimeno, 2004). Por ello el pronóstico se torna en un proceso complejo debido a que los fenómenos que se estudian tienen características no estacionarias, es decir son de alta variabilidad (Arenas, 2009).

Con la finalidad de mejorar la exactitud de pronóstico, han sido analizadas varias técnicas de pre procesamiento de los datos, por ejemplo: Barba *et al.* (2014), aplicaron técnicas de suavizado con series temporales por medio de Media Móvil(MA) utilizada para extraer la tendencia y por medio de la técnica de descomposición de valores singulares de Hankel(HSVD), con la finalidad de

extraer componente de alta y baja frecuencia como proceso previo al pronóstico (Barba, Rodriguez, & Montt, 2014), con tal implementación lograron mejorar la exactitud del pronóstico de accidentes de tránsito en Chile. Otra técnica de suavizado es el Suavizado Exponencial(SE) con dicha técnica Ferrero *et al.* (2017), lograron mayor exactitud de pronóstico del impacto de mortalidad de desastres en España entre los años 1950 -2012.

La Universidad Nacional de Chimborazo cuenta con un sistema de base de datos que registra la información de las actividades académicas e investigativas, constan datos correspondientes a las publicaciones científicas que han sido registradas en el observatorio de investigación, siendo esta una información de vital importancia en los procesos de evaluación con fines de acreditación, por ser indicadores de la calidad académica de las instituciones de educación superior. Una vez revisada la información se identifica que en esta base de datos se han registrado 12054 publicaciones en sus diferentes estados como son: publicaciones aceptadas, publicadas, en impresión, con evidencias incompletas, inhabilitadas, promoción docente y procesos electorales y patente en etapa de publicación y oposiciones; de estas publicaciones se realizó un proceso de depuración para seleccionar los datos para la investigación.

La metodología CRISP-DM que es convencionalmente utilizada para procesos de minería de datos, fue aplicada para guiar la investigación, esta metodología está dividida en seis fases: compresión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

Un modelo de pronóstico para la información concerniente a las publicaciones científicas de la UNACH contribuirá de manera sustantiva a la toma de decisiones hacia la mejora continua, sin

embargo, los modelos de pronóstico presentan dificultad al momento de alcanzar niveles óptimos de exactitud, por tal motivo en esta investigación se analizarán y probarán tres técnicas de suavizado de series temporales para mejorar la exactitud del modelo de pronóstico. Se aplicarán las técnicas de Media Móvil, Suavizado Exponencial y Descomposición de valores singulares de Hankel(HSVD) por medio de la aplicación Matlab. Como valor agregado a esta investigación se implementará un portal interactivo que realiza el suavizado de cualquier serie de tiempo ingresada por el usuario.

1. CAPÍTULO I

1.1 PLANTEAMIENTO DEL PROBLEMA

Según López (2007), en la última década el crecimiento de grandes volúmenes de datos y el uso de herramientas informáticas ha hecho que los análisis hacia determinadas técnicas ahora se engloben con el nombre de minería de datos.

La Universidad Nacional de Chimborazo cuenta con un sistema informático en el cual consta información académica de estudiantes y docentes, sin embargo, no se ha encontrado registro alguno de que esta información haya sido analizada con técnicas de minería de datos que apoye a la toma de decisiones, desaprovechándose así las oportunidades actuales que existen de obtener conocimiento a partir de los datos. Estas técnicas convierte a los datos adquiridos desde ambientes virtuales en información que podría representar gran impacto para la creación de un modelo de análisis, estos datos primero deben ser transformados para que sean compatibles con los distintos métodos o técnicas de suavizado (Huapaya, Lizarralde, Arona, & Massa, 2012).

La base de datos de publicaciones científicas de docentes constituye a una información clave debido a que estas publicaciones son un indicador para el proceso de acreditación de las universidades. Según el Art. 350 de la Constitución de la República del Ecuador, se establece que el sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica, estos indicadores de evaluación tienen como fin proporcionar a las autoridades resultados cuantitativos y cualitativos que constituyan un sustento válido para la toma de decisiones.

1.2 JUSTIFICACIÓN

En la en la actualidad las empresas, industrias, organizaciones e instituciones educativas tanto públicas como privadas generan grandes cantidades de información, que en muchos casos pueden llegar a cientos o miles de datos diarios, en la búsqueda de lograr mayor productividad y competitividad, ha puesto su mirada en los sistemas de extracción de conocimiento.

La exactitud del pronóstico es un desafío para los investigadores, existen varios modelos que sirven para el análisis de datos, estos modelos combinados con las técnicas de suavizado potencian la obtención conocimiento. Es por ello que se han elegido las herramientas adecuadas donde se analizaran los datos, para generar resultados con claridad con un lenguaje de comunicación ideal para no correr el riesgo de desvalorar esta investigación, esto porque el lector será quien juzgue la calidad de este contenido (Griffies, Perrie, & Hull, 2013).

La investigación con los datos mencionados anteriormente es factible debido a que existe el interés institucional de entregar la información requerida para los procesos investigativos planificados y que aportan al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica", en marcha en la UNACH.

1.3 OBJETIVOS

1.3.1 Objetivo General

- ✓ Analizar las técnicas de suavizado para series temporales para mejorar la exactitud del pronóstico de la información seleccionada de la base de datos del sistema académico de la UNACH.

1.3.2 Objetivos Específicos

- ✓ Aplicar las técnicas de suavizado Media Móvil, Suavizada Exponencial y HSVD a las series de tiempo correspondientes a producción científica, extraídas de la base de datos del sistema académico de la UNACH.
- ✓ Implementar un modelo de pronóstico a partir de los datos suavizados.
- ✓ Comparar la exactitud del pronóstico y determinar el modelo más eficiente para su uso en la UNACH con los datos seleccionados.
- ✓ Implementar un portal interactivo para suavizar cualquier serie de tiempo ingresado por el usuario.

2 CAPÍTULO II

2.1 Marco Teórico

2.1.1 Learning Analytics

Es un campo en donde se utiliza varias herramientas sofisticadas para el análisis con el fin de mejorar el aprendizaje y la educación, está estrechamente vinculado con otros campos de estudio como la inteligencia de negocios empresarial, analítica web, analítica académica, minería de datos y análisis de acción (Elias, 2011).

Learning Analytics es una herramienta novedosa relacionada directamente con la minería de datos que mediante el registro y estudio crítico de determinados indicadores, contribuye a la personalización y adaptación del aprendizaje así como también coopera en la planificación educativa con el objetivo de mejorar el desarrollo competencial y la significatividad de lo aprendido (Gutierrez, 2015).

A nivel de educación superior el análisis de los datos es imperativo, debido a que los procesos de gestión de la calidad están en constante evolución, Rodríguez (2018), manifiesta que el Learning Analytics no solo analiza datos sino también que es una disciplina emergente relacionada con el desarrollo de métodos para explorar series de datos procedentes de ecosistemas educativos. Los resultados del análisis de los datos de las instituciones de educación superior, permitirán entender de mejor manera la dinámica del sistema.

La ley Orgánica de Educación Superior (LOES), determina que, la calidad es un proceso para determinar las condiciones de las instituciones, mediante la recopilación sistemática de los datos cuantitativos y cualitativos porque va a permitir emitir un juicio o diagnóstico, analizando sus

componentes, funciones o procesos, esto con el fin que sus resultados sirvan para el desarrollo de las instituciones.

2.1.2 Minería de datos

Es un área multidisciplinaria en la cual se abarcan varios paradigmas computacionales como la programación lógica, redes neuronales artificiales, lógica difusa, inducción por reglas, además de ello trabaja con métodos como el clustering, estadística, clasificación, minera de textos, estos métodos permitirán descubrir información nueva y muy útil de grandes cantidades de datos (Huapaya, Lizarralde, Arona, & Massa, 2012).

La minería de datos pretende un descubrimiento automático del conocimiento con la información almacenada y ordenada en grandes bases de datos. Como se lo realiza en esta investigación las técnicas tienen un objetivo en específico que es la de descubrir patrones, perfiles y tendencias a través del análisis de los datos, utilizando tecnologías de inteligencia artificial, redes neuronales lógica difusa, algoritmos genéticos, series temporales y varias técnicas más que analizan los datos (Montero, 2007).

López (2007), determina que la minería de datos es solo una etapa para el proceso de extracción del conocimiento, la inteligencia artificial o el aprendizaje automático son de mucha importancia para generar un debido pronóstico, cabe recalcar que el proceso de extracción del conocimiento incorpora varias técnicas como los arboles de decisión, redes neuronales artificiales, técnicas bayesianas, máquinas de soporte vectorial y las técnicas de la presente investigación como son las series temporales combinadas con técnicas de suavizado que ayudan a generar una mayor exactitud al momento aplicarlas a cualquier modelo.

2.1.3 Minería de datos educativos

Es una disciplina emergente, que se centra en el desarrollo de métodos para explorar los datos procedentes del contexto educativo. En los últimos años desde distintos ámbitos que incluyen informática, estadística y educación se ha estado investigando sobre cómo la minería de datos puede mejorar la educación**.** Los datos que se utilizan proceden de diversas fuentes: clases en entornos tradicionales presenciales, software educativo, cursos online o pruebas acumulativas. Todas ellas proveen de una cantidad de datos en aumento, que puede ser analizada para dirigir preguntas que antes no era posible realizar y contemplan diferencias entre poblaciones de estudiantes o comportamientos específicos (Jiménez & Álvarez, 2010).

La minería de datos educativa ofrece numerosas ventajas comparándola con los paradigmas más tradicionales de investigación relativa a la educación. En particular la creación de repositorios de datos educacionales ha permitido generar bases de datos que hace posible la minería de datos en la educación, los métodos empleados en la minería de datos educativa suelen diferir de los métodos más generalistas, explotando explícitamente los niveles de jerarquía presentes en los datos (Jiménez & Álvarez, 2010).

2.1.4 Series temporales

Se las puede describir como procesos estocásticos o simplemente una sucesión ordenada a lo largo de un determinado tiempo de un conjunto de variables aleatorias, con una determinada realización de un proceso de series temporales se va a obtener un valor u observación de las variables que integran el sistema y estos valores a su vez evolucionarán a lo largo del tiempo de acuerdo con las leyes probabilísticas (Gras, 2001).

Pueden clasificarse según la forma en que se ofrecen los valores de las mediciones en dos tipos:

- **Continuas:** cuando los valores se ofrecen de forma permanente, de manera tal que cada uno de ellos representa el estado de la variable en un instante, el cual puede ser tan pequeño como teóricamente se quiera suponer (Gras, 2001).
- **Discretas:** cuando los valores se ofrecen para intervalos de tiempo, generalmente homogéneos y donde representan la magnitud acumulada del estado de la variable durante ese intervalo (Gras, 2001).

Otra característica se las series temporales, es que hace muy difícil su tratamiento mediante los métodos estadísticos habituales, pues en la mayoría de éstos se exige el cumplimiento del supuesto de independencia de las observaciones, mientras que las series generalmente se caracterizan por la dependencia existente entre observaciones sucesivas (Coutin, 2001). La figura 1 muestra un ejemplo de serie de tiempo, en ella se puede observar la evolución del PIB anual en Ecuador entre los años 2000 al 2018.

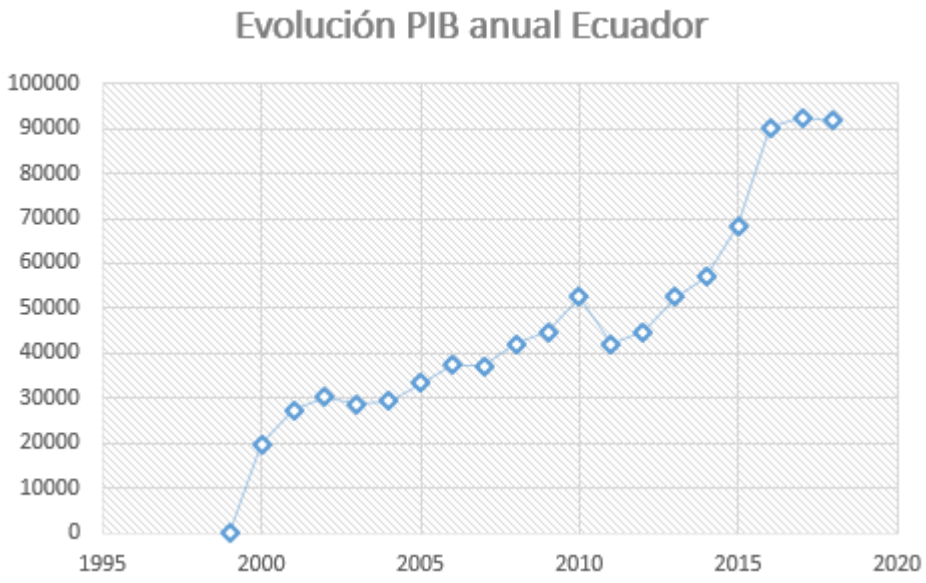

Figura 1: Evolución del PIB anual en Ecuador.

En esta serie se observa cómo ha ido evolucionando el PIB anual en el Ecuador por cada año por ejemplo en 2018 ha crecido un 1,4% respecto a 2017 y entre los años 2000 a 2018 ha tenido un crecimiento del 3.54%.

2.1.5 Técnicas de Suavizado

2.1.5.1 Media Móvil

La media móvil es un cálculo utilizado para analizar un conjunto de datos en modo de puntos para crear series de promedios. La media móvil contiene una secuencia de valores, cada valor es el promedio de un subconjunto de los datos originales, si el subconjunto de datos es de cinco, mostrará el promedio de los datos de cada cinco días dependiendo de dos factores, los valores que se están promediando y el horizonte temporal (Galán V. , 2015).

Esta técnica se calcula con la siguiente formula.

$$\bar{s}_k = \sum_{i=k-1}^{k+1} \frac{x_i}{3} \qquad (1)$$

Donde $\bar{s}_k$ es el enésimo elemento de señal suavizado para $k = 2, \ldots, n-1$, x_i es el elemento observado de la serie temporal original y los términos $\bar{s}_1$ y $\bar{s}_n$ tienen los mismos valores de x_1 y x_n respectivamente, para conformar la serie temporal $[\bar{s}_1, \bar{s}_2, \bar{s}_3 \ldots \bar{s}_n]$.

2.1.5.2 Suavizado Exponencial

Esta técnica se emplea tanto para suavizar como para realizar pronósticos, puede considerarse como la evolución de la técnica media móvil, tiene un mecanismo que ajusta los pronósticos en dirección opuesta a los errores pasados (Pérez, 2005).

Utiliza una constante llamada contante de suavizado que en este caso es el alfa (α), que tiene que ser mayor a 0 y menor a 1, para ello se utiliza la ecuación 2, donde se aprecia que la observación más próxima recibe el peso de α, y la siguiente observación más cercana recibe el peso de 1 – α,

además que el resultado tiende a ser valor absoluto, por eso el orden de estos dos componentes no influye. El suavizado exponencial se calcula con la siguiente ecuación.

$$\boldsymbol{Y' = \alpha x_{\square} + (1 - \alpha) x_{\square}'} \quad (2)$$

2.1.5.3 Descomposición de valores singulares

La técnica Descomposición de valores singulares ha sido apreciada desde mucho tiempo atrás en todo el mundo, aplicada en la teoría de matrices según Stewart en 1993, la SVD está estrechamente relacionada con la descomposición espectral, en 1930 Eckart y Young descubrieron que esta técnica se la puede utilizar para derivar la descomposición polar autónoma en la que una matriz se factoriza en el producto de una matriz ermitaña y una matriz unitaria, la SVD en un principio se aplicó para matrices cuadradas pero luego se extendió a matrices rectangulares.

SVD se ocupa del análisis de componentes para la reducción de la dimensionalidad cuyo cálculo se basa en una matriz simétrica positiva semidefinida. Golub (1965), publicó el primer algoritmo que proporciona información esencial sobre varios antecedentes matemáticos necesarios para la producción de software numérico.

La SVD en los últimos tiempos se ha utilizado en diferentes campos, desde la revisión de la literatura se encontraron aplicaciónes la SVD para la eliminación de ruido, la reducción de rasgos y la compresión de imágenes. Por ejemplo; Zhao (2009), demostraron como una señal se puede descomponer en la suma lineal de una serie de señales de componentes por SVD basada en el matiz de Hankel.

2.1.5.4 Descomposición de valores singulares de Hankel

La técnica HSVD fue utilizada por primera vez por Barba *et al.* (2014) para extraer los componentes intrínsecos de baja y alta frecuencia de una serie temporal. Este proceso es implementado en los tres pasos siguientes: embebido, descomposición y desembebido.

Embebido

Para el embebido se utiliza una matriz de Hankel en el primer paso del método HSVD. La serie temporal univariante observada x de los valores $[x_1 \dots x_N]$, esta embebida en una matriz $\mathrm{H_L} * k$ de forma de matriz de Hankel, esto significa que todas sus diagonales oblicuas con constantes.

$$\mathrm{H} = \begin{vmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1k} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2k+1} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ X_{L1} & X_{L2} & X_{L3} & & X_{Lk} \end{vmatrix} \tag{3}$$

L es la longitud de ventana y *k* se calcula con la siguiente ecuación:

$$k = N - L + 1 \tag{4}$$

La longitud de la ventana *L* es un numero entero, $2 \leq L \leq N$. la selección de *L* depende de las características de la serie temporal y del proceso de pruebas que se ejecutan en esta investigación.

Descomposición

Sea H una matriz $L * K$, entonces existe una matriz ortogonal U $L * L$, una matriz ortogonal V $K * K$, y una matriz diagonal $\sum L * K$, con entradas de diagonales $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L$ para $L < K$, tal que U^T HV = S y S = HH^T además los numero $\lambda_1, \lambda_2 \ldots \lambda_L$, están exclusivamente determinados por H.

$$H = U * \Sigma * V^T \tag{5}$$

U es la matriz de los vectores singulares izquierdos de H y V es la matriz de los vectores singulares derechos de H. además, la colección (λ_i, U_i, V_i) es el primer eigentriple del HSVD. Las matrices elementales $H_i, \ldots, H_L$ de igual dimensión (L x K) se obtiene de cada eigentriple (λ_i, U_i, V_i).

$$H_i = \lambda_i * U_i * V_i^T \tag{6}$$

Desembebido

Este proceso se lo desarrolla para extraer los componentes intrínsecos. Cada matriz elemental H_i contiene cada componente en su primera y última columna por lo tanto los elementos C_i.

$$C_i = [H_i(1,1), H_i(2,2), \ldots, H_i(1,K), \ldots, H_i(L,K)] \tag{7}$$

2.1.5.5 Modelo Autoregresivo

La finalidad del modelo autoregresivo es estimar los valores de una variable con base a valores conocidos de otra, en otras palabras, la forma de emplear una ecuación de regresión o en este caso,

una ecuación autoregresiva es para explicar los valores de una variable en términos de otra. El análisis de la regresión lineal únicamente indica que relación matemática podría haber (Walpole, 2012).

El pronóstico con el modelo autoregresivo tiene un objetivo en general, que es tratar de explicar la relación que existe entre una variable dependiente y un conjunto de variables independientes o variables explicativas, $x_1 \dots x_n$ (Carollo, 2012).

Las variables dependientes generan la respuesta que se observa en el estudio y que podrían estar influenciadas por los valores de las variables dependientes, esta variable es una característica que se trata de cambiar manipulando la variable independiente, en cambio, la variable independiente es aquella manipulada por el analista con el objetivo de estudiar cómo actúa sobre la variable dependiente (Walpole, 2012).

2.2.6 Metodología CRISP – DM

Esta metodología fue creada por el grupo de empresas SPSS, NCR y Darimer Chrysler en el año 2000 y es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. La metodología está estructurada en seis fases: comprensión de negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación, dichas fases cuentan con varias tareas generales de segundo nivel, es decir, CRISP – DM establece un conjunto de taras y actividades para cada fase del proyecto, pero no especifica cómo llevarlas a cabo (Moine, Haedo, & Gordillo, 2011).

La Metodología CRISP-DM, también describe una manera en la que los expertos en esta materia abordan el problema. Para implementar una tecnología en un negocio es necesaria una metodología. Estos métodos suelen venir de las experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos de implementación de minería

de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM y que representa el grado de utilización de las principales guías de desarrollo de proyectos de minería de datos según encuestas realizadas. CRISP-DM ha experimentado un ligero descenso en los últimos años, pero sigue siendo la más empleada de las distintas metodologías para la realización de proyectos con técnicas de minería de datos (Moine, Haedo, & Gordillo, 2011).

CRISP-DM incluye un modelo y una guía, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última (Galán V. , 2015).

2.2.7 Lenguaje de programación PHP

Este es un lenguaje de pre procesamiento de texto libre, se lo usa en la actualidad solamente para el desarrollo de aplicaciones que actúan al lado del servidor, que el más utilizado para crear contenido para la World Wide Web. PHP es popular entre los lenguajes para generar documentos HTML, el código es interpretado en el lado del servidor, pero también genera una página web para que los clientes puedan observarla. Es posible instalar PHP en la mayoría de sistemas operativos actuales, su tecnología ha evolucionado para ofrecer características adicionales a la de línea de comandos, por ello es competencia directa de la tecnología ASP que le pertenece a Microsoft. La licencia que maneja PHP es PHP License, pero aquí existe un pequeño inconveniente ya que no es compatible con la GNU debido a las propias restricciones de uso de PHP (Arias, 2017).

Su instalación es súper sencilla y permite el uso de servidores web. Al momento de la instalación varios módulos son creados conjuntamente con el repositorio de extensiones, varios de estos módulos son introducidos como patrones para nuevas versiones del mismo lenguaje, además su

sintaxis es muy parecida a otros lenguajes como C y C++. En la actualidad un lenguaje de programación debe tener conexión a una base de datos con PHP esto lo hace simple, da soporte a un gran número de bases de datos como son: ORACLE, POSTGRESQL, INTERBASE, MYSQL, SQLITE, ADOBD y varias más que existentes (Arias, 2017).

En la presente investigación se utiliza este lenguaje combinada con el gestor de base de datos MySQL para la creación del portal web interactivo.

2.2.8 Gestor de base de datos MySQL

Uno de los gestores de base de datos más populares en la actualidad a nivel mundial es MySQL, es muy conocido por su rendimiento y puede ser utilizado dentro de la arquitectura cliente servidor, el servidor mysqld es el encargado de captar las peticiones generadas por los clientes para luego transformar en un plan de educación, luego de ello recupera los datos según el plan de ejecución y al final devuelve los resultados al cliente (Combaudon, 2018).

Varios módulos de gestión componen esta arquitectura asi:

- Protocolos de comunicación con los clientes.
- Permisos de accesos o roles a administradores y clientes.
- Cachés para minimizar el acceso al disco.
- Varios tipos de registros para el servidor.
- El debido análisis a la optimización y la ejecución de las peticiones y por último el almacenamiento de los datos.

3. CAPÍTULO III

3.1.METODOLOGÍA

La metodología utilizada para el desarrollo del presente trabajo fue de tipo cuantitativa, porque en la investigación se utilizaron herramientas informáticas, cálculos matemáticos y cálculos estadísticos para poder analizar los datos y obtener resultados.

3.1.1 Tipo de investigación

Con la revisión de la literatura se realizó una investigación de tipo bibliográfica, que está basada en técnicas y estrategias que fueron empleadas para identificar, acceder y verificar aquellos documentos como artículos científicos, libros, tesis, entre otros con respecto al tema de estudio, como material de apoyo para el respaldo del trabajo de investigación y que se encuentran con las debidas citas según las normas APA.

3.1.2 Método de investigación

Para esta investigación se utilizó el método analítico, que contribuye como una manera de proceder para llegar a un resultado mediante la descomposición de un fenómeno en sus elementos constitutivos, además ayudo a conocer el objeto de estudio, sus características con las cuales se pudieron explicar, hacer analogías, comprender mejor su comportamiento y establecer nuevas teorías (Echaverría, Gómez, Aristizábal, & Vanegas, 2010).

3.1.3 Procesamiento y análisis

Para el análisis de los datos se utilizó la metodología CRISP-DM, esta metodología ha sido desarrollada exclusivamente para la elaboración de proyectos de minería de datos, la cual funciona como un ciclo de vida y se divide en seis fases, cabe recalcar que la sucesión de estas fases no

necesariamente tienen un orden sino que puede varias según las tareas o niveles de su uso (Ochoa, Britos, & Martinez, 2006).

Fase de comprensión del negocio

En esta fase la finalidad es determinar los objetivos y requisitos de la investigación desde una perspectiva de negocio.

En referencia a la situación actual de la universidad al principio de esta investigación se puede decir que se cuenta con la base de datos del sistema académico, que consta de dos grupos de datos, información académica de estudiantes y publicaciones científicas de docentes, sin embargo, no existe ningún informe sobre estudios anteriores de la aplicación de técnicas de minería de datos sobre esta información. Por ello el objetivo principal de esta investigación es generar un conocimiento por medio del pronóstico aplicando técnicas de minería de datos con la información del sistema.

En cuanto a los requisitos de software de la investigación se dispone del programa Matlab que proporciona herramientas eficientes para la aplicación de las técnicas de minería de datos.

Fase de comprensión de datos

En la segunda fase se realiza la exploración inicial de los datos para así poder establecer un primer contacto con el problema, familiarizarse con los datos y verificar su calidad para descubrir cuál de los dos grupos de datos aporta más en nuestra investigación, los datos académicos de estudiantes o los de investigación científica de docentes.

En el sistema académico se registran datos académicos de los estudiantes, información personal, rendimiento, periodo, nivel, situación actual entre otras variables, por otro lado, en la investigación de docentes se registran datos como el estado de publicación, tipo de publicación, titulo, área de investigación, la fecha de aceptación, fecha de registro y la fecha de publicación.

Por lo tanto, una vez analizadas estos dos grupos de datos mediante una ETL encargada de extraer, transformar y cargar la información. Se ha determinado que la base de datos de estudiantes no aporta suficiente información para nuestra investigación debido a que esta base registra información con una frecuencia semestral y como el objetivo de la investigación es realizar un pronóstico y estudios lo más reales posibles, se necesita de un mayor número de registros o con frecuencias más cortas para que el trabajo sea exitoso, por eso se ha decidido trabajar con la base de datos de publicaciones de docentes que es la única variable que registra información de las publicaciones diarias en los últimos cinco años, desde el 2014 hasta el 2018.

Fase de preparación de datos

En la tercera fase de la metodología se preparar la información para adecuarla a las técnicas de suavizado y de pronóstico que se van a emplear en la investigación, para ello se ha decidido agrupar los datos en tres frecuencias, diaria, semanal y mensual, con la finalidad de realizar comparaciones e identificar la mejor aproximación, obteniendo:

- De la información diaria 1825 valores de los cinco años como elementos de la serie temporal con frecuencia diaria.
- De la información semanal 260 valores de los cinco años como elementos de la serie temporal con frecuencia semanal.
- De la información mensual 60 valores de los cinco años como elementos de la serie temporal mensual.

Estas series temporales presentan comportamientos no lineales, esto hace posible que se puedan formular los modelos matemáticos basados en leyes de estadística. La tendencia es un componente de estos modelos que representa el comportamiento de la serie. La media o promedio es una medida con la que se puede presentar dicha tendencia, la varianza también es una medida muy importante,

porque representa la variabilidad de la serie de tiempo con respecto a la media y la desviación estándar mide cuanto se separan los datos de la serie, en las ecuaciones ocho , nueve y diez se describen como se calculan estas medidas estadísticas y medidas de tendencia central (Rodríguez, 2016).

Media $$x = \frac{\sum x_i}{\mathrm{n}} \quad (8)$$

Varianza $$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (9)$$

Desviación estándar $$\sigma = \sqrt{s^2} \quad (10)$$

Tabla 1. Medidas estadísticas y de tendencia central (frecuencia diaria).

	2014	**2015**	**2016**	**2017**	**2018**
Media	*8,29*	*14,77*	*17,20*	*30,17*	*16,50*
Varianza	*2427,91*	*394,65*	*703,41*	*2875,79*	*219,73*
Desviación estándar	*49,27*	*19,86*	*26,52*	*53,62*	*14,82*

Tabla 2. Medidas estadísticas y de tendencia central (frecuencia semanal).

	2014	**2015**	**2016**	**2017**	**2018**
Media	3,63	*22,73*	*41,03*	*91,44*	*65,71*
Varianza	*360,94*	*1349,80*	*2404,03*	*44670,68*	*2681,89*
Desviación estándar	*18,99*	*36,73*	*49,03*	*211,35*	*51,78*

Tabla 3. Medidas estadísticas y de tendencia central (frecuencia mensual).

	2014	**2015**	**2016**	**2017**	**2018**
Media	*45,16*	*98,50*	*177,83*	*389,75*	*210,41*
Varianza	*14343,60*	*3995,18*	*8290,69*	*99072,93*	*7886,03*
Desviación estándar	*119,76*	*63,20*	*91,05*	*314,75*	*88,80*

A partir de las tablas 1,2,3 se determina que la media no permanece constante a lo largo del tiempo, en cada tabla la media cambia de acuerdo al número de publicaciones que tiene cada año, también se observa que la desviación estándar tampoco es constante porque depende de la media.

Fase de modelización

Para esta fase de la metodología se seleccionan las técnicas más apropiadas, para luego aplicar dichas técnicas sobre los valores de la serie temporal generada en la fase de preparación de los datos y así generar el modelo, y por ultimo tendremos que evaluar si el modelo ha cumplido con los objetivos con éxito o no.

Las técnicas de suavizado se utilizan para mejorar las características de variabilidad subyacentes en los datos contribuyendo a revelar características importantes.

Se utilizaron tres técnicas de suavizado como son media móvil, suavizado exponencial y descomposición por valores singulares de Hankel, cada una de estas técnicas de suavizado tiene una funcionalidad diferente, por ejemplo:

- La media móvil generalmente se utiliza con valores de serie de tiempo para suavizar a corto plazo y para resaltar tendencias o ciclos a largo plazo.
- El Suavizado Exponencial asume que tiene una media estable por lo que este suavizado no funciona para la predicción de valores que tienen estacionalidad o una tendencia (Chen, 2016).
- La Descomposición por Valores Singulares de Hankel es diferente a las dos técnicas anteriores, porque descompone a la serie temporal original en componentes de baja y alta frecuencia. El componente de baja frecuencia $C_L$se extrae de la primera matriz elemental H□ que se calcula con.

$$H_i = \lambda_i * U_i * V_i^T \tag{11}$$

Mientras que la componente de alta frecuencia se calcula por sustracción simple con la siguiente ecuación.

$$C_H = X(n) - C_L(n) \tag{12}$$

Estos dos componentes $C_L(n)$ y $C_H(n)$ se utilizan para obtener el pronóstico $\widetilde{X}$(n).

Algoritmo HSVD

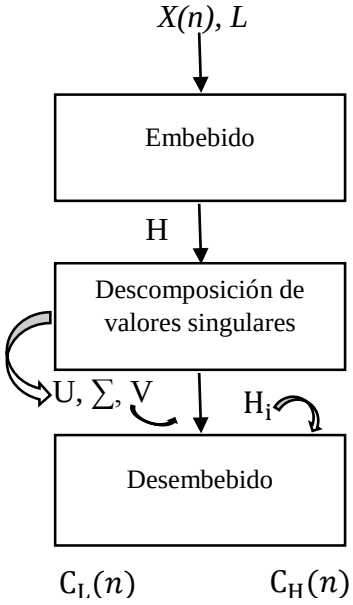

Figura 2: Algoritmo HSVD (Barba,2018).

El algoritmo HSVD primero realiza un embebido usando una matriz de recorrido, para luego descomponerla en valores singulares, matriz de vectores singulares izquierdos, matriz de vectores propios derecho, luego de esto se extraen las componentes C_L y C_H.

Después de la aplicación de las tres técnicas a la serie temporal, se procede a buscar un modelo. El mejor modelo que se adapta a nuestra investigación es el autoregresivo porque uno de nuestros objetivos que queremos resolver es el pronóstico, este modelo será calibrado para el horizonte próximo (one-step ahead forecasting), el pronóstico de horizonte próximo es ir un paso delante de los valores reales (Xiong, 2013).

El modelo autoregresivo de orden P; donde P es el número de observaciones previas (valores históricos), para una serie temporal el modelo se implementará de la siguiente manera.

$$\widehat{X}_{(n)} = \sum_{i=1}^{p} \alpha i \, Z_i \qquad (13)$$

Donde P es el orden de los polinomios autoregresivos y de la parte del promedio móvil respectivamente, α es el coeficiente de los términos AR(Modelo AutoRegresivo), (n) es el instante de tiempo y Z_i es la primera columna de la matriz regresiva que se forma con los componentes $C_L(n)$ y $C_H(n)$ previamente extraídos en el suavizado (Barba L. M., 2018).

Sin embargo, este modelo de pronóstico cuando se trata de las técnicas media móvil y suavizado exponencial se compone por valores suavizados como se lo observa en la ecuación (14) y cuando se trata de HSVD se conforma por las componentes de alta y baja frecuencia como lo muestra la ecuación (15).

$$Z = (\tilde{x}_n, \tilde{x}_{(n-1)}, \dots, \tilde{x}_{(n-P+1)}) \tag{14}$$

$$Z = (CL_{(n)}, CL_{(n-1)}, \dots, CL_{(n-P+1)}, (CH_{(n)}, CH_{(n-1)}, \dots, CH_{(n-P+1)}) \tag{15}$$

También se parametrizó el modelo autoregresivo en cuanto a variables explicativas, en este caso el número de semanas que vienen a ser los lags; y la muestra de entrenamiento y validación, estas dos variables cambiarán de acuerdo a cada técnica de suavizado hasta encontrar los valores que brinden el mejor resultado.

Fase de evaluación

En esta fase de la metodología se verifica si el modelo creado se ajusta a las objetivos establecidos en la primera fase, para ello una buena forma de evaluar la efectividad del modelo utilizado es utilizando indicadores, se calculan las tres métricas siguientes; MAPE (Error Porcentual Absoluto Medio) que relaciona el error en el pronóstico con la demanda de manera conceptual, RMSE (Error Cuadrático Medio) que mide la cantidad de error que existe en toda la serie temporal y R^2 (Coeficiente de Determinación) que es el porcentaje de variación de respuesta que explica la relación con una o más variables predictores, mientas mayor sea el R^2 mejor será el ajuste del

modelo, el R^2 siempre se encuentra entre 0 y 100% (Ramayah, 2003), estas métricas se calculan con las siguientes ecuaciones.

$$\text{MAPE} = \frac{\sum|x_n - \hat{x}_n|}{\text{n}} \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{(x_n - \hat{x}_n)}{\text{n}}} \tag{17}$$

$$R^2 = 1 - \frac{\sum(x_n - \hat{x}_n)^2}{\sum(x_n - \hat{x}_n)^2} \tag{18}$$

Donde:

x_n= valor observado de la muestra de validación.

$\hat{x}_n$ = valor pronosticado de la muestra de validación.

n = número de valores de la muestra de validación.

Fase de implementación

El objetivo en la última fase de la metodología CRISP-DM es el de explicar al cliente como poner en funcionamiento el modelo que se ha construido en las fases anteriores, así también el de exponer los resultados obtenidos para que se pueda entender fácilmente, los mismos que se mostrarán en el capítulo IV de resultados.

4. CAPÍTULO IV

4.1 RESULTADOS Y DISCUSIÓN

En este capítulo se construyeron nueve modelos de pronóstico, basados en las tres técnicas de suavizado como son media móvil, suavizado exponencial y descomposición por valores singulares de Hankel, cada técnica fue evaluada con el modelo autoregresivo con la finalidad de encontrar y comparar cuál de ellos genera la mayor exactitud, para contar con un modelo que logre explicar de mejor manera el fenómeno estudiado.

En la figura 3 se observan las tres series temporales de donde parte el análisis, estas series corresponden a valores extraídos de la base de datos de publicaciones científicas de docentes en sus diferentes frecuencias, diaria, semanal y mensual, en estas series se aplicaron cada técnica de suavizado.

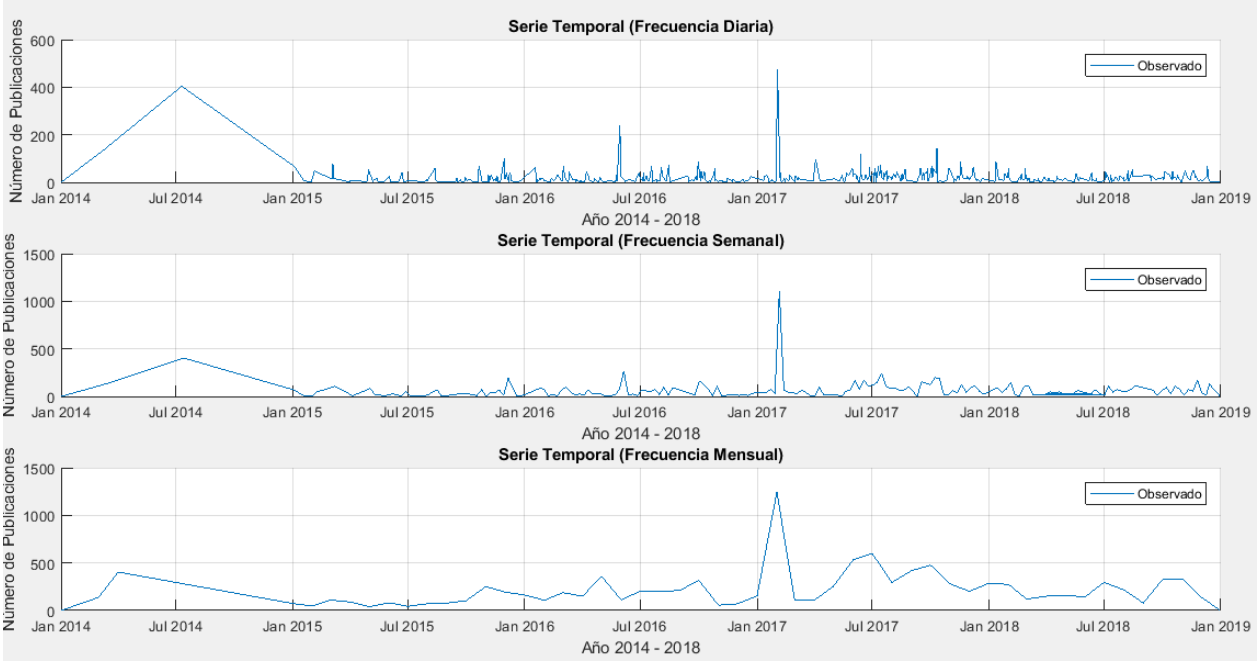

Figura 3: Publicaciones científicas de docentes frecuencia diaria, semanal y mensual.

4.1.1 Análisis con la técnica de suavizado Media Móvil

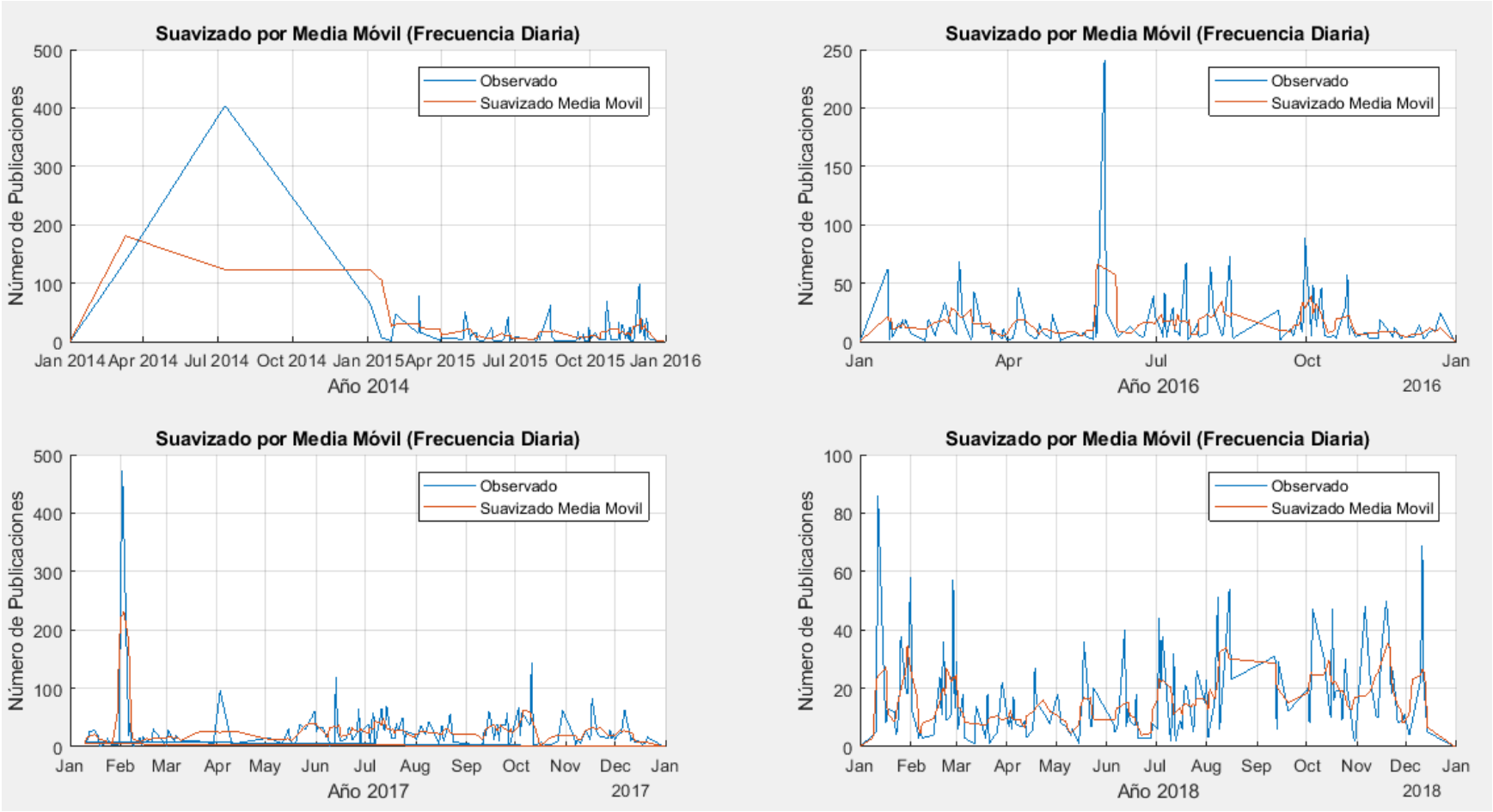

Figura 4: Serie temporal suavizada con Media Móvil (publicaciones diarias).

Para la aplicación de esta técnica de suavizado se utilizó una media móvil de orden 3, para tener una mejor visualización se optó por dividir el análisis en 4 gráficas distintas. Cada gráfica presenta de color azul el valor observado y de color naranja el valor suavizado, en el que la serie temporal muestra menor variabilidad.

La primera figura está comprendida entre los años 2014-2015 que son los años en los que se registran menor número de publicaciones, presentando solo en los días 03 y 07 de octubre del 2014 un mayor registro de publicaciones con un total de 442 publicaciones atribuible a que posiblemente en este periodo no se registraron los datos en todo el año, después para el año 2015 se presenta un cambio de nivel a bajo. La segunda gráfica muestra el año 2016, este año tiene un registro normalizado de las publicaciones, es decir se han registrado un valor más o menos constante de publicaciones diariamente. La tercera gráfica es del año 2017, de igual manera en los días 01, 02, 03 de febrero, existe un mayor registro de publicaciones, por ello se observa una alta variabilidad en la serie temporal. La cuarta gráfica es del año 2018 en esta serie temporal no se observa mucha variabilidad a lo largo de todo el periodo. Con la división de las 4 figuras se puede observar apreciar como la serie temporal y el suavizado cambia drásticamente para cada año lo que quiere decir que esta serie no es estacionaria.

4.1.2 Análisis con la técnica de Suavizado Exponencial

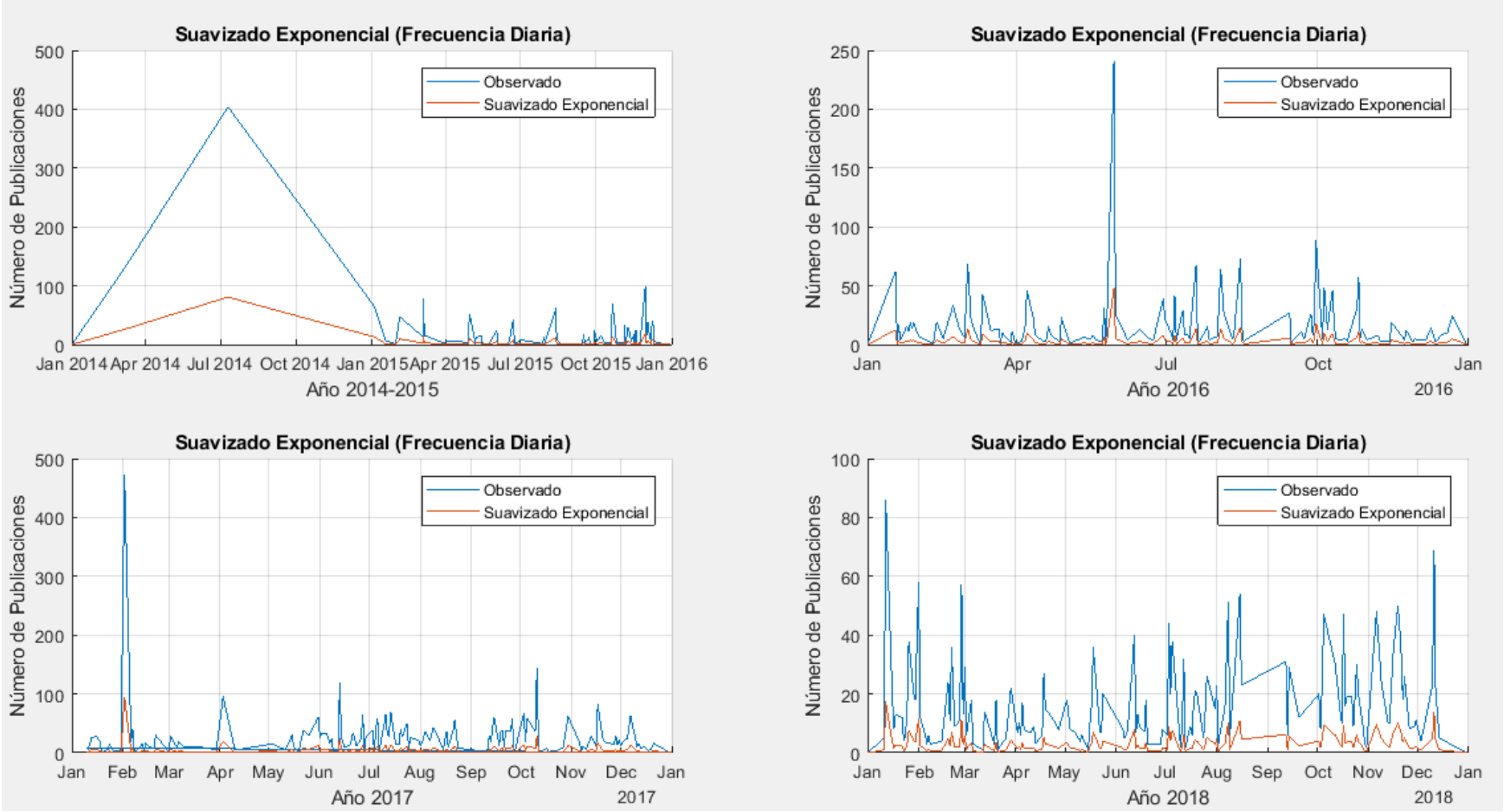

Figura 5: Serie temporal suavizada con Suavizado Exponencial (publicaciones diarias).

Siguiendo los mismos pasos de la técnica anterior se realiza este segundo análisis, por medio de la técnica del suavizado exponencial como lo muestra la figura 5. De igual forma para tener una mejor visualización se optó por dividir el análisis en 4 gráficas distintas como se lo hizo para la media móvil.

Antes de aplicar el proceso para el suavizado con esta técnica se decide el valor que le daremos a la constante de suavizado α (alpha). Para este análisis vamos a considerar que $\alpha = 0.3$ y que el suavizado del día 1 es igual a la demanda real observada en el mismo día, estos valores que seleccionamos para α y la estimación inicial influyen en los resultados del suavizado.

En este análisis de igual forma se puede apreciar en las cuatro gráficas alta variabilidad, además de ello se puede observar que el suavizado exponencial dibujado de color naranja no se mantuvo tan cerca a los valores reales. Sin embargo, se verificará la exactitud de esta técnica en el modelo de pronóstico que se generará en los próximos pasos.

4.1.3 Análisis con la técnica Descomposición por valores singulares de Hankel

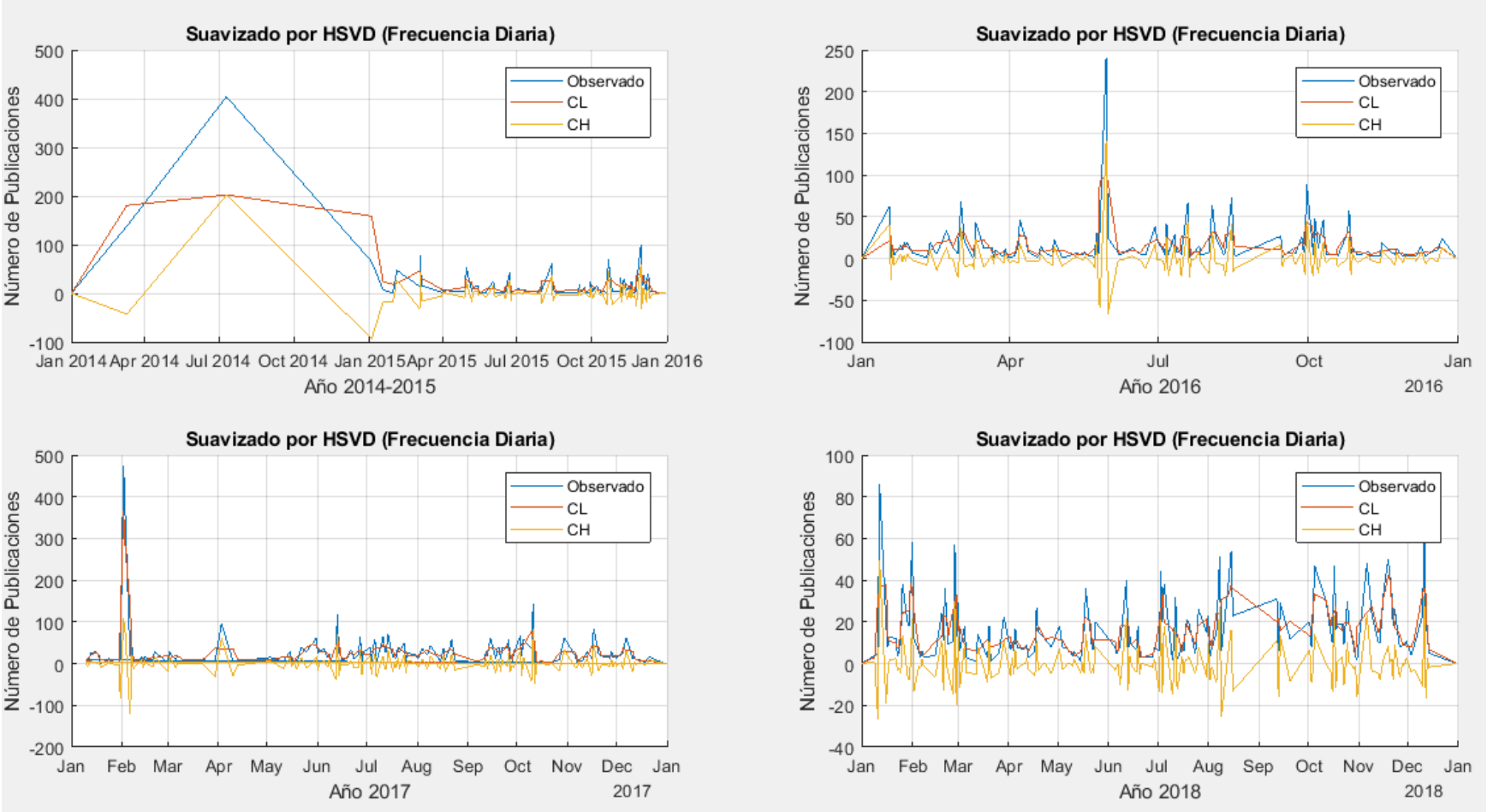

Figura 6: Serie temporal suavizada con HSVD (publicaciones diarias).

Para el tercer suavizado se utilizó la técnica Descomposición por valores singulares de Hankel, al igual que en las dos técnicas anteriores la figura 6 muestra 4 gráficas para su mejor visualización.

En este análisis cada gráfica muestra tres series temporales diferentes, la observada y los componentes C_L y C_H, se puede apreciar la serie C_L de baja frecuencia muestra menor variación, considerando que representa aquellos componentes de mayor duración; mientras que la serie C_H de alta frecuencia muestra mayor variación considerando que representa aquellos componentes de menor duración(duración corta).

Vale recalcar que para el pronóstico de las tres técnicas de suavizado se utilizó el análisis de la serie temporal original con los 1825 valores.

4.1.4 Análisis con las tres técnicas de suavizado con intervalos semanal y mensual

Siguiendo con los mismos pasos de las tres técnicas anteriores se procedió a realizar un análisis con los mismos datos de publicaciones científicas de docentes, pero en este caso dividiendo los datos con una periodicidad semanal entre los años 2014-2018, en la figura 7, se observan tres gráficas con la aplicación de las tres técnicas de suavizado. Para este análisis se aprecia una serie temporal con menor variación que en los análisis anteriores, esto debido a que disminuyen los valores que están siendo suavizados también se observa que el mejor modelado de los datos reales de la serie se obtiene con la técnica HSVD.

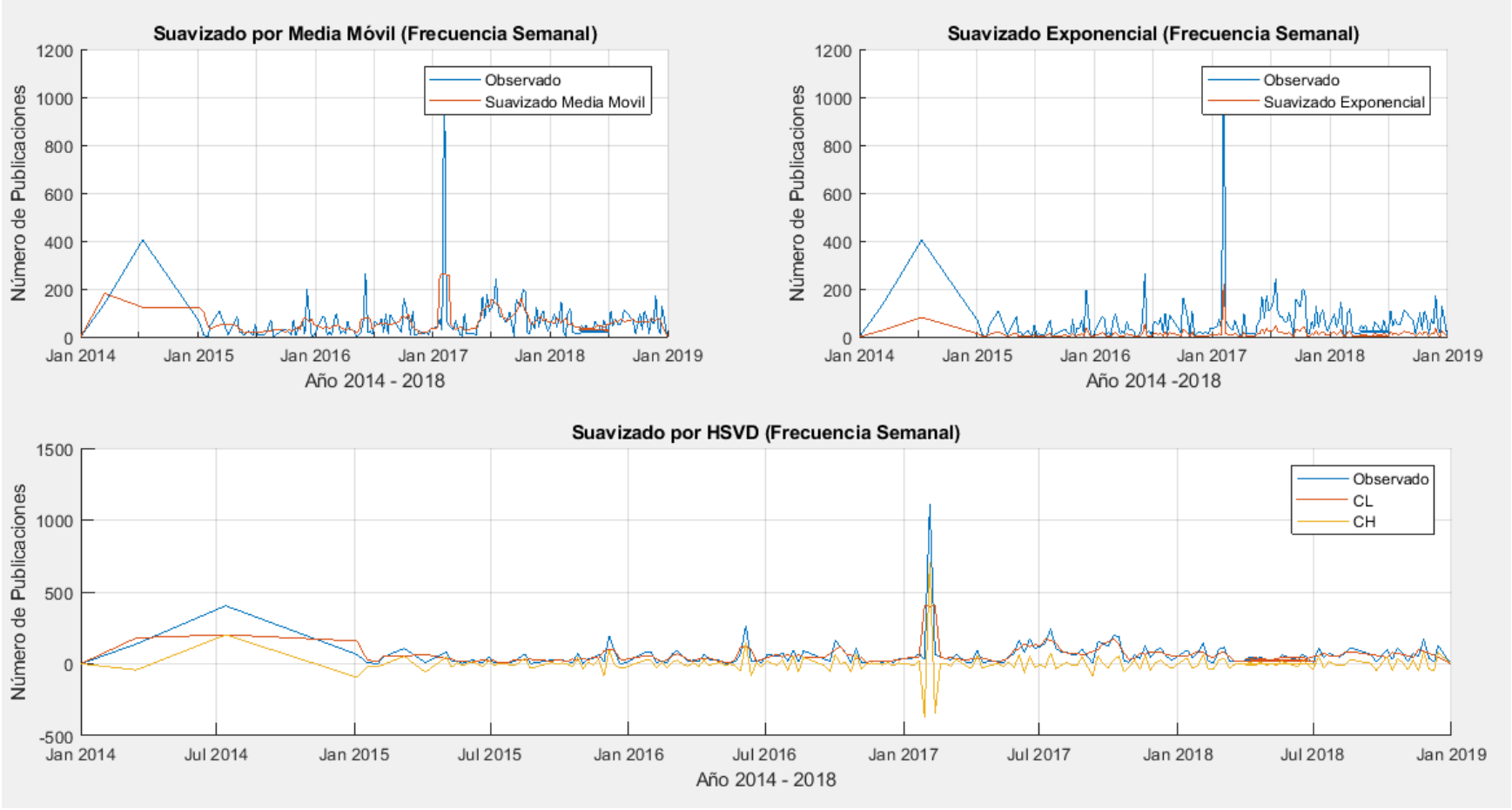

Figura 7: Serie temporal suavizada con las tres técnicas (publicaciones semanales).

Al igual que los análisis anteriores posteriormente se realizó un último análisis, pero ahora con una periodicidad mensual de las publicaciones, es decir se dividió la serie temporal original en los diferentes meses que existen registros de publicaciones, obteniendo así 60 meses entre los años 2014 y 2018. Estos datos son los que ahora forman la serie temporal para el análisis. En la figura 8 se pueden observar los tres análisis con las distintas técnicas, media móvil, suavizado exponencial y HSVD, estas figuras al igual que las anteriores muestran las líneas de los valores observados y los estimados, para este caso, la serie temporal tiene menor variabilidad que en los análisis anteriores esto se debe a que cada mes cuenta con un número similar de publicaciones, sin embargo, en el mes de junio del 2017 se puede observar una subida de nivel, debido a que este mes tiene un mayor registro de publicaciones.

En estas figuras también se observa en la serie suavizada de color naranja como actúa el suavizado en cada una de las técnicas, se puede visualizar que el mejor suavizado nos está brindando la técnica HSVD y el peor suavizado la técnica Suavizado Exponencial.

Con estos datos suavizados se procede a realizar el pronóstico para saber que técnica genera mejor exactitud para el modelo creado y así poder dar un mejor criterio de cada una de las técnicas, cabe resaltar que cada técnica actúa de forma diferente de acuerdo a la investigación que sean aplicadas.

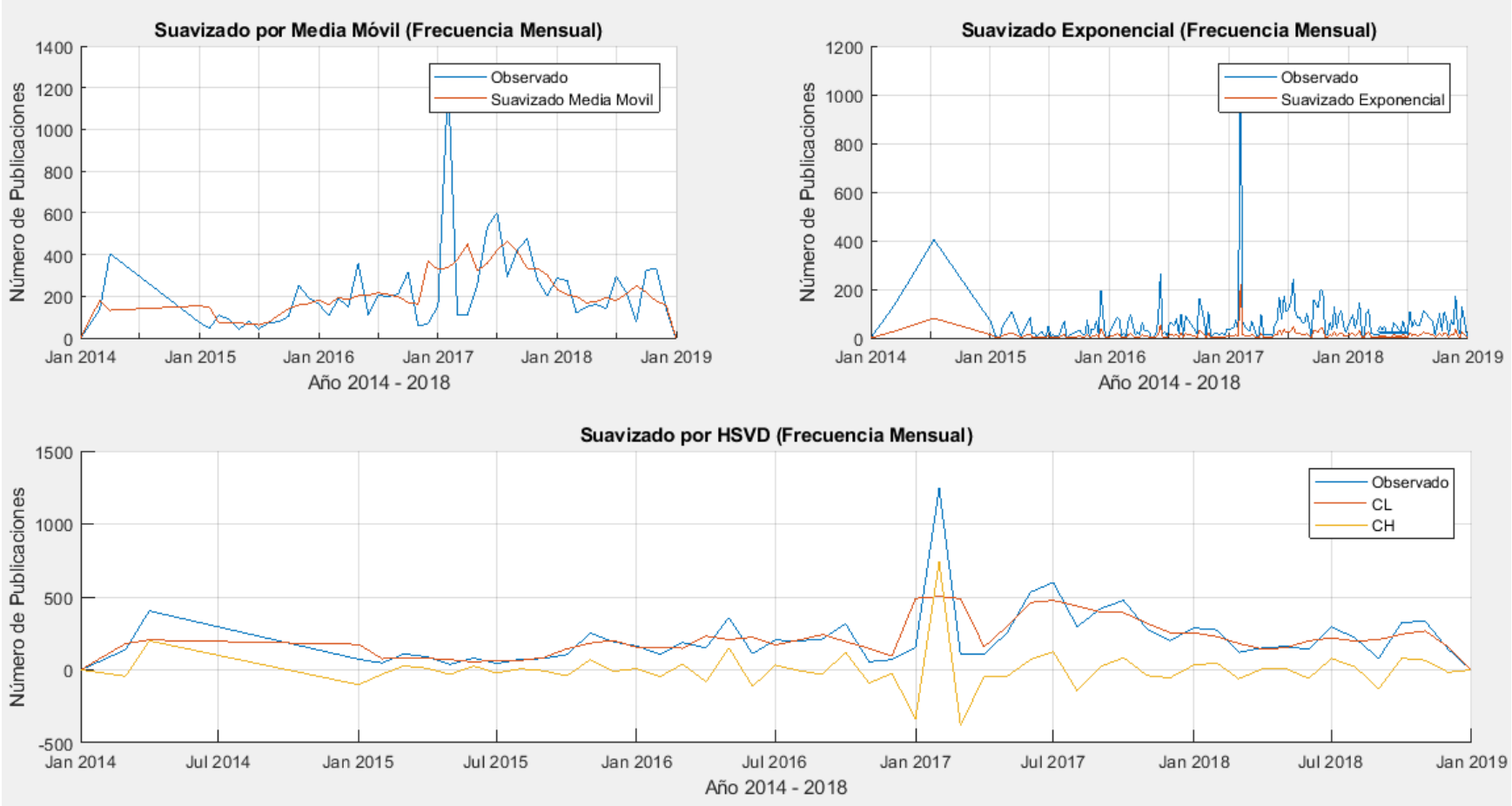

Figura 8: Serie temporal suavizada con las tres técnicas (publicaciones mensuales).

4.1.5 Pronóstico autoregresivo utilizando las tres técnicas de suavizado e intervalos diarios

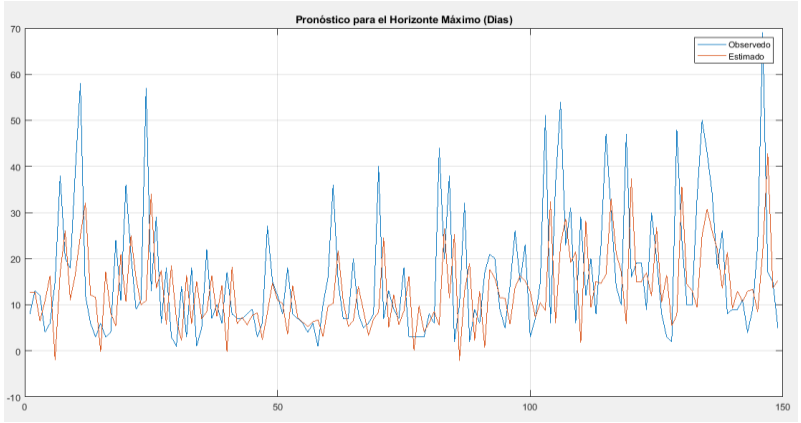

Figura 9: Pronostico basado en Media Móvil (diaria).

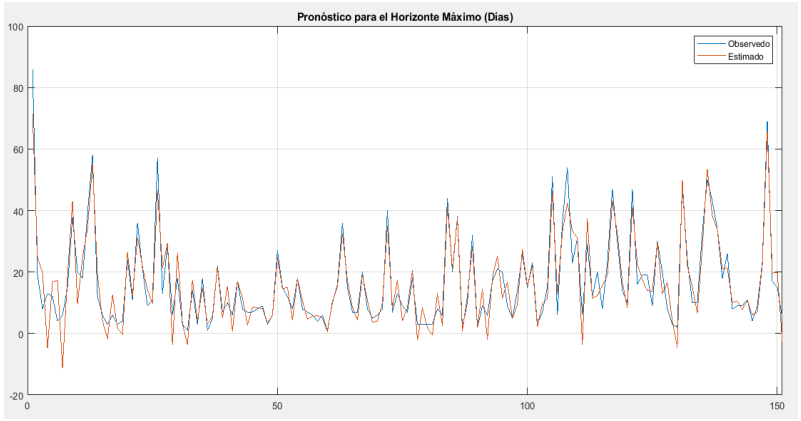

Figura 10: Pronostico basado en HSVD (diaria).

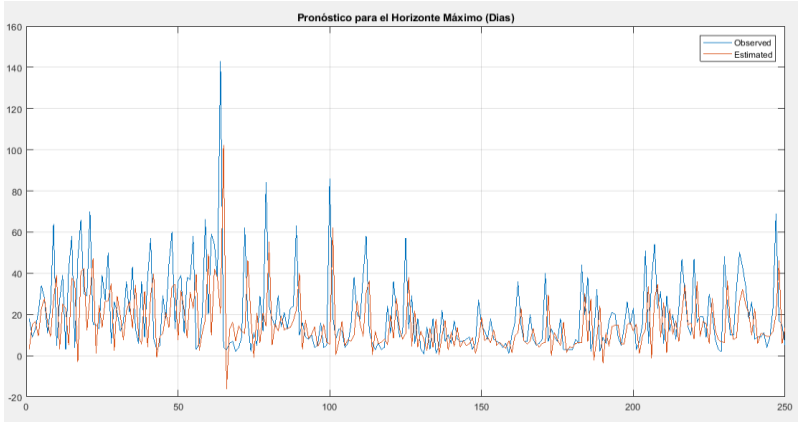

Figura 11: Pronostico basado en Suavizado Exponencial (diaria).

A continuación, se puede observar en las figuras 9, 10 y 11 la aplicación del modelo de pronóstico autoregresivo o regresión lineal, se han utilizado los valores suavizados de las tres técnicas para posteriormente poder comparar que técnica genera mayor exactitud para el pronóstico.

Por ello cada modelo contiene parámetros que fueron calibrados hasta obtener el mejor resultado de las métricas que van a ser comparadas;

- Para el pronóstico con media móvil se calibró el modelo con un porcentaje de entrenamiento de los datos = 0.70, la memoria de pronóstico (lags) = 9, y para un máximo horizonte = 1.
- Para el pronóstico con suavizado exponencial se calibro el modelo con un porcentaje de entrenamiento = 0.80, lags = 15 y un máximo horizonte = 1.
- Para el pronóstico con HSVD se calibro el modelo con un porcentaje de entrenamiento de datos = 0.70, lags = 19 y un máximo horizonte = 1;

Una vez calibrados los modelos con los mejores parámetros se obtiene los resultados que muestra la tabla 4.

4.1.6 Interpretación del pronóstico

Tabla 4. Métricas de exactitud con frecuencia diaria.

	MAPE	RMSE	R^2
Media Móvil	*68.89*	*5.21*	*80.48*
Suavizado Exponencial	*93.61*	*14.43*	*72.99*
HSVD	*10.80*	*2.85*	*98.85*

En la tabla 4 se visualizan los resultados obtenidos de los tres modelos de pronóstico con las tres técnicas de suavizado, con estos datos se puede realizar la comparación para identificar cuál de los tres modelos generan mayor exactitud para el pronóstico.

Se puede observar en la tabla 4 que el mejor resultado está dado por la combinación del modelo autoregresivo con la técnica HSVD, generando un MAPE = 10.80%, RMSE = 2.85% y un R^2 = 98.85%; seguido de la combinación del modelo autoregresivo con la técnica Media Móvil con un MAPE = 68.89%, RMSE = 5.21% y un R^2 = 80.48%; el peor resultado obtenido en este análisis es la combinación del modelo autoregresivo con el Suavizado Exponencial con un MAPE = 93.61%, un RMSE = 14.43% y un R^2 = 72.99%.

Siguiendo la misma metodología, se implementaron 6 modelos de pronóstico, a partir de 2 nuevas series de tiempo, en las cuales se utilizaron las frecuencias semanal y mensual respectivamente; los resultados obtenidos se muestran en las tablas 5 y 6.

Tabla 5. Métricas de exactitud con frecuencia semanal.

	MAPE	RMSE	R^2
Media Móvil	*64.12*	*18.62*	*73.63*
Suavizado Exponencial	*65.60*	*48.15*	*33.13*
HSVD	*21.48*	*7.21*	*94.55*

En la tabla 5 se puede apreciar que el mejor resultado sigue dando la combinación del modelo autoregresivo con la técnica HSVD con un MAPE = 21.48%, RMSE 7.21% y un R^2 = 94.55%;

como segundo mejor resultado tenemos la combinación del modelo autoregresivo con la técnica Media Móvil con un MAPE = 64.12%, RMSE = 18.62% y un R^2 = 73.63%; sin embargo, la combinación del modelo autoregresivo con la técnica Suavizado Exponencial generan valores parecidos a la técnica Media Móvil con un MAPRE 65.60%, RMSE 48.15% y un R^2 = 33.13%, entonces con este pronóstico se puede deducir que los resultados de estos tres modelos obtenidos con la serie temporal con frecuencia semanal tiene un parecido a los resultados de los modelos con la frecuencia diaria.

Tabla 6. Métricas de exactitud con frecuencia mensual.

	MAPE	RMSE	R^2
Media Móvil	*31,67*	*15,06*	*60.15*
Suavizado Exponencial	*58.60*	*11.25*	*42.97*
HSVD	*24.67*	*14.96*	*86.41*

En la tabla 6 se muestra los resultados obtenidos con la serie temporal de la frecuencia mensual, en esta tabla los resultados no cambian a diferencia de los 6 modelos anteriores. El mejor resultado de igual forma brinda la combinación del modelo autoregresivo con la técnica HSVD con un MAPE = 24.67%, un RMSE = 14.96% y un R^2 = 86.41%; como segundo mejor resultado está la combinación del modelo autoregresivo con la técnica media móvil con un MAPE = 31.87%, RMSE = 15.06% y un R^2 = 60.15%; y como el peor resultado tenemos la combinación con la técnica Suavizado Exponencial con un MAPE = 58.60%, RMSE = 11.25% y un R^2 = 42.97%.

Promedio de los 9 modelos de pronóstico

Al realizar un promedio de todos los modelos realizados en esta investigación se obtiene la tabla 7 con la cual se llega a la conclusión que, para la investigación realizada con los datos del sistema académico de la UNACH, la combinación del modelo autoregresivo con la técnica HSVD brinda la mayor exactitud al momento de realizar pronóstico, el segundo resultado genera la combinación del modelo auto regresivo con la técnica media móvil, y la combinación que peor exactitud genera es la combinación del modelo autoregresivo con la técnica de suavizado exponencial.

Tabla 7. Promedio de métricas del Pronóstico.

	MAPE	RMSE	R^2
Media Móvil	*54.89*	*12.96*	*71.42*
Suavizado Exponencial	*72.60*	*24.61*	*49.69*
HSVD	*18.91*	*8.37*	*93.27*

4.1.7 Portal Interactivo

Figura 12: Portal interactivo.

El portal interactivo está diseñado en el lenguaje de programación PHP conjuntamente con el gestor de base de datos MySQL donde se almacenarán los datos de la serie temporal original y suavizada. En la figura 12 se puede observar la interfaz del portal, el cual contiene un inicio de sesión donde permitirá el registro de los usuarios que se lo puede visualizar en anexos en la figura 18 y la página principal donde se apreciará el suavizado de las series temporales creadas por el usuario.

La serie temporal la pueden cargar los usuarios en un archivo de Excel con el formato .csv, en el botón seleccionar archivo se almacenará la serie creada, y generará de manera automática el suavizado con Media Móvil y Suavizado Exponencial.

Cabe recalzar que el usuario no debe tener conocimientos avanzados para poder utilizar este portal interactivo y será de gran aporte a la toma de decisiones deacuerdo al propósito que sea utilizado.

5. CONCLUSIONES

- Se logró analizar el sistema académico de la UNACH específicamente la base de datos de publicaciones de investigación de docentes, aplicando las técnicas de minerías de datos, obteniendo resultados positivos con lo cual se cumplió con el objetivo plantado de contribuir al desarrollo de la institución y garantizar la integridad de los datos analizados.
- Las técnicas de suavizado fueron aplicadas con la herramienta Matlab, la cual cumple con las políticas en minería de datos para realizar procesos de predicción, a través de estas técnicas se logró determinar que técnica de suavizado generar mayor exactitud a la hora de pronosticar.
- Se implementó un modelo autoregresivo de pronóstico, el cual fue de gran aporte a la investigación, a través de las técnicas de suavizado, se generó información valiosa para así garantizar los resultados finales.
- Para la implementación de estos modelos de pronóstico se aplicó la metodología CRISP-DM, esta metodología es una de las más aplicadas al momento de realizar pronósticos con minería de datos, está compuesta por seis fases que ayudan a verificar la calidad de los datos, generando una confiabilidad e integridad de los resultados obtenidos.
- La investigación permitió desarrollar nuevos conocimientos en lo referente a la analítica de datos, específicamente en las técnicas de suavizado las cuales permitieron la investigación, aportando también con nuevos hallazgos, por ejemplo, la técnica HSVD genera mejor exactitud con el grupo de datos de investigaciones científicas de docentes de la UNACH.

6. RECOMENDACIONES

- Al momento de realizar un análisis de datos es indispensable primero conocer y comprender las distintas técnicas a ser utilizadas, por ello se recomienda contar con fuentes necesarias a donde puedan recurrir en caso de necesitarlas, para que estas ayuden como base en el proceso de la aplicación de la minería de datos.
- Es necesario determinar los objetivos principales y delimitar el alcance de la investigación, para evitar la pérdida de interés o enfoque de la investigación.
- Es recomendable para la aplicación del modelo autoregresivo realizar varias interacciones cambiando el número de semanas o lags, la muestra de validación y entrenamiento hasta encontrar el modelo que genere mejor exactitud.
- Se debería profundizar en estudios de analítica de Datos, debido a que en la actualidad todas las empresas, industrias, organizaciones e instituciones educativas tanto públicas como privadas, generan datos valiosos para la sostenibilidad de las mismas.
- Extender el presente estudio a horizontes de pronostico mayores, con la finalidad de contribuir de manera sostenible la toma de decisiones en las actividades académicas y de investigación de la UNACH.

7. BIBLIOGRAFÍA

Arenas, J. S. (2009). *DESARROLLO DE UN MODELO DE PRONOSTICO DE CAUDALES.* Medellin: Universidad Nacional de Colombia.

Arias, M. Á. (2017). *Aprende Programación Web PHP y MySQL*. IT Campus Academy - Segunda Edición.

Barba, L. M. (2018). *Forcasting Based on Hankel Singular Value Descomposition.* Valparaíso, Chile: Pontífica Universidad Católica de Valparaíso.

Barba, L., Rodriguez, N., & Montt, C. (2014). Smoothing Strategies Combined with ARIMA and Neural networks to improve the forecasting of traffic accidents. *The Scientific World Journal*.

Bianka, H. . (2012). Minería de datos en educación.

Carollo, C. M. (2012). Regresion Lienal Simple . *Departamento de estadistica e investigación operativa* , 2 - 10.

Chen, W. M. (2016). SVD-based technique for interference cancellation and noise reduction in NMR measurement of time-dependent magnetic fields. *Sensors*, 16(3), 313.

Combaudon, S. (2018). *MySQL 5.7 Administración y Optimización.* Barcelona: ENI.

Coutin, G. (2001). *Las series temporales.* Habana.

Echaverría, J. D., Gómez, C. A., Aristizábal, M. U., & Vanegas, J. O. (2010). El método analítico como método natural. *Universidad de Antioquia, Colombia*.

Eckart, C. y. (1930). Una transformación de eje principal para matrices no hermitianas. *Boletín de la American Mathematical Society*, 45 (2), 118-121.

Elias, T. (2011). Learning analytics. Learning 1-22.

Ferrero, E., Castro, R., Pérez, J., & Arcos, P. (2017). *La mortalidad por desastres en España: Un análisis del periodo 1950-2012.* Granada: Index Enferm vol.26 no.1-2.

Galán, V. (2015). *Aplicación de la Metodología CRISP-DM a un Proyecto.* Madrid: Universidad Carlos III de Madrid.

Galindo, Á. J., & Garcia, H. Á. (2010). Mineria de Datos en la Educación . *Universidad Carlos III*, 1-8.

Golub, G. H. (1965). Descomposición de valores singulares y soluciones de mínimos cuadrados. *En Álgebra lineal* , 134-151.

Gras, J. A. (2001). *Diseños de series temporales: técnicas de análisis (Vol. 46).* Edicions Universitat Barcelona.

Griffies, S., Perrie, W., & Hull, G. (2013). Elements of style for writing scientific journal articles. *Publishing Connect, Elsevier*, 20-50.

Gutierrez. (2015). Learning Analytics o analítica de aprendizaje.

Huapaya, C. R., Lizarralde, F. A., Arona, G. M., & Massa, S. M. (2012). Minería de datos educacional en ambientes virtuales de aprendizaje. *In XIV Workshop de Investigadores en Ciencias de la Computación.*

Hussien, H. H., Eissa, F. H., & Awadalla, K. E. (2017). Statistical methods for predicting malaria incidences using data from Sudan. *Malaria research and treatment*.

Jiménez, A., & Álvarez, H. (2010). Minería de Datos en la Educación. *Inteligencia en Redes de Comunicación* , 30.

Kalekar, P. S. (2004). Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology*, 4329008(13).

López, C. P. (2007). *Minería de datos: técnicas y herramientas.* Editorial Paraninfo.

Mauricio, J. A. (2007). Introducción al Análisis de Series Temporales. *Universidad Complutense de Madrid*, 1-33.

Moine, J. M., Haedo, A. S., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *In XIII Workshop de Investigadores en Ciencias de la Computación*, http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y.

Montero, J. (2007). *Mineria de Datos, Tecnicas y Herramientas.* Madrid - España: Clara M de la Fuente Rojo.

Ochoa, M. A., Britos, P. V., & Martinez, R. G. (2006). Una Protofase de Entendimiento del Negocio para Metodologías de Desarrollo de Sistemas. *In XII Congreso Argentino de Ciencias de la Computación*.

Olmedo, E., Valderas, J. M., Mateos, R., & Gimeno, R. (2004). Utilización de redes neuronales en la caracterización, modelización y predicción de series temporales económicas en un entorno complejo. *Inteligencia Artificial, 8(23)*, 7-25.

Pajuelo, J. G., & Lorenzo, J. M. (1995). Análisis y predicción de la pesquería demersal de las Islas Canarias mediante un modelo ARIMA. *Departamento de Biología, Universidad de Las Palmas de Gran Canaria,*.

Pérez, C. (2005). Métodos estadísticos avanzados con SPSS. *Thompson. Madrid*.

Ramayah, T. J. (2003). Receptiveness of internet banking by Malaysian consumers. *The case of Penang. Asian Academy of Management Journal*, 8(2), 1-29.

Rodríguez, C. (2016). *Modelos no lineales de pronóstico de series temporales.* Córdoba, Argentina: UNIVERSIDAD NACIONAL DE CÓRDOBA.

Rodriguéz, K. V. (2018). Datos abiertos para el desarrollo de. *Universidad de Alicante*, 8-20.

Walpole, R. E. (2012). *Probabilidad y Estadística para ingeniería y ciencias* . México: Pearson Educación - Novena Edicion .

Xiong, T. B. (2013). *Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices.* 40, 405-415: Energy Economics.

8. ANEXOS

8.1. Pronóstico con regresión lineal utilizando las tres técnicas de suavizado e intervalos semanales

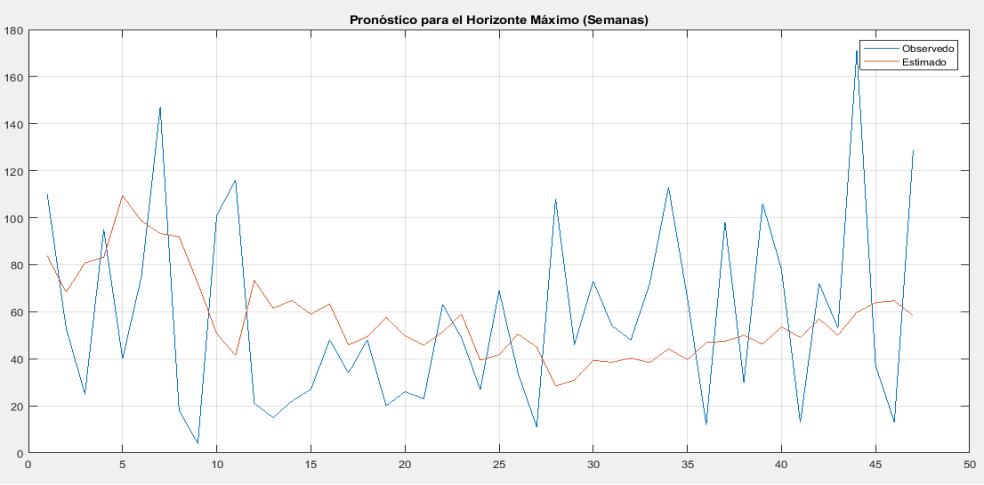

Figura 13: Pronostico basado en Media Móvil (semanal).

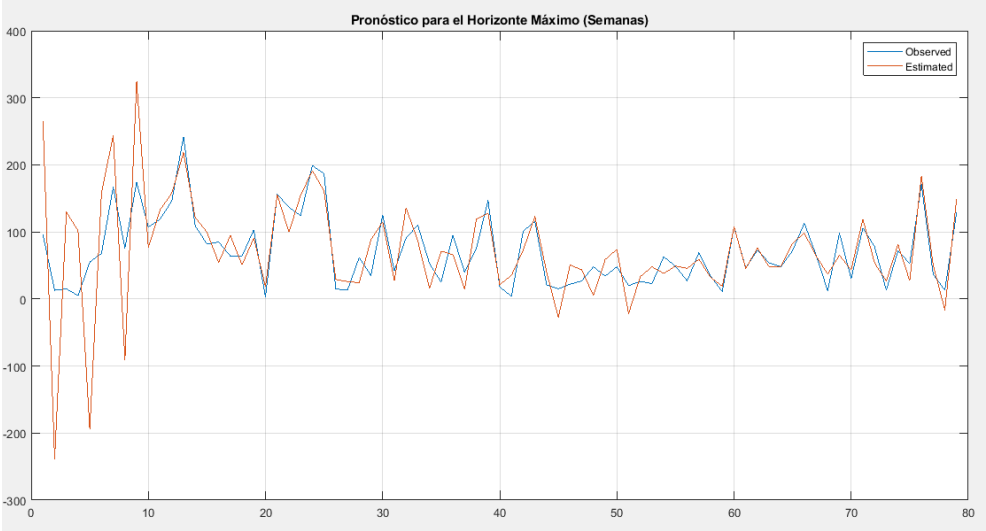

Figura 14: Pronostico basado en Suavizado Exponencial (semanal).

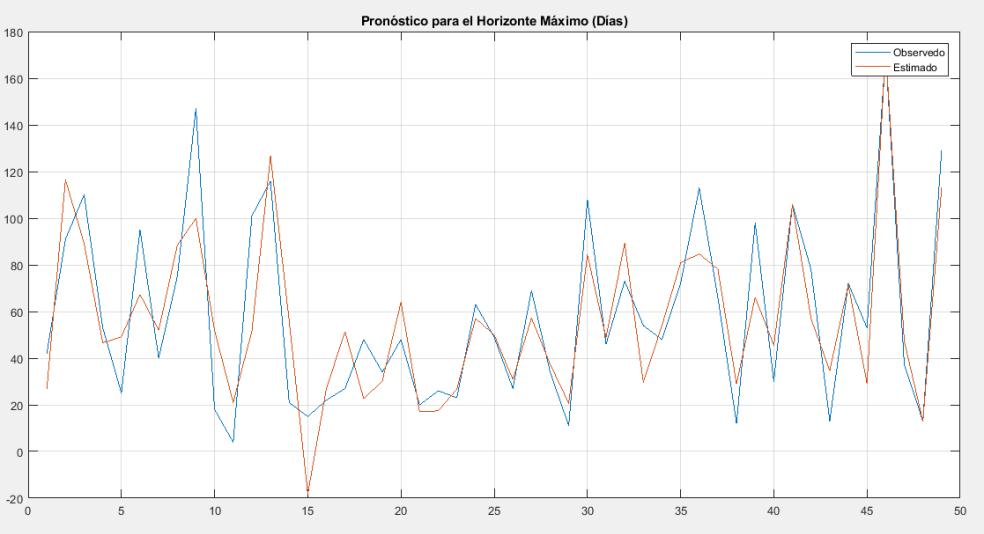

Figura 15: Pronostico basado en HSVD (semanal).

8.2.Pronóstico con regresión lineal utilizando las tres técnicas de suavizado e intervalos mensuales

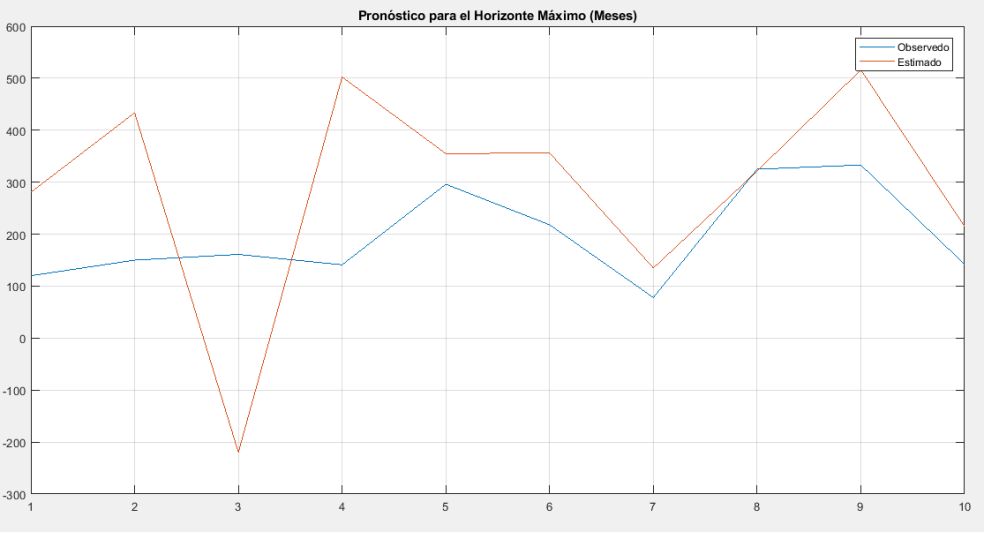

Figura 16: Pronostico basado en Media Móvil (mensual).

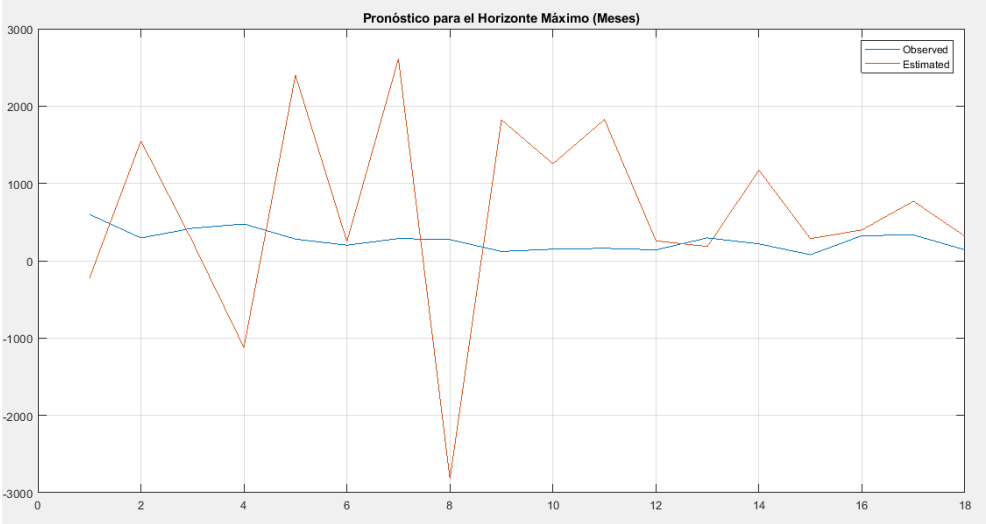

Figura 17: Pronostico basado en Suavizado Exponencial (mensual).

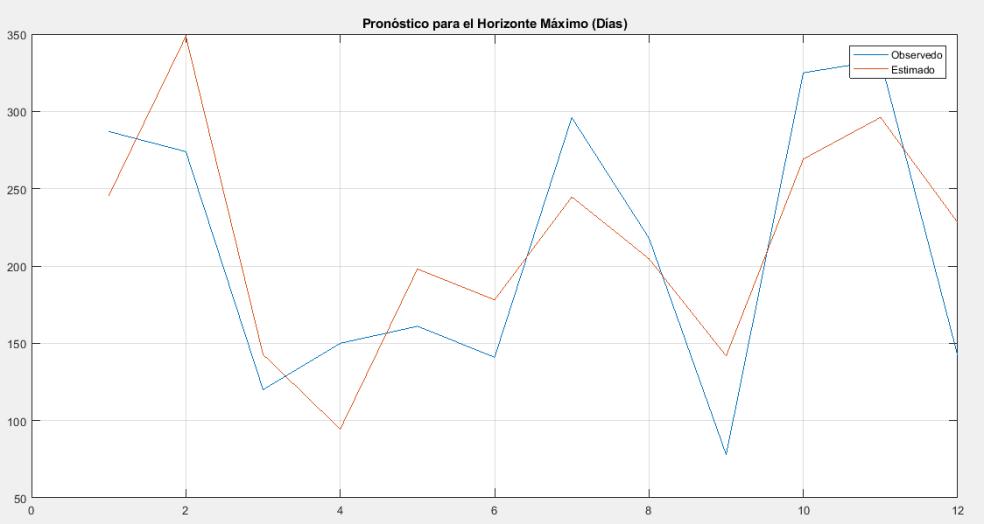

Figura 18: Pronostico basado en HSVD (mensual).

8.3.Inicio de sesión del portal interactivo

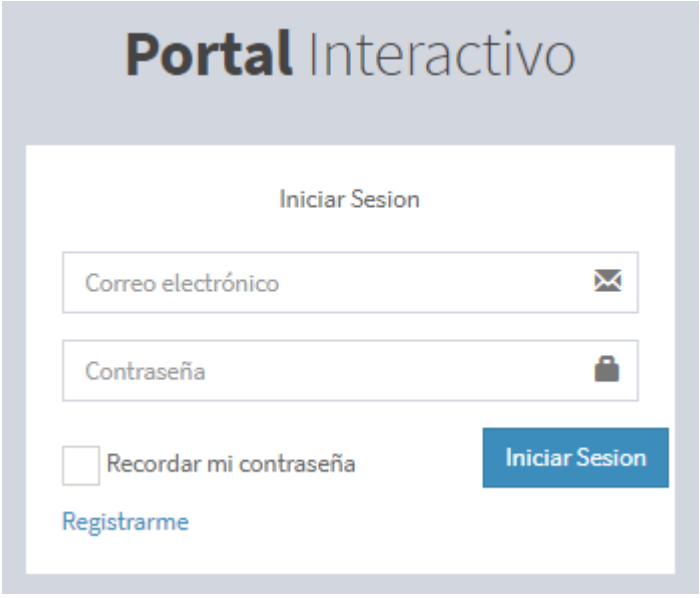

Figura 19: Inicio de sesión portal interactivo.

8.4.Botón para seleccionar serie temporal a suavizar

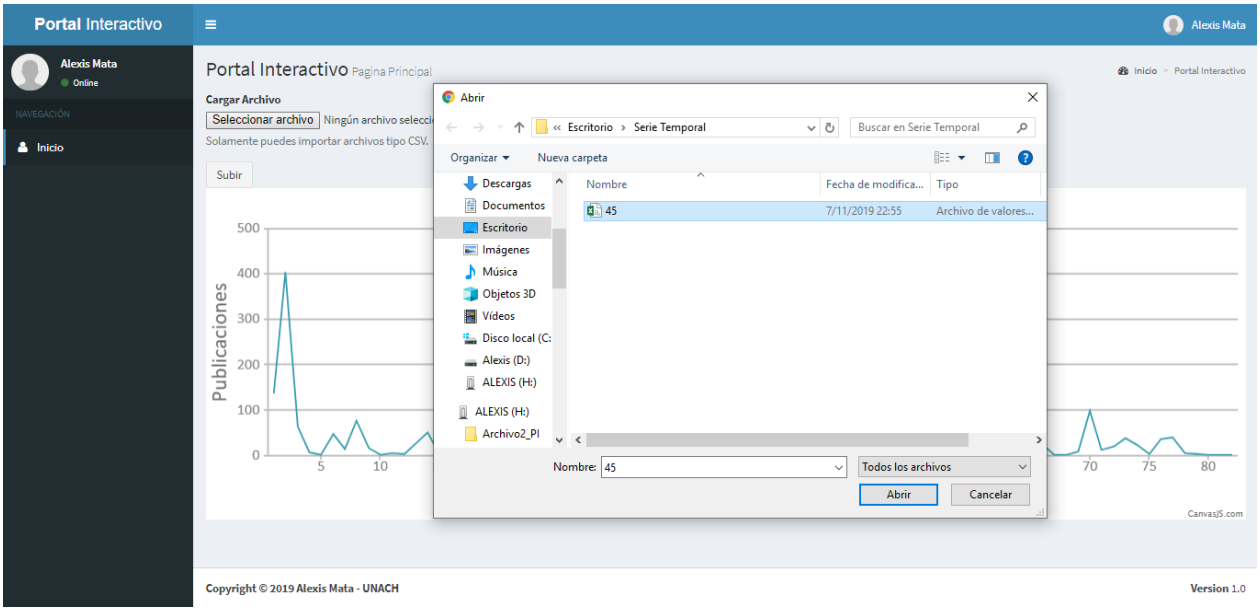

Figura 20: Botón para cargar la serie temporal.

8.5.Programación del portal interactivo en el lenguaje de programación en PHP

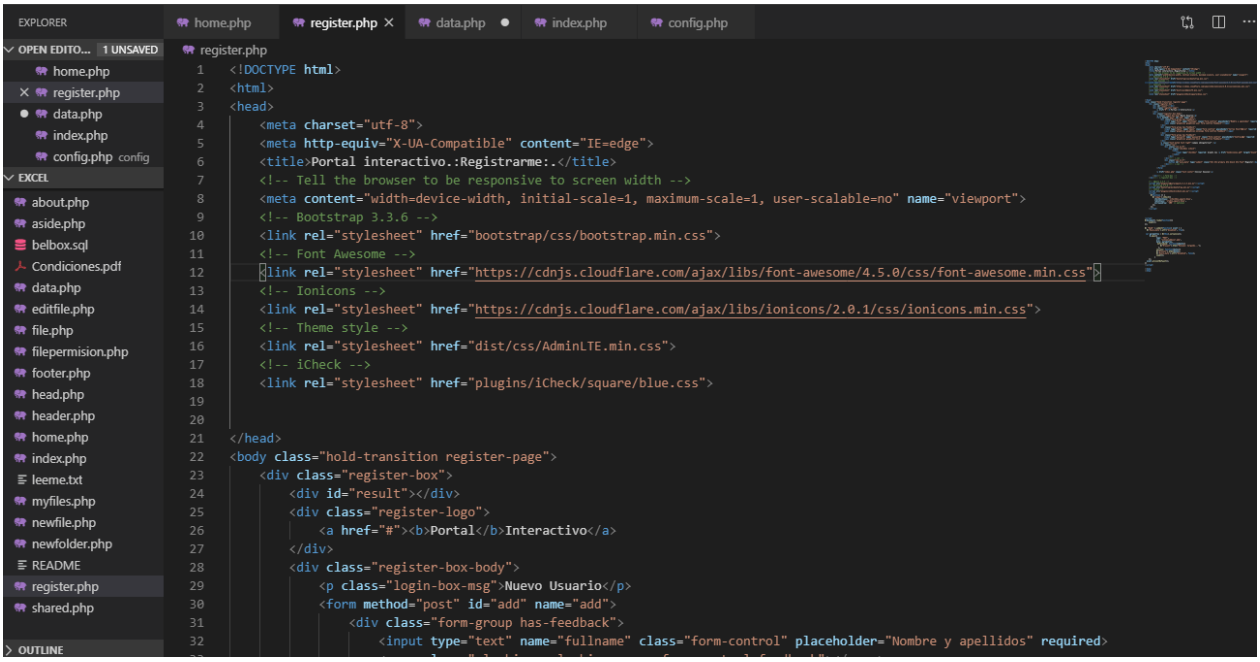

Figura 21: Programación del portal interactivo.

8.6.Base de datos del portal interactivo en MySQL

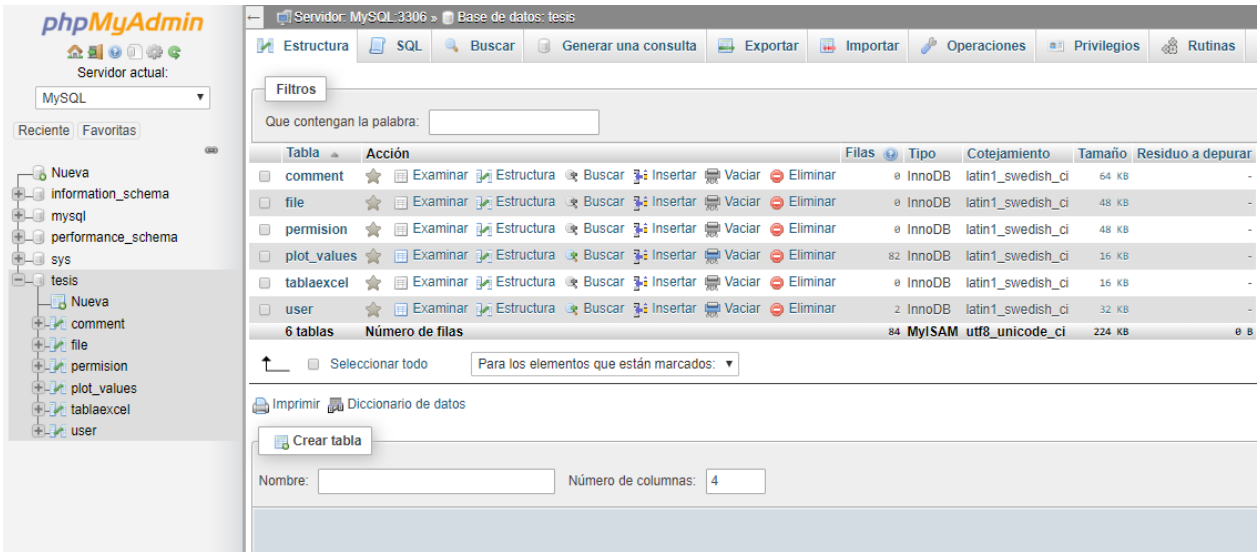

Figura 22: Base de datos en MySQL.