Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Agrupamiento no supervisado de series de tiempo epidemiológicas de México entre 2005 y 2015

por

José Alberto Benavides Vázquez

como requisito parcial para obtener el grado de

maestría en ciencias

en Ingeniería de Sistemas

Agosto 2019

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección de Estudios de Posgrado

Agrupamiento no supervisado de series de tiempo epidemiológicas de México entre 2005 y 2015

por

José Alberto Benavides Vázquez

como requisito parcial para obtener el grado de

maestría en ciencias

en Ingeniería de Sistemas

Agosto 2019

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis "Agrupamiento no supervisado de series de tiempo epidemiológicas de México entre 2005 y 2015", realizada por el alumno José Alberto Benavides Vázquez, con número de matrícula 1373079, sea aceptada para su defensa como requisito parcial para obtener el grado de Maestría en Ciencias en Ingeniería de Sistemas.

El Comité de Tesis

Dr. José Arturo Berrones Santos	Dra. Satu Elisa Schaeffer
Co-Asesor	Co-Asesora

Dra. María Guadalupe Villarreal Marroquín

Revisora

Vo. Bo.

Dr. Simón Martínez Martínez

Subdirector de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, agosto 2019

A mis padres, hermano y esposa.

Agradecimientos

Deseo agradecer a la Universidad Autónoma de Nuevo León (UANL) la oportunidad que me ha brindado de realizar mis estudios de posgrado. A la Facultad de Ingeniería Mecánica y Eléctrica (FIME) por el apoyo brindado durante mis estudios de maestría. Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado mediante una beca de estudios de tiempo completo.

Quedo agradecido al Posgrado en Ingeniería de Sistemas (PISIS) por darme la oportunidad de realizar mis estudios de maestría y en especial a mis asesores Arturo Berrones y Elisa Schaeffer quienes me orientaron en este proceso y me compartieron sus conocimientos e inquietudes. También a mi revisora, Guadalupe Villarreal, por aceptar formar parte del comité de esta tesis. Valoro los comentarios y correcciones de mis compañeros de las clases Redacción Científica en Inglés e Inteligencia Artificial, y de mis compañeros de generación en la maestría.

Agradezco a Gabriela Sánchez por proporcionarme una plantilla de `beamer`, a Miguel Mata por publicar una plantilla para tesis, a Alejandro Benavides por facilitarme una plantilla actualizada del mismo documento, y a José Vargas, quien me ha guiado en los vericuetos digitales donde la Secretaría de Salud de México comparte sus datos.

Resumen

José Alberto Benavides Vázquez.

Candidato para obtener el grado de Maestría en Ciencias en Ingeniería de Sistemas.

Universidad Autónoma de Nuevo León.

Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio: Agrupamiento no supervisado de series de tiempo epidemiológicas de México entre 2005 y 2015.

Número de páginas: 65.

Objetivos y método de estudio: El objetivo consiste en agrupar series de tiempo de registros epidemiológicos semanales de México entre 2005 y 2015 con la finalidad de describirlos en términos de sus características, además de analizar si existe alguna relación estadísticamente significativa que permita asociar dichas series de tiempo a la clasificación asignada a cada enfermedad por la Organización Mundial de la Salud. Esto se logra, primero, mediante la extracción, limpieza y depurado de los datos que contienen esos reportes; en segundo lugar, convirtiendo tales datos en series de tiempo semanales por enfermedad a nivel nacional normalizados contra el total de derechohabientes interpolado entre los valores obtenidos del INEGI de los años 2005 y 2015; después se interpolan los registros faltantes de dichas series de

tiempo; posteriormente, se eligen series de tiempo con al menos cinco años de registros completos para estandarizar el mínimo de información que ofrecen las series de tiempo; luego, se extraen la pendiente, ordenada en el origen y autocorrelaciones de cada serie de tiempo; enseguida, se analizan dichas características por una matriz de correlación y con el método del umbral de varianza se seleccionan aquéllas con varianza superior a la mediana de los datos; ahora, se separan en conjuntos de entrenamiento y prueba a partir de tamaños determinados por la combinación de tamaños con menores sesgo y varianza; con los datos seleccionados separados en tamaños de entrenamiento y prueba determinados, se determina el número de centroides del algoritmo de agrupamiento con el método del codo; a continuación, se agrupan los registros elegidos utilizando el número de centroides determinado por el método del codo; por último se describieron los grupos resultantes por sus características y se examinó si los grupos resultantes guardaban relación con la clasificación propuesta por la Organización Mundial de la Salud.

Resultados: Se logran extraer y limpiar datos de archivos `PDF` publicados por la Secretaría de Salud durante el decenio 2005–2015. Estos datos constituyen información relevante para el estudio de focos epidemiológicos a nivel nacional y estatal. A partir de estos datos, se obtienen series de tiempo de los casos registrados por cada enfermedad normalizados por la derechohabiencia registrada a nivel nacional a lo largo del decenio señalado. Estas series de tiempo fueron agrupadas de modo tal que pueden caracterizarse por sus propiedades temporales. El 60 % de los registros, la mayoría, son relativos a enfermedades infecciosas parasitarias, mientras que la mayoría de las consultas atendidas fueron relacionadas con afecciones respiratorias seguidas por accidentes y envenenamientos, en tanto que son mínimas las consultas por enfermedades originadas en el periodo perinatal. Se identificaron tres enfermedades con tendencia creciente durante el periodo: La infección asintomática

por VIH, la tos ferina y el cólera, mas es importante recalcar que el cólera aparece al alza debido a que en 2013 hubo una reaparición de dicha enfermedad en Hidalgo debida al paso de los huracanes Ingrid y Manuel. Del análisis de autocorrelaciones por matriz de correlación se intuye que hay una fuerte correlación entre el número de consultas realizadas en un mismo mes, mientras que se esperaría una correlación inversa entre consultas de seis a ocho meses de diferencia lo que da la idea de que los registros de consultas son estacionarios. Las autocorrelaciones con retrasos de cuatro a ocho meses suelen ser estadísticamente no significativas, causa de que se muestren muy correlacionadas entre sí. Además, las autocorrelaciones del primer mes tienen correlación positiva con las de diez a doce meses de diferencia puesto que corresponden a valores estadísticamente significativos y positivos. Se generaron cinco grupos por el algoritmo de agrupamiento, los cuales no guardan relación estadísticamente significativa con la clasificación propuesta por la Organización Mundial de la Salud para la versión 10. Pero en cuanto a su descripción, resalta la existencia de grupos estacionarios, pronosticables con frecuencias anuales o semestrales, o grupos con series de tiempo no pronosticables.

Firmas de los asesores:

Dr. José Arturo Berrones Santos	Dra. Satu Elisa Schaeffer
Co-Asesor	Co-Asesora

Abstract

José Alberto Benavides Vázquez.

Candidate for obtaining the degree of Master in Engineering with Specialization in Systems Engineering.

Universidad Autónoma de Nuevo León.

Facultad de Ingeniería Mecánica y Eléctrica.

Title of the study: Unsupervised clustering of time series deseases in Mexico between 2005 and 2015.

Number of pages: 65.

Objectives and methods: The objective consists in grouping epidemiological weekly records from Mexico between 2005 and 2015 time series with the purpose of describing them in terms of their characteristics, as well as to analyze if there is a statistically significant relationship strong enough to associate said time series to the classification assigned according to the World Health Organization. This can be achieved, first, through the extraction, cleaning and debugging of the data held in those records; secondly, transforming said data in weekly time series by sickness to a national level, normalized to the total of health beneficiaries interpolated between the obtained values from the INEGI from the years of 2005 and 2015; next, the mis-

sing records from said time series are interpolated; later, time series are chosen with at least five years of complete records to standardize the minimum information the time series offer; then, the slope, its intercept, and autocorrelations are extracted from each time series; afterwards, said characteristics are analyzed by a correlation matrix and with the variance threshold method, the ones with a higher variance than the median of the data are selected; now, they are separated in training and testing datasets determined sizes by the combination of sizes with less bias and variance; with the selected data separated in those datasets, the number of centroids is determined by the elbow method; lastly, the resulting groups were described by their characteristics and it was proven by hypothesis test if the resulting groups maintain some relation with the proposed clasification by the World Health Organization.

Results: We achieve to extract and clean data from PDF files published by the Secretaría de Salud of Mexico during 2005 and 2015. Said data is considered relevant information for the study of epidemiological focus at local and nationwide scope. From these dataset we obtain time series from the cases reported for each disease normalized by the number of health insured people registered at nationwide scope between 2005 and 2015. Those time series were grouped so they can be characterized by their temporal properties. The 60 % of the registries are relatives to infectious parasitary diseases, while the majority of medical consultations were due to respiratory diseases, followed by accidents and poissonings, whilst the diseases originated in the perinatal period were minimum. Three diseases with positive trends were identified within the studied decade:: The asymptomatic HIV infection, the whooping cough and the cholera. The cholera appears in this list because in 2013 there was an outbreak in Hidalgo, México after hurricanes Ingrid and Manuel striked the region. From the autocorrelation matrix analysis we can say that there is a strong correlation between the number of consultations in the same month,

whilst there is an inverse correlation between consultations with six to eight months of lag from, wich suggests that the time series are seasonal and maybe stationary. The autocorrelations with lags between four and eigth monts tend to be statistically not significant, which explain why they are correlated with each other. The autocorrelations from the first month usually have a positive correlations with the autocorrelations of lags of ten to twelve months. Five groups were generated by the clustering algorithm. Those groups do not have a statistically significant relationship with the 10th version of the classification proposed by the World Health Organization. Nevertheless, the groups obtained showed patterns characteristic of seasonal, stationary non-seasonal and non-stationary time series.

Signatures of supervisors:

Dr. José Arturo Berrones Santos	Dra. Satu Elisa Schaeffer
Co-Supervisor	Co-Supervisor

Índice general

Agradecimientos v

Resumen vi

Abstract ix

1. Introducción 1

1.1. Hipótesis 2

1.2. Objetivo 2

1.3. Estructura de la tesis 3

2. Marco teórico 4

3. Revisión bibliográfica 9

3.1. Agrupamiento por k-medias 11

3.2. Agrupamiento temporal dinámico 13

3.3. Modelos ARIMA 14

3.4. Otras implementaciones 15

3.5. Área de oportunidad 16

4. Metodología **18**

4.1. Recursos 18

4.2. Origen de los datos 19

4.3. Obtención de los datos 23

4.4. Preprocesamiento 26

4.5. Caracterización y selección de características 29

4.6. Procedimiento k-medias 30

5. Resultados **32**

6. Conclusiones **45**

6.1. Contribuciones 45

6.2. Trabajo a futuro 47

Bibliografía **48**

A. CIEs y sus nombres de enfermedades **57**

Índice de figuras

4.1. Cuadro de la página 13 del reporte correspondiente a la semana epidemiológica 6 de 2012. 21

4.2. Tres ejemplos de series de tiempo de los casos registrados normalizados por año, con marcas en rojo donde fueron interpolados los valores de los casos faltantes. 28

5.1. La figura contiene el conteo de CIEs generales de los registros seleccionados. 32

5.2. Logaritmo de casos normalizados por derechohabiencia. 33

5.3. Series de tiempo (en azul) con su pendiente (en rojo) y la serie de tiempo menos la tendencia (negro). 34

5.4. Autocorrelaciones de las enfermedades cuyas tendencias crecen significativamente. 35

5.5. Matriz de correlación entre características de las series de tiempo estudiadas. Sobresalen las fuertes correlaciones entre las autocorrelaciones de hasta dos semanas, las de las primeras seis semanas entre sí, las de los últimos dos meses y, por otro lado, las de retrasos semestrales por tratarse de correlaciones negativas con las autocorrelaciones de las primeras seis semanas y las últimas ocho semanas del año. 37

5.6. Características dadas sus varianzas y el umbral en 0.06 representado por una recta horizontal. 38

5.7. Errores con respecto al conjunto de prueba en diagramas de caja y bigotes para cada k elegida en el conjunto de entrenamiento. 40

5.8. PCA de dos componentes principales de las enfermedades estudiadas (círculos) coloreadas con base al grupo generado por k-medias al que pertenecen y, dentro de cada círculo, la letra impresa de la CIE general que se les asigna. 41

5.9. Autocorrelaciones de los grupos de enfermedades generados por k-medias. 44

Índice de tablas

4.1. Ejemplo de los datos extraídos desde los boletines epidemiológicos de la Secretaría de Salud de México a nivel estatal. 25

4.2. Muestra de los datos extraídos desde los boletines epidemiológicos de la Secretaría de Salud de México a nivel nacional. 25

4.3. CIEs generales y su descripción a partir de la CIE rev. 10 [68]. 26

5.1. Cifras de los conjuntos de entrenamiento y desarrollo. 39

A.1. CIEs y el nombre de la enfermedad correspondiente presentes en la población de 23 721 registros tomados de los datos obtenidos a nivel nacional. 57

Capítulo 1

Introducción

Los algoritmos de agrupamiento son una herramienta rápida y de bajo costo computacional que permiten describir y conocer las relaciones entre conjuntos de datos. Por ello, su uso ha sido muy extendido a lo largo de los últimos cuarenta años, periodo en el cual se han utilizado en una gran diversidad de tipos de datos pertenecientes a ámbitos biológicos, financieros, visuales, médicos, y entre otros también figuran los agrupamientos realizados a series de tiempo cuya finalidad es comprender bajo una metodología rigurosa el comportamiento de estas series. La relevancia de estos algoritmos radica en que los agrupamientos ofrecen, además de este carácter descriptivo, una herramienta sólida sobre la que probar hipótesis, cuyos resultados pueden ser utilizados para mejorar la precisión de otros algoritmos de clasificación, por ejemplo.

Por otro lado, las series de tiempo que se analizan en esta investigación provienen de boletines epidemiológicos que por la manera en que se distribuyen (formato `PDF`) presentan dificultades para extraer la información que contienen. Se extraen los datos de dichos boletines digitales por medio de herramientas computacionales especializadas en recuperar y preparar este tipo de información para ser utilizada por la ciencia de datos, rama de la ciencia a la que pertenecen los algoritmos de

agrupamiento antes mencionados. Ahora que se cuenta con estas series de tiempo epidemiológicas, su descripción y análisis se vuelve una tarea relevante y de interés para las ciencias de la salud y la ciencia de datos.

1.1 Hipótesis

La agrupación a partir de las características de las series de tiempo de los registros semanales de morbilidad en México publicados entre 2005 y 2015 ofrece información estadísticamente significativa que permite describir dichos registros epidemiológicos con base en sus propiedades temporales para futuras investigaciones de interés general.

1.2 Objetivo

Esta investigación se lleva a cabo para obtener datos epidemiológicos de interés respecto al reporte de enfermedades por parte de derechohabientes de la república mexicana durante el periodo de 2005 a 2015. Esto, a su vez, permite proponer una metodología de extracción y limpieza de datos que, por su presentación, son considerados de difícil manipulación. De igual manera, se busca establecer un procedimiento de preprocesamiento de datos epidemiológicos cuya frecuencia sea dada por el concepto de *semana epidemiológica*. Además, se quiere encontrar un conjunto de características temporales de los datos que los representen para, ulteriormente, ofrecer una descripción de los datos de estudio a partir de su agrupamiento con base en las características determinadas.

1.3 Estructura de la tesis

En el capítulo 2 se revisan los conceptos fundamentales que cimentan la investigación de este trabajo que incluyen la definición de CIE, metodologías de agrupamiento y series de tiempo con sus características. En el capítulo 3 se realiza un recorrido por los estudios más relevantes acerca de agrupamiento de series de tiempo y sus usos posteriores y, específicamente, agrupamientos de series de tiempo por k-medias. En el capítulo 4 se describen los procesos llevados a cabo para satisfacer la prueba de la hipótesis propuesta. Después, en el capítulo 5 se muestran los datos extraídos de cada proceso descrito en la metodología, entre los que destacan las propiedades de las series de tiempo, los grupos generados por k-medias y su caracterización. Finalmente, en el capítulo 6 se concluye el trabajo realizado y se plantea el trabajo a futuro.

Capítulo 2

Marco teórico

En este capítulo se definen los conceptos teóricos y formulaciones matemáticas que sustentan las metodologías y experimentos computacionales realizados en esta investigación.

Primeramente, resulta indispensable mencionar que a partir de reportes semanales de epidemiología[1] [2] publicados en PDF por la Secretaría de Salud de México durante 2005 y 2015, se desea extraer datos limpios de interés para las instancias gubernamentales, académicas, científicas y médicas interesadas.

Este organismo cuenta con un instrumento estadístico y sanitario para identificar enfermedades llamado Clasificación Internacional de Enfermedades (CIE)[2] [69], cuya finalidad es entender las causas de morbilidad y mortalidad de la población y así mejorar la calidad de vida de la misma [45]. Con base en un criterio epidemiológico y sanitario establecido por Farr a finales del siglo XIX [46], esta clasificación agrupa enfermedades en epidémicas, generales, locales ordenadas por origen (geográfico), trastornos del desarrollo y lesiones [46]. Para distinguirlas se utiliza un código alfa-

[1] Una semana epidemiológica es un estándar de medición temporal que se utiliza, principalmente en ámbitos médicos, para comparar datos en ventanas de tiempo definidas. La primera semana epidemiológica del año termina el primer sábado de enero de cada año.

[2] Actualmente en la versión 11; sin embargo, puesto que los datos estudiados corresponden a la versión 10, se utiliza el manual de ésta para definir los criterios de clasificación.

numérico consistente en una letra en la primera posición, seguida de dos dígitos, un punto decimal y un último dígito. El rango de valores va de A00.0 a Z99.9, reservando la U para causas de morbilidad o mortalidad cuya clasificación aún se desconoce [46].

Por otro lado, existen otras metodologías pertenecientes a la estadística para procesar grandes cantidades de datos [27]: las *descriptivas* y las *inferenciales*. En las últimas se hacen inferencias sobre la población utilizando una muestra de la población, pero también se quiere determinar las características de los mismos a través de las metodologías descriptivas.

Dentro de las metodologías inferenciales se encuentra el reconocimiento de patrones, usualmente diferenciado en *supervisado* y *no supervisado*. Las metodologías supervisadas cuentan con una característica a partir de la cual se pueden clasificar los datos, por ejemplo la especie a la que pertenecerían conjuntos de flores o las marcas de vehículos. Por su parte, los métodos no supervisados carecen de este tipo de información [27].

Como parte de las metodologías no supervisadas se encuentra el *agrupamiento de datos*, cuyo objetivo es buscar estructuras en conjuntos de datos a través de sus características [27], de modo que se parte de n objetos y se tratan de asociar en k grupos a partir de la similitud de una determinada medida de sus características. El agrupamiento de datos se utiliza principalmente con tres finalidades, a saber: encontrar estructuras subyacentes de datos, agrupar conforme a un orden natural, y reducir la cantidad de datos con los que se trabaja.

Estas aproximaciones usan la idea de *grupo* entendida en este contexto como una colección de puntos cuyas distancias entre sí son menores con respecto a las distancias entre los puntos de las otras colecciones [6]. El algoritmo no supervisado

más utilizado para agrupar datos es llamado k-medias [6, 27], mismo que parte de $X = \{x_i\}, i = 1, ..., n$ puntos d-dimensionales a tomar en cuenta por el algoritmo. De manera general, las características se pretenden agrupar en k grupos con $C = \{c_j, j = 1, ..., k\}$ centros. Esto se logra al minimizar la distancia cuadrada entre la media μ_j de los puntos $x_i \in c_j$ asociados a cada grupo. Para cada centro esto es

$$J(c_j) = \sum_{x_i \in c_j} \|x_i - \mu_j\|^2, \tag{2.1}$$

de modo que para todos los grupos se tiene

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_j} \|x_i - \mu_k\|^2. \tag{2.2}$$

La función objetivo de este algoritmo consiste en minimizar $J(C)$, o sea

$$\text{mín}(J(C)). \tag{2.3}$$

Esta función objetivo representa un problema NP-duro [27] que se resuelve al

1. seleccionar un número k de grupos;
2. asignarles una posición C_j inicial aleatoria;
3. asociar cada punto con el centro C_j más cercano;
4. encontrar la media de cada grupo μ_j;
5. mover cada centro C_j a dicha media μ_j;
6. medir $J(C)$ y si es menor que el anterior, repetir desde el paso 4 [6, 27].

En este estudio, los datos se obtienen de *series de tiempo*, entendidas como un conjunto de observaciones $\{o_t\}$ tomadas en un tiempo t determinado [7], en las que

cada observación o_{et} expresa los casos registrados de cada enfermedad e a lo largo de todas las semanas epidemiológicas t reportadas durante el periodo especificado. En general [7], para estudiar series de tiempo se obtiene la tendencia, los componentes estacionales, la autocorrelación de sus residuales estacionarios y sus componentes de Fourier.

La tendencia w_0 de una serie de tiempo se puede obtener a partir de una *regresión lineal* de la misma. Una regresión lineal [11] es una metodología inferencial supervisada que busca predecir valores y dado un vector de variables de entrada t por medio del ajuste de coeficientes ω de la función lineal

$$\hat{y}(t, \omega) = \omega_0 + \omega_1 x_1 + \ldots + \omega_t x_t. \tag{2.4}$$

Estos coeficientes ω son los que minimizan el error cuadrado entre los valores de y y sus estimados $\hat{y}$, esto es

$$\text{mín} \left(\sum_t (y_t - \hat{y}_t)^2 \right). \tag{2.5}$$

A su vez, la autocorrelación $\hat{p}$ es usada para conocer el grado de dependencia de las observaciones de una serie de tiempo y el modelo al que se ajustan. El concepto de autocorrelación $\hat{p}$ con retraso h parte de la autocovarianza

$$\gamma_o(h) = \text{CoV}(X_{t+h}, X_t) \tag{2.6}$$

para definirse como

$$\hat{p}_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(X_{t+h}, X_t). \tag{2.7}$$

La medición de esta función permite conocer si una serie de tiempo es aleatoria

y, en dado caso, impredecible e imposible de modelar [7]. Esto sucede cuando la varianza entre las observaciones es constante y sus valores de correlación con el resto de las observaciones son de cero. Cuando una serie de tiempo cumple con estas condiciones se llama *ruido blanco* [7].

Capítulo 3

Revisión bibliográfica

Este capítulo incluye una revisión bibliográfica de 1979 a 2017 sobre temas relacionados con el agrupamiento de series de tiempo, los modelos utilizados, las características relevantes y los preprocesamientos necesarios para lograr mejores resultados. En general, predomina la preferencia de agrupar series de tiempo por k-medias y agrupamiento temporal dinámico aunque es frecuente el uso de modelos ARMA y ARIMA e incluso figuran trabajos que agrupan por modelos jerárquicos, correlacionales y por redes complejas. En cuanto a las características empleadas para agrupar predominan las autocorrelaciones para k-medias y las series de tiempo para el agrupamiento temporal dinámico. Aún así, figuran características como las ondículas de Haar y los coeficientes de Fourier.

Estos algoritmos se utilizan en muy variadas ramas de las ciencias, economía y humanidades. Por ejemplo, en 2004 Focardi y Fabozzi [18] utilizaron agrupamientos de series de tiempo como herramienta para elegir portafolios financieros a partir de distintos óptimos arrojados por los algoritmos y métricas que estudiaron. Mientras que en 2011, Li y Prakash [36] agruparon series de tiempo de capturas de video puesto que, por su naturaleza, son difíciles de clasificar manualmente. Su aproximación implementa un sistema lineal dinámico de variables complejas y matrices de transición

para luego usar un algoritmo de esperanza-maximización con el que agrupar dichas series de tiempo. También figuran estudios que asocian autoría a textos con base en implementación de algoritmos de k-medias como el efectuado por Layton et al. [35] en 2010 en donde se logran atribuir textos menores o iguales de 140 caracteres provenientes de publicaciones de Twitter a los usuarios que las escribieron.

Entre las dificultades principales de trabajar con series de tiempo se encuentra el tamaño de las mismas tanto en la cantidad de datos registrados como en el número de características asociadas a cada dato. Por lo mismo, existen maneras de reducir características. Entre ellas, en 2005, figura la de Bagnall y Janacek [4] que consiste en tratar series de tiempo con información recortada provenientes del modelo AR-MA por k-medias y llegaron a la conclusión de que reducir dimensiones y utilizar autocorrelaciones mejora la certeza en el agrupamiento para estas series de tiempo. Un año después apareció otra de estas aportaciones de parte de Zhang et al. [74] quienes utilizaron las transformaciones ortogonales de las ondículas de Haar como característica para reducir las dimensiones de series de tiempo.

En la misma línea del párrafo anterior resaltan estas aproximaciones. Primero la explorada por Wang et al. [66] en 2006. Ellos utilizaron como características la tendencia, estacionalidad, frecuencia, correlaciones, oblicuidad, kurtosis y la no linealidad. A partir de esta aproximación se reducen las dimensiones de grandes series de tiempo y se logra un mejor agrupamiento de los datos de las series de tiempo como tales. Un año después, este estudio se expandió para series de tiempo multivariadas por Wang et al. [67]. En este respecto y más recientemente, en 2014, Fulcher y Jones [21] estudiaron la reducción de características y agrupamiento de series de tiempo con base en un clasificador lineal que toma como entrada características de una serie de tiempo tales como la autocorrelación, distribución y tendencia. El clasificador elige características y separa las series de tiempo con base en las similitudes y diferencias

de dichas características.

Una rápida revisión de este tema se puede consultar en dos publicaciones realizadas en 2010. Una trata de algoritmos de agrupamiento por k-medias Jain [27] y otra específicamente de agrupamiento de series de tiempo Kavitha y Punithavalli [29].

3.1 Agrupamiento por k-medias

El algoritmo de agrupamiento por k-medias es el históricamente más utilizado para agrupar todo tipo de datos, entre los que figuran las series de tiempo. Fue propuesto en 1979 por Hartigan y Wong [22] como un proceso para agrupar x puntos d-dimensionales en k grupos previamente definidos a partir de la minimización de la suma de errores cuadrados de dichos puntos. Desde entonces, se ha utilizado extensamente, como lo denota la bibliografía relacionada, y por la rapidez y claridad de sus resultados es considerado un buen algoritmo de exploración inicial de los datos.

Un ejemplo de su uso se documenta en 2002 cuando Singhal y Seborg [59] modificaron el algoritmo de k-medias para agrupar series de tiempo a partir de los factores de similitud obtenidos del *análisis de componentes principales* (PCA por sus siglas en inglés) y de su distancia de Mahalanobis. Al respecto, cabe señalar que las distancias de k-medias, pese a que por definición son euclidianas, pueden modificarse para ajustarse a los datos con que se trabaja.

Otra de las características que se utiliza para agrupar series de tiempo por este algoritmo es la elegida por Vlachos et al. [64] en 2003 quienes publicaron un artículo en el que utilizan *ondículas* (conocidas como *wavelets* por su traducción en inglés)

Un año después, Lin et al. [37] dieron una conferencia en la que demostraron que el uso de las descomposiciones de ondículas de Haar en lugar de las series de tiempo en sí mejoran la precisión y tiempos de cómputo de los algoritmos de agrupamiento de k-medias.

Con todo, hay estudios que concluyen en que agrupar series de tiempo da resultados no significativos. El más llamativo apareció antes de terminar 2005 realizado por los autores Keogh y Lin [30], mismos que sostuvieron que el agrupamiento de subsecuencias de series de tiempo es irrelevante. Sin embargo, estos mismos investigadores adjuntan en su artículo un método para agrupar algunas series de tiempo que consiste en distinguir sus motivos, entendidos como subsecuencias recurrentes distanciadas de manera no trivial de otras ocurrencias, y utilizarlos como subsecuencias susceptibles de ser agrupadas por algoritmos como k-medias, resolviendo así el procedimiento que ellos mismos describieron como irrelevante.

El artículo de Keogh y Lin [30] tuvo varias respuestas, de entre las que sobresale la dada por el investigador Chen [9] quien demostró que es posible obtener agrupamientos significativos de subsecuencias de series de tiempo utilizando retrasos de las mismas con el fin de encontrar patrones similares que, posteriormente, se utilizan como entrada para el algoritmo de k-medias.

Las mejoras al algoritmo de k-medias incluyen también ajustes para agrupar series de tiempo de pequeña duración, problema abordado por Ernst et al. [16] en el mismo año de 2005. Este grupo de científicos resolvieron este problema combinando el algoritmo de k-medias y los coeficientes de correlación entre las mismas series de tiempo. En esta misma línea de investigaciones orientadas a la mejora del algoritmo, se cuentan las aportaciones para mejorar tiempos de ejecución o precisión en el algoritmo, siendo destacable la realizada por Ratanamahatana et al. [55] en 2005 que consiste en convertir las series de tiempo en secuencias binarias: los valores

mayores a la media se convierten en unos y el resto en ceros. También en 2009, Lai et al. [34] utilizaron k-medias para generar grupos de series de tiempo de datos financieros a partir de selección de características por prueba F, grupos que luego fueron usados para predecir índices de mercado a partir de árboles de decisión difusos junto a algoritmos genéticos.

Otra aproximación al agrupamiento por k-medias de series de tiempo aparece en 2015 cuando Paparrizos y Gravano [47] propusieron un modelo que denominaron k-Shape para agrupar series de tiempo comparándolas mediante una normalización de la covarianza entre dichas series, lo que permite mantener su forma y características a cambio de una mayor exigencia computacional. Dos años después, en 2017, Paparrizos y Gravano [48] desarrollaron dos técnicas de agrupamiento que parten de una medida normalizada de correlaciones entre series de tiempo. La denominada k-Shape produce un centroide por grupo, mientras que k-*MultiShapes* produce varios centroides relacionados con su proximidad y distribución espacial.

3.2 Agrupamiento temporal dinámico

El uso del algoritmo de agrupamiento temporal dinámico, abreviado DTW por sus siglas en inglés, empezó a extenderse a finales del siglo XX, específicamente en 1998, cuando Keogh y Pazzani [31] propusieron una representación segmentada de series de tiempo que promovió una mayor precisión para clasificarlas y agruparlas, además de que permitía visualizar rápidamente valores relevantes. A partir de entonces se ha mejorado la precisión y representación de este algoritmo mediante esfuerzos tales como el de Oates [42] por este médoto y, un año después, Oates et al. [43] utilizaron este mismo método para determinar el número de *modelos ocultos de Markov* (HMMs, dada su abreviatura en inglés) en una serie de tiempo lo cual

permite eliminar secuencias en las series que no pertenecen a las mismas.

Nuevamente fueron Keogh y Pazzani [32] quienes, un año después, propusieron una mejora al método de alineamiento temporal dinámico que consiste en comprimir una serie de tiempo a partir de obtener la media de segmentos del mismo tamaño mejorando el tiempo de cómputo y la certeza del agrupamiento.

Más adelante, en 2011, Zhang et al. [75] realizaron agrupamientos de series de tiempo a partir de características obtenidas por el método de vecinos cercanos a partir de la métrica de similitud coseno entre series de tiempo y, posteriormente, agrupando por alineamiento temporal dinámico y agrupamiento jerárquico. Pocos años después, Izakian et al. [26] propusieron una métrica difusa para el alineamiento temporal dinámico con la que determinar los grupos de series de tiempo.

3.3 Modelos ARIMA

Los *modelos autorregresivos integrados de media móvil* o ARIMA, por su abreviatura en inglés, también comprenden un vasto catálogo de aproximaciones para estudiar series de tiempo y, aunque principalmente se utilizan para pronosticarlas, también se documentan usos para su agrupamiento, como el de Kalpalis et al. [28] en 2001, investigadores que midieron la similitud entre distintas series de tiempo pertenecientes a ARIMA utilizando las distancias euclidianas entre los coeficientes cepstrales de sus codificaciones predictivas lineales, a saber, la inversa de la transformada de Fourier de la amplitud logarítmica más baja del espectro. En su investigación demostraron que el uso de estas distancias permite un mejor agrupamiento sin necesidad de que las series de tiempo sean del mismo tamaño. En ese mismo año, Xiong y Yeung [70] utilizaron el algoritmo de esperanza-maximización para conocer los valores faltantes de series de tiempo y posteriormente pasarlas por un algorit-

mo de agrupamiento basado en el *modelo autorregresivo de media móvil* (abreviado ARMA en inglés).

En 2004, Xiong y Yeung [71] agruparon de series de tiempo de distintos tamaños obteniendo el número de grupos iniciales mediante el criterio de información bayesiana y determinando los grupos por el algoritmo de esperanza-maximización a partir de mezclas de modelos ARMA. Cuatro años más tarde, Corduas y Piccolo [10] trabajaron con series de tiempo desde el paradigma de las distancias autorregresivas de sus modelos ARIMA tanto para agrupar como para clasificar dichas series de tiempo. Aparte, Hautamäki et al. [23] propusieron un método de agrupamiento consistente en minimizar distancias mediante alineamiento temporal dinámico y optimizar este proceso por una heurística de búsqueda local.

3.4 Otras implementaciones

Finalmente se comparten estudios que utilizan otros algoritmos características e implementaciones. En primer lugar figura el trabajo realizado en 2003 por Möller-Levet et al. [38] quienes propusieron un algoritmo para agrupamiento difuso para series de tiempo de corta duración y cuyos datos no están equitativamente distribuidos a lo largo del tiempo.

Luego, Rodrigues et al. [56] exploraron el agrupamiento jerárquico de series de tiempo por árboles binarios con la finalidad de encontrar conjuntos de variables altamente correlacionados. Además, por su parte, Frühwirth-Schnatter y Kaufmann [20] agruparon múltiples series de tiempo a partir de parámetros extraídos de cadenas bayesianas de Markov aplicadas a simulaciones por el método de Montecarlo. Por su parte, D'Urso y Maharaj [15] utilizaron, en 2009, un sistema basado en autocorrelaciones difusas de series de tiempo que podrían ser de tamaños distintos para

agruparlas con base en los cambios de comportamiento que registran a lo largo del tiempo.

Más adelante, Rakthanmanon et al. [53] propusieron una nueva aproximación para agrupar subsecuencias de series de tiempo basada en el principio bayesiano de descripción mínima, descartando el agrupamiento de la serie en sí por considerarla destinada al fracaso. Un año después, los mismos autores [54] ampliaron su estudio con la implementación de un algoritmo MDL cuya mejora es que requiere una mínima cantidad de características para poder agrupar series de tiempo. Una aportación en este mismo sentido fue realizada por Zakaria et al. [73]. Estos investigadores propusieron un método de agrupamiento que toma en cuenta patrones locales de las series de tiempo (*shapelets*) para, a partir de ellos, medir distancias respecto a otras series de tiempo de la misma longitud o diferentes.

Por último, Ferreira y Zhao [17] convirtieron las series de tiempo en vértices de una red compleja para luego agruparlas conforme a algoritmos de detección de comunidades que generan aristas entre vértices cercanos constituyendo los vértices conectados los grupos generados.

3.5 Área de oportunidad

Esta revisión de la literatura relacionada permite presentar el cuadro **??** (p. **??**) que sintetiza, por metodología y datos de entrada, los trabajos relacionados al agrupamiento de series de tiempo entre 1979 y 2017. En dicho cuadro se constata que k-medias es el algoritmo más utilizado para agrupar series de tiempo, en tanto que las características utilizadas principalmente son las autocorrelaciones (abreviadas ACF en el cuadro), seguidas por las ondículas de Haar. El motivo por el que los autores de estas investigaciones se decantan por el algoritmo de k-medias es debido

a que ofrece muy rápidamente grupos con características relevantes que se ajustan a sus objetivos de investigación. Igualmente, el uso de las autocorrelaciones como datos de entrada de los algoritmos de agrupamiento ha permitido a los investigadores que las han utilizado obtener buenos resultados y grupos de series de tiempo con características temporales similares. Por estos motivos, en esta investigación se opta por utilizar el algoritmo de k-medias y las autocorrelaciones de las series de tiempo para agrupar dichas series y describir su comportamiento temporal.

En otro respecto, el estudio del estado del arte revela que una de las primeras aproximaciones para la descripción de datos y su preprocesamiento para investigaciones posteriores, consiste analizarlos mediante algoritmos de agrupamiento. Específicamente en el caso del agrupamiento de series de tiempo con base en sus características, se logran encontrar relaciones entre las series de tiempo estudiadas, mismas que han sido utilizadas para corroborar preconcepciones sobre series de tiempo de interés o para describir nuevos patrones y relaciones insospechadas entre las mismas. Con base en el sustento anterior, se propone como objeto de estudio el análisis mediante algoritmos de agrupamiento de estos datos cuyo análisis o descripción, en el caso georreferenciado de México, no se encuentra registrado en la literatura existente.

Capítulo 4

Metodología

En este capítulo se detalla, en primer lugar, cómo se han extraído y preparado los datos para su manipulación computacional. Luego, la manera en que se manipulan las características de dichos datos para convertirlos en series de tiempo. Posteriormente, se extraen características a partir de propiedades temporales de las series de tiempo. Se reduce el número de estas características. A continuación, se agrupan los registros con base en sus características. Por último, se describen los grupos y se mide el impacto que tienen con respecto a las clasificaciones existentes de las enfermedades a las que se asocian.

4.1 Recursos

En esta investigación se utiliza una computadora portátil Asus X556U con sistema operativo Windows 10 Home Single Language de 64 bits, procesador Intel Core i7-7500U a 2.70 GHz, con 8 GB de memoria RAM y un disco de estado sólido Kingston SA40037480G.

Para extraer los datos de los archivos `PDF` se usa la herramienta `tabula-py`

[3] y PyPDF2 [50] que se ejecutan en lenguaje Python [51]. Con la primera se extrae el contenido de archivos PDF mediante especificación de algunos parámetros, y la segunda permite leer archivos PDF y, entre otras funciones, extraer el número de páginas.

4.2 Origen de los datos

La Secretaría de Salud de México publica boletines epidemiológicos en los que se detalla semanalmente el número de casos registrados por enfermedad en cada estado de la república mexicana. Estos boletines pueden descargarse desde la página de la Secretaría de Salud [58]. Estos archivos se descargaron por medio de un programa que accede de manera iterativa a las direcciones URL de los archivos. En general, las direcciones de los archivos pudieron predecirse a partir de la observación de patrones en algunas direcciones por año.

Por ejemplo, el boletín de la semana epidemiológica 25 de 2008 está disponible en la dirección http://www.epidemiologia.salud.gob.mx/doctos/boletin/2008_sem25.pdf y el del boletín 51 del mismo año en http://www.epidemiologia.salud.gob.mx/doctos/boletin/2008_sem51.pdf, de donde se puede observar que sólo cambian los dígitos que siguen a sem y preceden a .pdf. Esto no sucede en el año 2013, cuyas direcciones no parecen seguir ningún patrón rápidamente predecible, por lo que se opta por descargar manualmente los boletines de ese año. Los boletines se encuentran en formato PDF [1], salvo los del año 2011, cuyas páginas por separado se hallan en formato PDF comprimidas en formato ZIP. Con el fin de tener los boletines en el mismo tipo de archivo y formato, las páginas de los boletines de 2011 se extraen y agrupan en un solo archivo PDF por semana epidemiológica mediante el uso de la herramienta PDF Mergy [65].

Cada boletín corresponde a una semana epidemiológica del año, de modo que generalmente se tienen 52 boletines por año, excepto en 2008 y 2014 en que se cuenta con 53. La información semanal de casos registrados para cada enfermedad y estado de la República se muestra por página en forma de cuadro. Antes de la aparición de los cuadros de enfermedades, se tienen algunas páginas que pueden contener la portada del boletín, una presentación del mismo, y uno o más reportes científicos relacionados con casos de morbilidad nacional o temas afines al sector salud nacional. Tras esta información, en la mayoría de los boletines hay un cuadro con un resumen de los nuevos casos registrados durante la semana epidemiológica del boletín. En la página siguiente de este boletín, aparecen los cuadros con los casos de enfermedades que nos interesan en este trabajo de investigación y cuya estructura es similar a la de la figura 4.1. Después de éstos, se suelen aparecer cuadros con información social diversa. La página final muestra información relacionada con la edición e impresión del boletín.

En los cuadros de casos registrados, las filas contienen los estados de la república mexicana con una fila al final que representa el total, y en las columnas se tienen las enfermedades. Cada página que contiene estos cuadros incluye todos los estados de la República y de una a cuatro columnas de enfermedad. Cada columna puede estar subdividida en uno, dos o tres años; cada año contiene una subdivisión que puede incluir la cantidad de casos registrados en la semana epidemiológica del boletín, el acumulado del año a la fecha de la semana epidemiológica del boletín o la cantidad de hombres y mujeres que reportaron la afección en la semana epidemiológica o a lo largo del año, esto último cuando se añade un encabezado para indicarlo.

En cuanto a la información de las celdas, los estados y enfermedades se escriben en mayúsculas con una tipografía que agranda la primera letra de cada nombre; las celdas de las enfermedades incluyen el nombre, la edición de revisión de la **CIE**,

Vigilancia Epidemiológica Semana 6, 2012 13

CUADRO 5.1 *Casos por entidad federativa de* ***Enfermedades Infecciosas*** *del Aparato Respiratorio hasta la semana epidemiológica 5; Influenza hasta la 6 del 2012*

ENTIDAD FEDERATIVA	Neumonías y Bronconeumonías CIE-10ª REV. J12-J18 excepto J18.2				Influenza (A H1N1) CIE-10ª REV. J09			Influenza Estacional CIE-10ª REV. J10-J11		
	2012			2011	2012			2012		2011
	Sem.	Acum.		Acum.	Sem.	Acum.		Acum.		Acum.
		M	F			M	F	M	F	
Aguascalientes	79	170	233	677	17	27	19	1	1	5
Baja California	161	445	434	909	18	17	13	6	11	3
Baja California Sur	36	98	91	176	30	74	67	8	8	-
Campeche	20	50	48	86	8	11	16	-	3	-
Coahuila	65	232	217	692	7	10	5	-	-	-
Colima	48	100	111	207	30	50	69	1	1	6
Chiapas	73	160	179	343	21	60	78	9	7	1
Chihuahua	246	654	695	1 632	19	10	18	-	-	6
Distrito Federal	180	806	779	2 154	34	226	282	13	21	65
Durango	69	205	227	625	21	16	14	1	-	-
Guanajuato	184	414	425	1 105	19	22	27	1	1	-
Guerrero	98	210	194	417	9	20	21	2	1	3
Hidalgo	91	198	200	610	12	127	156	7	6	8
Jalisco	377	941	1 044	3 318	114	163	236	17	26	17
México	182	373	417	984	32	174	158	35	19	12
Michoacán	90	173	258	415	41	71	97	8	6	3
Morelos	41	102	85	203	13	53	74	2	-	22
Nayarit	22	86	53	140	9	17	18	3	5	-
Nuevo León	431	1 115	1 171	2 100	90	133	143	11	13	52
Oaxaca	70	177	179	578	12	127	155	38	32	3
Puebla	134	375	336	778	20	97	91	2	1	5
Querétaro	48	137	133	400	79	200	205	6	7	18
Quintana Roo	41	107	76	235	15	21	19	1	2	-
San Luis Potosí	162	387	380	726	37	52	60	9	11	9
Sinaloa	76	215	203	435	11	15	35	5	9	1
Sonora	324	987	967	1 707	74	67	101	3	2	-
Tabasco	60	133	92	331	48	24	64	-	9	-
Tamaulipas	222	400	445	902	20	20	32	1	2	2
Tlaxcala	18	33	35	86	5	25	30	6	1	9
Veracruz	141	355	356	550	15	52	62	1	10	1
Yucatán	74	155	128	182	14	25	16	9	13	2
Zacatecas	89	213	179	427	14	9	12	2	5	-
TOTAL	3 952	10 206	10 370	24 130	908	2 015	2 393	208	233	253

FUENTE: SINAVE/DGE/SALUD 2012. Información preliminar.

Figura 4.1: Cuadro de la página 13 del reporte correspondiente a la semana epidemiológica 6 de 2012.

acrónimo de Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud [69], y la CIE asignada por la Secretaría de Salud de México. Los casos se registran con números arábigos, pero puede aparecer una siglas cuya nomenclatura se especifica en los boletines como sigue:

- -: No se presentaron casos en la semana epidemiológica.
- n. d.: Información no disponible.
- n. e.: Información no enviada por la entidad federativa.
- n. a.: No aplica.
- s. n.: Sin notificación por la entidad federativa.

Dicha nomenclatura suele variar en ausencia de espacios entre palabras y ausencia de uno o los dos puntos de manera difícil de predecir e incluso observar a simple vista. Asimismo, cuando los números de casos exceden los cientos, pueden encontrarse separaciones en grupos de tres dígitos por comas o espacios (uno o más), o carecer de ellas. Cabe destacar que las tipografías varían a lo largo de los años y se presentan boletines que utilizan comillas de manera irregular para enmarcar datos.

En relación al aspecto visual de los cuadros de casos reportados, suelen estar demarcados por líneas gruesas, sin embargo las separaciones entre columnas y filas no siguen un formato constante. Finalmente, se presentan casos con errores de impresión en donde el contenido de las celdas puede aparecer fuera del lugar que le corresponde o invadiendo celdas contiguas. En total, se tienen entre veinte y cuarenta páginas con datos de interés por cada boletín en formato `PDF`.

4.3 Obtención de los datos

Extraer información de cuadros en este tipo de archivos se considera complejo al punto de que esta tarea constituye un campo de estudios denominado *Table Extraction* [72]. Con el uso de las librerías descritas en el cuadro 4.1 se extrajeron los datos por año de modo que primero se procesaron todos los `PDF` semanales año por año. Por cada reporte se lee cada página y se busca aquélla que contenga las cadenas de texto `CUADRO 3` y `MENINGITIS` puesto que es la primera enfermedad contenida en todos los reportes. Una vez alcanzada dicha página, se leen todos los datos de la página con los rectángulos que los contienen. Dichos rectángulos están determinados por el pixel superior y a la izquierda, el ancho y el alto de pixeles de dicho rectángulo. A partir de esos datos, se buscan las posiciones del nombre de las enfermedades en la página `PDF`, a partir de allí, es busca la palabra `Sem`, el encabezado de la columna de interés. Con los pixeles de esos rectángulos se especifican los anchos de columna de interés y se extraen los casos registrados por estado y semana por cada página. En general, se siguió el procedimiento mostrado en el algoritmo 1 (p. 24).

En el algoritmo se elige como punto de partida para la lectura de datos la aparición de las cadenas de texto `CUADRO 3` y `MENINGITIS` puesto que todos los documentos `PDF` presentaban esta información en el primer cuadro de interés. Sin embargo y pese a esta extracción, los datos volcados en archivos de formato `CSV` contenían numerosos errores de lectura debidos a las diferencias de formato arriba señaladas, por ello se pasaron por otro archivo escrito en `Python` que limpia para cada registro los números de casos reportados, el estado de la república mexicana, los nombres de enfermedad y las CIEs asignadas. Tras limpiar cada archivo generado por el programa que extraía la información, se buscaron errores mediante expresiones regulares, técnica de búsqueda de caracteres o estructuras de caracteres definida por

```
para cada directorio en año hacer
    para cada archivo en directorio hacer
        si termina con .pdf entonces
            leer páginas con PyPDF2 [50];
            para cada página en archivo hacer
                si contiene cuadro de interés entonces
                    extraer contenido con tabulapy [3];
                    extraer posiciones del contenido en JSON;
                    seleccionar pixeles de columnas de interés;
                    para cada columna en página hacer
                        ajustar anchos de columna;
                        leer filas;
                    fin
                fin
            fin
        fin
    fin
    exportar datos en CSV;
fin
```

Algoritmo 1: Algoritmo de extracción de datos.

Thompson [63]. Esta búsqueda evidenció errores de formato en las CIEs del grupo T63, además de errores en el número de casos reportados en miles de registros, mismos que fueron corregidos manualmente.

Después de limpiados por este proceso, se usa `awk` [19] para visualizar el contenido de los archivos generados agrupados por algún dato de interés. Resaltaron los grupos por nombre de enfermedad y CIE ya que, a lo largo de los años, dichos nombres variaron para la misma enfermedad, como sucedía con el VIH que a veces aparecía como *Virus de Inmunodeficiencia Humana*. También se utilizó esta herramienta para ordenar de mayor a menor el número de casos registrados puesto que existen registros obtenidos con valores superiores a la mitad de la población mexicana. Estos registros se cotejaron directamente con los `PDF` correspondientes y se corrigieron de manera manual.

Como resultado de este proceso, se extrajeron 784 660 registros con 169 CIEs

Cuadro 4.1: Ejemplo de los datos extraídos desde los boletines epidemiológicos de la Secretaría de Salud de México a nivel estatal.

Año	SE	Estado	Enfermedad	Casos	CIE
2006	12	Querétaro	Shigelosis	0	A03
2010	8	Zacatecas	Conjuntivitis	165	B30
2014	29	Durango	Asma	85	J45

Cuadro 4.2: Muestra de los datos extraídos desde los boletines epidemiológicos de la Secretaría de Salud de México a nivel nacional.

Año	SE	Enfermedad	Casos	CIE
2013	37	Cólera	0	A01
2006	52	Mordeduras por otros mamíferos	117	W55
2014	20	Paludismo por P. Vivax	8	B51

distintas que incluyen el año reportado, la semana epidemiológica (SE) correspondiente, el estado de la república mexicana donde se informó de los incidentes, el número de casos registrados, el nombre de la enfermedad y la CIE asignada por la OMS. Un ejemplo de estos datos aparece en el cuadro 4.1 (p. 25).

De estos registros, se seleccionan los registros agrupados por tipo de enfermedad a nivel nacional con el objetivo de reducir el número de registros. Este agrupamiento consiste en 23 722 registros que contienen las mismas columnas que los originales, salvo por el estado de la república mexicana. Una muestra de eso se halla en el cuadro 4.2 (p. 25).

Cuadro 4.3: CIEs generales y su descripción a partir de la CIE rev. 10 [68].

CIE general	Descripción
A–B	Enfermedades infecciosas y parasitarias
C–D48	Neoplasmas
D50–D89	Enfermedades de la sangre
E	Enfermedades endocrinas, metabólicas y nutricionales
F	Desórdenes mentales y del comportamiento
G	Enfermedades del sistema nervioso
H00–H59	Enfermedades del ojo y anexas
H60–H95	Enfermedades del oído
I	Enfermedades del sistema circulatorio
J	Enfermedades del sistema respiratorio
K	Enfermedades del sistema digestivo
L	Enfermedades de la piel y tejidos subcutáneos
M	Enfermedades del sistema musculoesquelético y tejido conectivo
N	Enfermedades del sistema genitourinario
O	Embarazo y nacimiento
P	Enfermedades originadas en el periodo perinatal
Q	Malformaciones congénitas, deformaciones y anormalidades cromosómicas
R	Anormalidades no clasificadas
S–T	Heridas y envenenamientos
V–Y	Causas externas de morbilidad y mortalidad
Z	Factores que influyen en el estado de salud y el contacto con servicios de salud

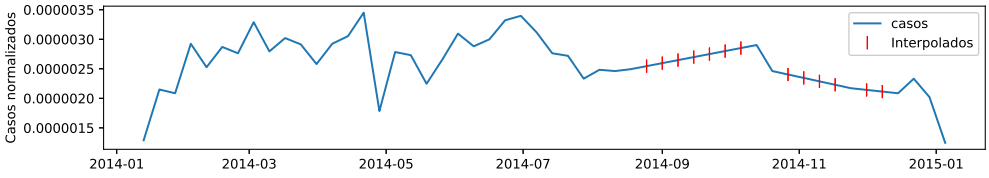
4.4 Preprocesamiento

Debido a que estos datos representan una extensa colección de información asociada al paso del tiempo, se ha decidido tratarlos en conjuntos como series de tiempo con el fin de procesarlos y reducir sus características para, ulteriormente, utilizarlos como datos de entrada en un algoritmo de k-medias. Por lo tanto, en primer lugar se separaron los datos por CIE. Se obtuvieron 143 CIEs distintas, asociadas a un nombre de enfermedad que pueden consultarse en el cuadro A.1 del apéndice A, mientras que las CIEs generales se hallan en el cuadro 4.3 (p. 26).

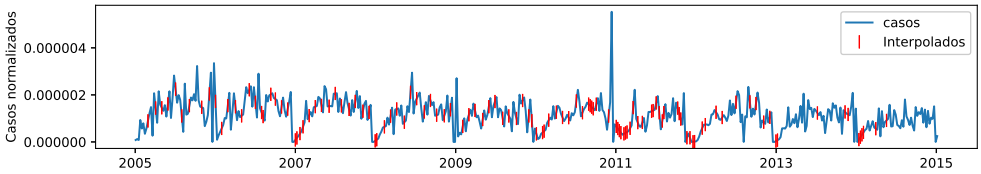
Ahora se normalizan los casos registrados entre el número de derechohabientes

del año correspondiente [24] con el fin de tener una medición comparable de enfermedades. Como sólo se cuenta con la cantidad de derechohabientes a nivel nacional de los años 2010 y 2015, se interpolan y extrapolan para cada año en el periodo comprendido en la investigación con la herramienta `linregress` de la librería `SciPy` [61]. Cabe señalar que previa a esta normalización se intentó utilizar el número de habitantes en México [25] para ajustar el parámetro de los casos registrados, mas los resultados obtenidos indicaban que algunas enfermedades tendían a la alza, como la fiebre tifoidea, tendencia que desaparece al normalizar contra derechohabientes por año.

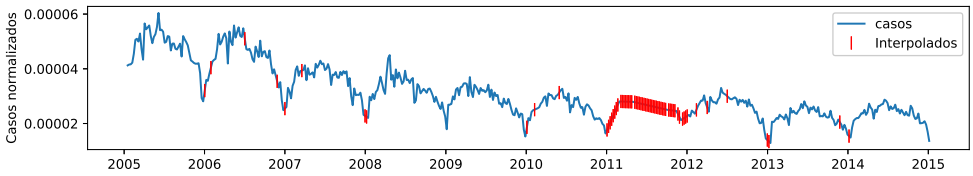
Para trabajar con estas series de tiempo, es necesario convertir el valor de su año y semana epidemiológica en una fecha con formato compatible con Python, el lenguaje que se utiliza para realizar esta investigación. Para dicho fin, se procesaron los datos temporales por las librerías `pandas` [41] y `datetime` [52]. Así, verbigracia, la semana 10 del año 2005 correspondería con la el 7 de marzo de 2005 (2005-03-07 en formato compatible). Posteriormente, se enumeran las semanas del periodo de modo que a la primera semana de 2005 le corresponde el número 1 y así consecutivamente hasta la semana 52 del año 2014 que toma el número 522 en este orden. Como no todas las series de tiempo de las CIEs se encuentran completas, se interpola con base en el tiempo mediante la función `interpolate` [40] que toma como parámetro la semana epidemiológica y realiza una interpolación lineal de los casos faltantes con base en los presentes. Un ejemplo de estas series de tiempo se muestra en la figura 4.2 (p. 28), donde se puede ver que la figura 4.2a presenta un intervalo de tiempo menor a las otras dos series de tiempo, mientras que la 4.2c carece de datos registrados entre el año 2011 y 2012 los cuales fueron interpolados mediante la función mencionada y marcados con líneas verticales rojas. Con este incremento de datos interpolados, se pasa a 28 049 registros.



(a) Giardiasis



(b) SIDA



(c) Mordeduras por perro

Figura 4.2: Tres ejemplos de series de tiempo de los casos registrados normalizados por año, con marcas en rojo donde fueron interpolados los valores de los casos faltantes.

Debido a que en algunas series de tiempo carecen de información para todo el periodo estudiado, se eligen aquellas que al menos tengan cinco años de semanas registradas, o sea series de tiempo que cuenten con al menos 260 semanas. Al hacerlo, las CIEs se reducen de 143 a 40 y los registros a 26 242.

4.5 Caracterización y selección de características

Con estas reducciones, se pueden extraer características de las series de tiempo por cada CIE. Así, las semanas ordenadas secuencialmente se utilizan para encontrar la regresión lineal de la serie de tiempo contra los casos registrados y normalizados contra derechohabiencia. Esto se hace mediante la función `linregress` de la librería `SciPy` [61] que incluye el valor de la ordenada en el origen y la pendiente de la regresión lineal. Este último valor, además, muestra si las enfermedades presentan tendencia al alza o baja en el periodo de 2005 a 2015. Esta tendencia se elimina de cada serie de tiempo con la función `detrend` de la librería `SciPy` [60] y con ello se extraen las autocorrelaciones con retraso de 52 semanas correspondientes a un año mediante `acf` de la librería `StatsModels` [49]. El número de registros obtenidos con esta caracterización es de 40 con 54 características cada uno, a saber: la pendiente, ordenada en el origen y las autocorrelaciones con retraso de hasta 52 semanas para cada CIE. Ahora, pese a la reducción del número de registros, se aumentó considerablemente el número de características para cada registro, sin embargo esta dimensión también se puede reducir mediante algoritmos de selección de características. Se utiliza el algoritmo de *umbral de varianza* de `scikit-learn` [13] para seleccionar características. Finalmente, este algoritmo elimina características cuya varianza sea inferior a la media de las varianzas de las características.

4.6 Procedimiento k-medias

Una vez elegidas las mejores características para ejecutar el algoritmo de k-medias, se siguieron las recomendaciones de Ng para separa los datos en conjuntos de entrenamiento y desarrollo [39]. Para agrupar los datos se utiliza el algoritmo de **k-medias** de la librería `scikit-learn` [12]. Este método de agrupamiento no supervisado requiere especificar el número k de agrupamientos y toma como medición del error la suma de los cuadrados de las distancias entre los puntos y sus respectivos centroides, medida definida en la ecuación 2.2 (p. 6).

Para elegir el número de agrupamientos, se realizaron cincuenta réplicas donde se midió el error para cada conjunto de entrenamiento y prueba especificando un número k desde tres hasta once (la cantidad de CIEs generales presentes en los datos). Las medias de cada experimento dado el número de k de grupos se utilizan como parámetros de entrada del método del codo desarrollado por Satopää et al. [57] en 2011 en que se busca, en un conjunto de puntos, el punto de mayor curvatura medido como el punto cuya distancia es la mayor respecto a la recta que une los puntos extremos del conjunto. Dicho punto es el número de agrupamientos que se eligió para cada conjunto de datos.

A continuación, se realizan cincuenta iteraciones en las que, a partir del número de grupos definidos por el método del codo, se obtienen los errores del algoritmo de k-medias para los conjuntos de prueba y desarrollo variando la cantidad de datos en el conjunto de entrenamiento desde el número de grupos definido por el método del codo hasta el total de registros por conjunto de datos con incrementos de la décima parte de dicho total. Por ejemplo, si hubiera diez grupos determinados por el método del codo y un total de cien registros, el tamaño de muestra de entrenamiento en cada iteración sería de diez en diez hasta cien, o sea los tamaños

$\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

Una vez determinados tanto el número de agrupamientos por el método del codo y el tamaño de muestra que minimizan el error del algoritmo, se determina el agrupamiento al que pertenece cada uno de los 40 registros consistentes en las características seleccionadas por el umbral de varianza para cada CIE general.

Capítulo 5

Resultados

En este capítulo se presentan resultados de las pruebas descritas en la metodología. En primer lugar se muestra la cantidad de registros por cada CIE general en la figura 5.1 (p. 32). En esta figura se puede observar una gran presencia de las CIEs cuya clasificación general corresponden a la letra **A** y la **B**, quienes representan 42 % y 21 % de los registros. Estas letras corresponden a enfermedades infecciosas y parasíticas [68] de las que podría esperarse esta cantidad de registros dada la facilidad de su propagación, lo que puede cotejarse en Bailey [5]. Por año, en cambio, los registros permanecen constantes, lo cual puede comprobarse en la figura **??** (p. **??**).

En cuanto a los casos registrados normalizados agrupados por CIE general, se presenta una predominancia de la CIE general **J**, lo que puede cotejarse en la figura **??** (p. **??**). La CIE **J** es la asociada a las enfermedades respiratorias [68], cuya rápida

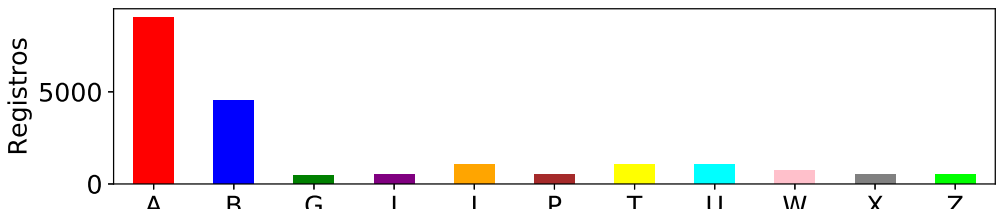


Figura 5.1: La figura contiene el conteo de CIEs generales de los registros seleccionados.

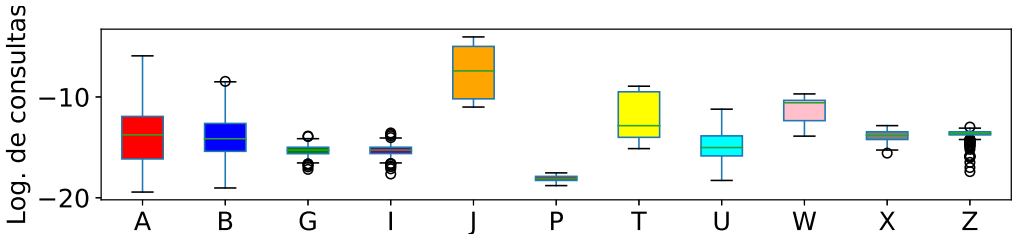
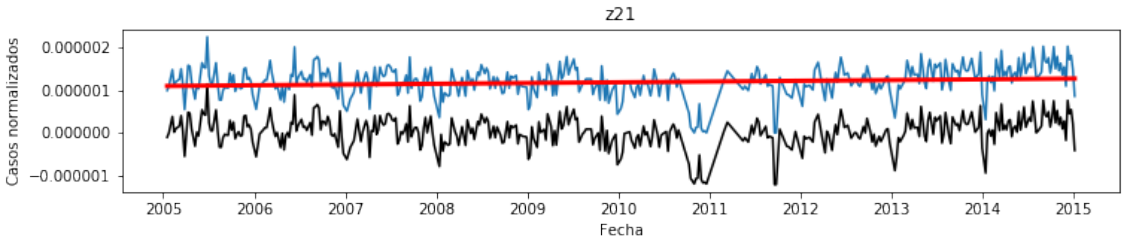


Figura 5.2: Logaritmo de casos normalizados por derechohabiencia.

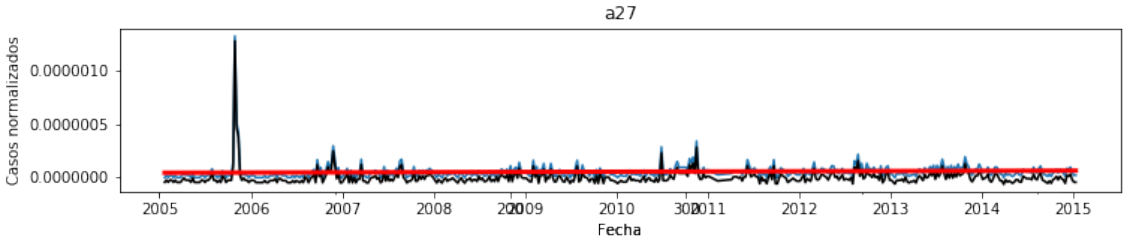
transmisión ha sido estudiada por autores como Cauchemez et al. [8], Klovdahl et al. [33], lo que explica la gran cantidad de casos registrados. Una mejor visualización de esta información aparece en la figura 5.2 (p. **??**).

Estos registros consisten en series de tiempo a los que se les extraen la pendiente y ordenada en el origen a partir de su regresión lineal y, tras restar la tendencia a la serie de tiempo, se pueden calcular las autocorrelaciones con retrasos de 1 hasta 52 semanas (un año). Tras este preprocesamiento se observa que algunas de las enfermedades estudiadas presentan una tendencia a la alza durante el intervalo de tiempo seleccionado para hacer esta investigación al rechazarse la hipótesis nula tal que la pendiente es igual a cero con intervalo de confianza del 95 %. Dichas enfermedades son, a saber, la infección asintomática por VIH, la tos ferina, y el cólera. Esta última presenta cero casos registrados por muchos años hasta el 2013 cuando surgió un brote de cólera en Hidalgo tras el paso de los huracanes Ingrid y Manuel [44]. Las gráficas de estos resultados pueden observarse en la figura 5.3 (p. 34), mientras que las de sus autocorrelaciones pueden consultarse en la figura 5.4 (p. 35)

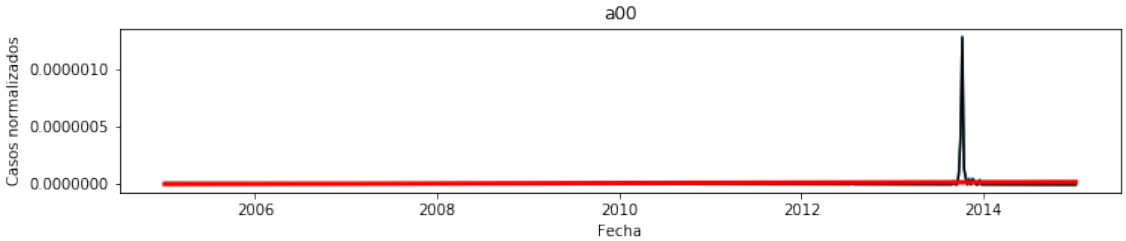
Esto deja con un total de 54 características por series de tiempo de cada CIE, sin tomar en cuenta ni el nombre de la enfermedad a la que pertenecen ni su CIE. Estos datos se pueden representar en una matriz de correlaciones que, a su vez, permite conocer las relaciones lineales entre las características que se tienen para los datos de interés. Dicha matriz se presenta en la figura 5.5 (p. 37), donde se aprecia



(a) Infección asintomática por VIH.

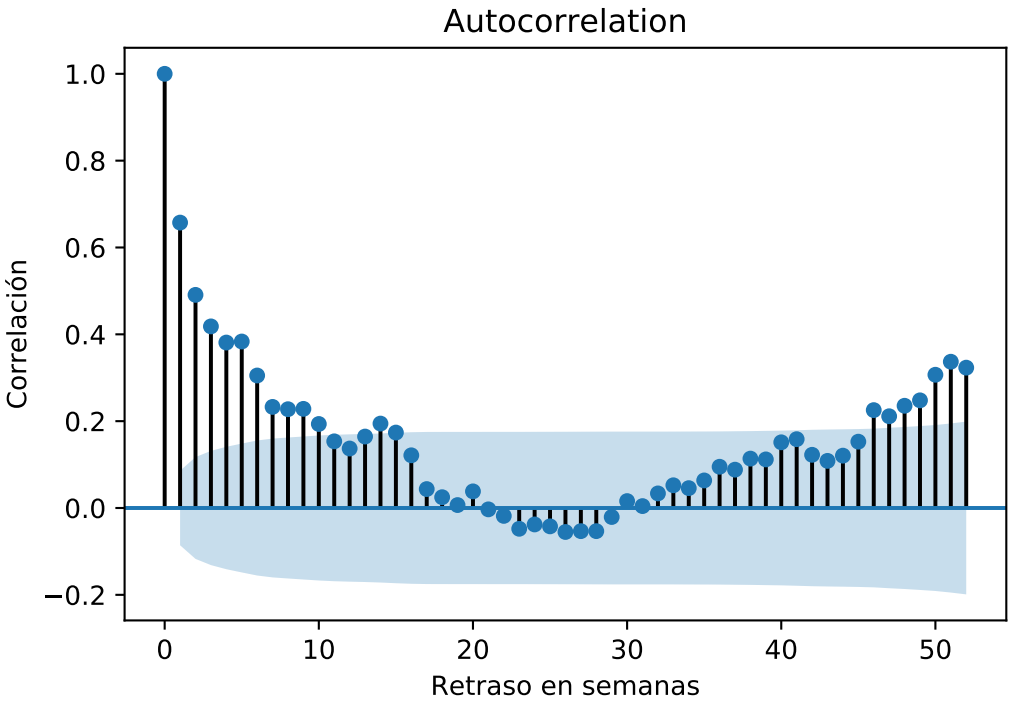


(b) Tos ferina.

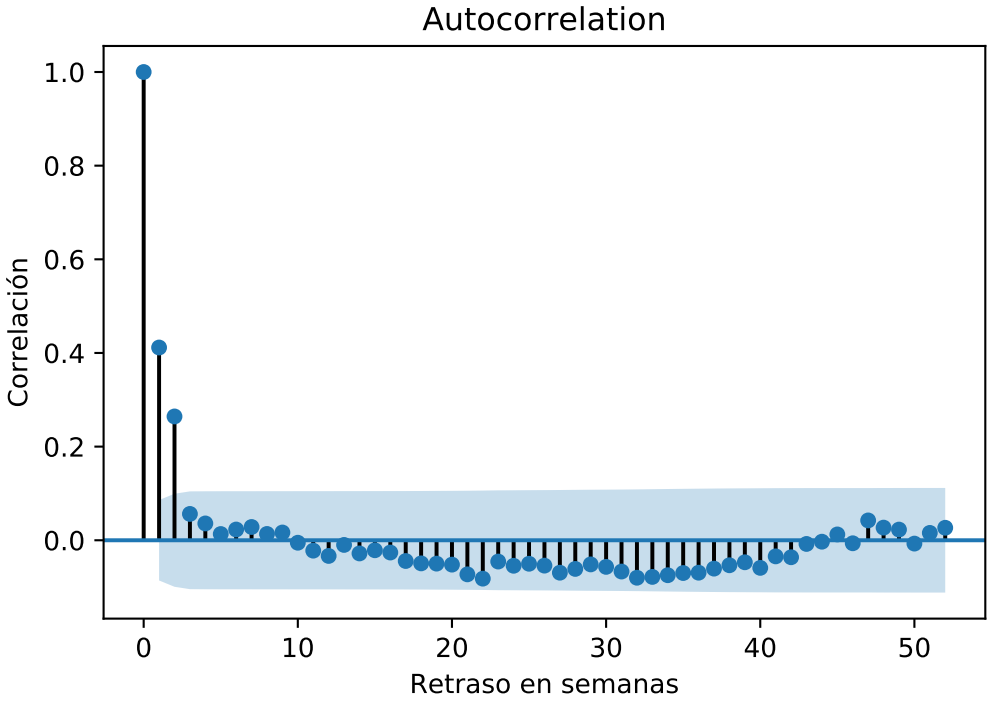


(c) Cólera.

Figura 5.3: Series de tiempo (en azul) con su pendiente (en rojo) y la serie de tiempo menos la tendencia (negro).



(a) Infección asintomática por VIH.



(b) Tos ferina.

Figura 5.4: Autocorrelaciones de las enfermedades cuyas tendencias crecen significativamente.

que las correlaciones positivas más fuertes se dan entre grupos de autocorrelaciones con retrasos muy cercanos entre sí. Además se encuentran dos grandes grupos de correlaciones positivas en las autocorrelaciones de retrasos menores a seis semanas, es decir, de hasta un mes y medio, y los de entre 44 y 52 semanas, asociados a los retrasos de diez a doce meses. Esto indica que durante estos periodos, los casos normalizados de las series de tiempo están influidos por la frecuencia con que se registraron casos de hasta un mes y medio de diferencia. Además, existen también correlaciones positivas significativas entre los retrasos de los primeros dos meses y los dos últimos meses del año, lo que marca una periodicidad anual entre los datos y la posibilidad de pronosticar casos registrados con estacionalidad mensual y anual. Luego, respecto a retrasos con el primer trimestre, tiende a no haber correlaciones, así que no podría predecirse mediante modelos lineales el comportamiento de los casos registrados entre cambio de estaciones del año. Pero cada semestre y hasta el octavo mes de diferencia, respecto al comportamiento de las primeras semanas, se tienen correlaciones negativas, lo que indica que la forma en que se registran enfermedades es inversamente proporcional entre ambos periodos, tal que si en un mes incrementa el número de consultas, un semestre después debería decrecer el número de consultas, y viceversa. Esta intuición refuerza la presencia de series estacionales con periodicidad anual.

En cuanto a la pendiente y ordenada en el origen de las regresiones lineales, pese a que entre ellas se hayan correlacionadas inversamente con mucha fuerza, no mantienen esta propiedad con el resto de las autocorrelaciones. Este tipo de datos suelen considerarse despreciables en los análisis estadísticos por ser atípicos dentro del conjunto al que pertenecen. A continuación y antes de proceder a la agrupación por k-medias de estos registros, se disminuyen sus características por medio del algoritmo del umbral de varianza, que también elimina aquellas características cuya varianza sea menor a un umbral determinado. Este algoritmo requiere que las

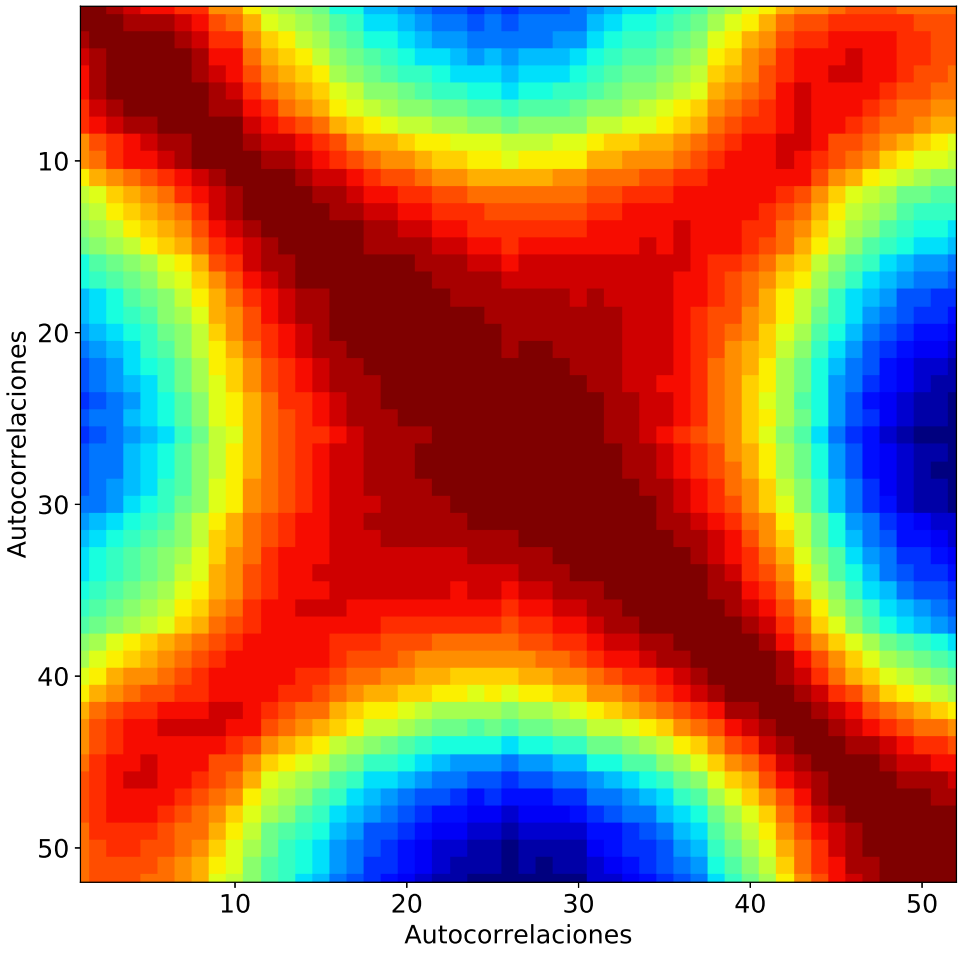


Figura 5.5: Matriz de correlación entre características de las series de tiempo estudiadas. Sobresalen las fuertes correlaciones entre las autocorrelaciones de hasta dos semanas, las de las primeras seis semanas entre sí, las de los últimos dos meses y, por otro lado, las de retrasos semestrales por tratarse de correlaciones negativas con las autocorrelaciones de las primeras seis semanas y las últimas ocho semanas del año.

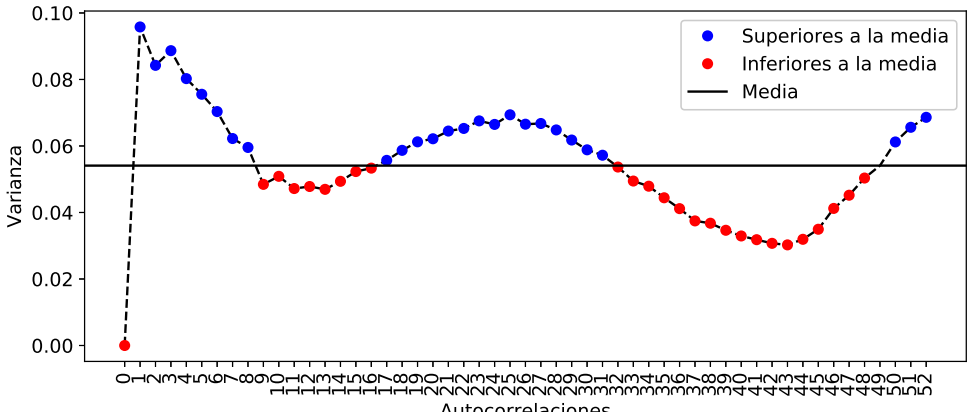


Figura 5.6: Todas las características cuya varianza se halla por encima del umbral, son seleccionadas para agrupar por k-medias.

variables sean normalizadas, por lo que se primero se normalizan por el método `MinMaxScaler` de `scikit-learn` [14]. Para estos datos se eligió como umbral de varianza el valor 0.06 dado por la mediana de los datos. Las características contra sus varianzas y el umbral denotado por una línea horizontal aparecen en la figura 5.6 (p. 38). Allí puede apreciarse que las características por debajo del umbral dado son eliminadas de las características significativas para el algoritmo de k-medias. Entre las descartadas se encuentran la pendiente y ordenada en el origen de las tendencias de las series de tiempo, que intuitivamente se esperaba despreciar desde la visualización de la matriz de correlación. Además, se conservan las autocorrelaciones de las primeras seis semanas, las de retrasos de un semestre y, finalmente, las de retrasos de diez a doce meses que también fueron destacadas por intuición visual en la discusión de la matriz de autocorrelaciones.

Los registros aparecen mezclados entre las distintas CIEs generales a las que pertenecen, por lo que se intuye que la agrupación por k-medias podría ajustarse poco a esta clasificación propuesta por la OMS. Sin embargo, el agrupamiento de estas series de tiempo permitirá conocer las características que comparten y lo que las diferencia. Así, se procede a la preparación de los datos para agruparlos por el

Cuadro 5.1: Cifras de los conjuntos de entrenamiento y desarrollo.

Conjunto	Porcentaje	Cantidad
Entrenamiento	67	25
Prueba	33	13

algoritmo propuesto. En primer lugar se separa el conjunto de datos en un conjunto de entrenamiento y uno de prueba. Como la cantidad de registros es pequeña, no hace separar el conjunto de prueba en uno de desarrollo como propone Ng [39]. Así, los conjuntos de entrenamiento quedan separados en los porcentajes mostrados en el cuadro 5.1 (p. 39).

Con base en esta separación de datos, se puede elegir el mejor número k de grupos para el algoritmo de k-medias con base en la medida del error definida en la ecuación 2.2 de suma de errores cuadrados y el método del codo desarrollado por Satopää et al. [57] en la que se ejecuta el algoritmo de k-medias con diferentes k hasta encontrar la distancia mayor de entre las distancias de las k y sus correspondientes errores hacia la recta que forman la primera y última medición del error de k. El algoritmo de k-medias se ajusta con el conjunto de entrenamiento, mientras que su error se mida con base en el conjunto de prueba. Los resultados para 50 réplicas de esta experimentación computacional se resumen en la figura 5.7 (p. 40) en donde se marca con una línea vertical el número $k = 4$ de agrupamientos, es decir: el número de agrupamientos que da más información sin comprometer los resultados.

Tras ejecutar el algoritmo de k-medias, se obtienen cinco grupos cuya distribución se muestra a través de un *análisis de componentes principales* (o PCA por sus siglas en inglés). Un análisis de componentes principales permite realizar una visualización bidimensional de los registros restantes por CIE general a fin de contar con una manera de cotejarlos. Este algoritmo ofrece una solución para este problema al presentar una proyección de cada registro a partir de la regresión lineal que mejor

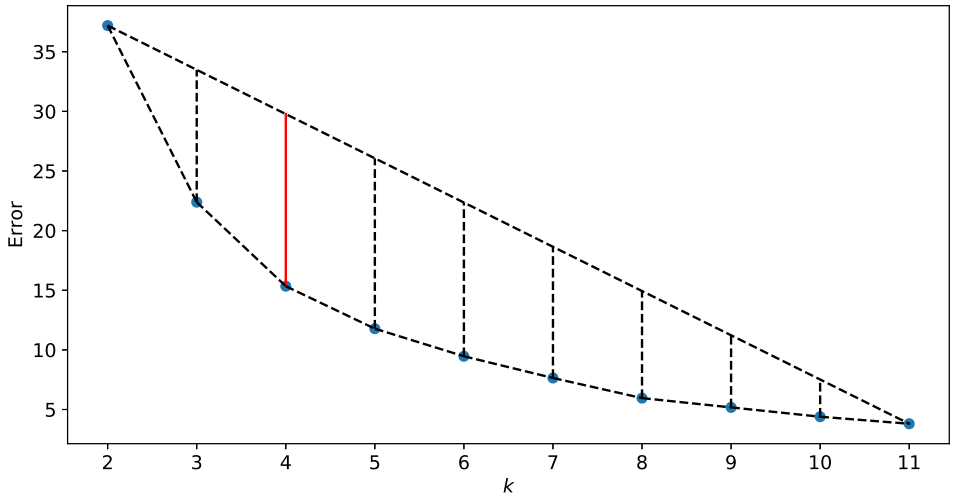


Figura 5.7: Errores con respecto al conjunto de prueba en diagramas de caja y bigotes para cada k elegida en el conjunto de entrenamiento.

se ajuste al conjunto de datos para, posteriormente, definir la perpendicular de dicha recta a partir del punto medio del segmento de recta definida entre los valores más extremos de los datos a los que la regresión lineal se ajusta. El análisis de componentes principales, además, ofrece la variación que logran recoger los componentes definidos a partir de la suma de errores cuadrados dividida entre la cantidad total de registros. Para el presente conjunto de datos, el primer componente recoge el 89 % de la variación de los datos, y el segundo componente un 7 % de la misma, de modo que el total de la variación recogida por estos componentes, 96 %, contiene casi la totalidad de la información ofrecida por las características de los datos.

La gráfica con los datos plasmados con base en estos dos componentes principales aparece en la figura 5.8 (p. 41). En dicha figura se observan las enfermedades representadas por círculos coloreados con base en el grupo al que pertenecen y en su centro presentan la letra más general de la CIE que les corresponde. Un vistazo la figura permite la intuición de que los grupos generados por k-medias no guardan relación con la CIE general dada por la OMS y una prueba de Wilcoxon [62]

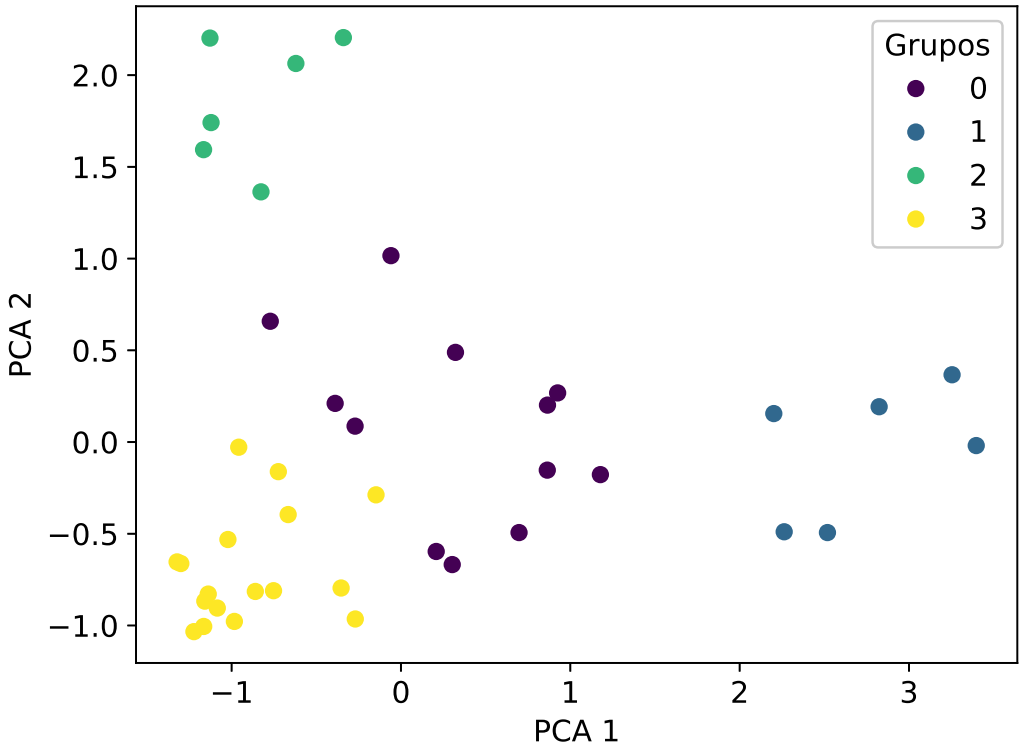
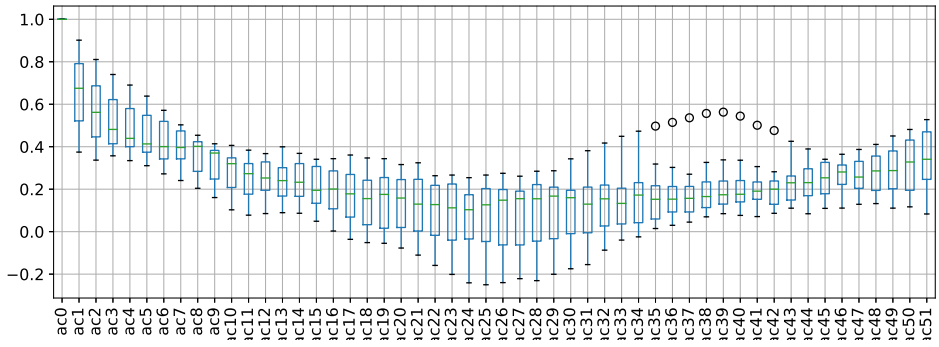


Figura 5.8: PCA de dos componentes principales de las enfermedades estudiadas (círculos) coloreadas con base al grupo generado por k-medias al que pertenecen y, dentro de cada círculo, la letra impresa de la CIE general que se les asigna.

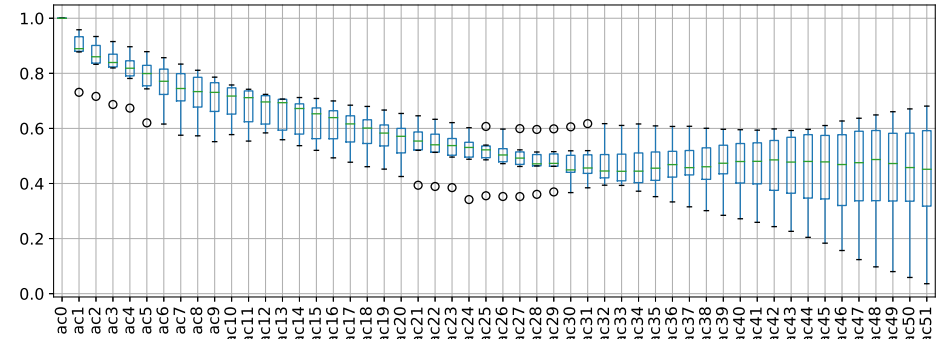
con $\alpha = 0.050$ entre los grupos y los factores de las CIEs generales arroja un valor $p = 0.257$ con lo que podemos concluir que ambos conjuntos de datos pertenecen a la misma distribución y, por lo tanto, no tienen relación estadísticamente significativa entre sí.

Ahora, se muestran diagramas de caja y bigotes de cada característica de las series de tiempo por cada grupo dado por k-medias, información hallada en la figura 5.9 (p. 44). En ella se puede observar que los grupos 0 y 4 de las figuras 5.9a y **??** contienen un componente estacional semestral denotado por la forma de campana que tienen sus autocorrelaciones y que alcanzan los valores más altos en las semanas correspondientes a retrasos de seis meses. También que la figura 5.9b muestra

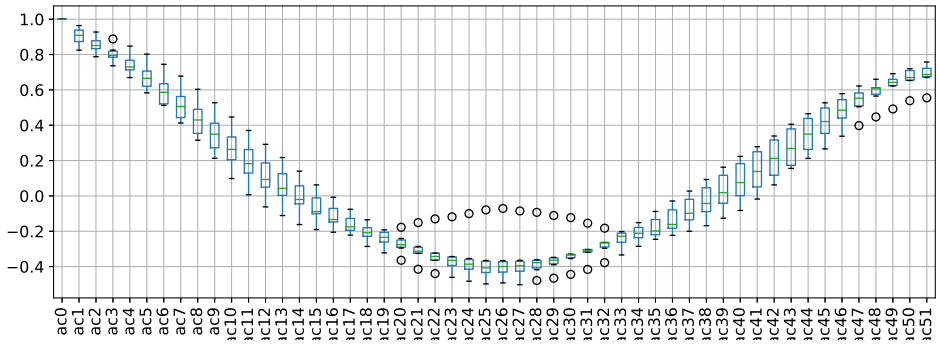
una curva que pareciera ser logarítmica pero cuyas autocorrelaciones nunca llegan a valores cercanos a cero. Cuando esto ocurre, las series de tiempo podrían ser estacionarias o contar con componentes residuales que las vuelvan difíciles de pronosticar. Por su parte, la figura 5.9c asociada al grupo 2 reúne las enfermedades que presentan un componente estacional anual claramente marcado por las altas autocorrelaciones del primer mes y último mes del año. Para terminar, la figura **??** tiene series de tiempo que mantienen sus autocorrelaciones constantes, señal indicativa de que se trata de series de tiempo ruidosas, aleatorias y generalmente impredecibles.



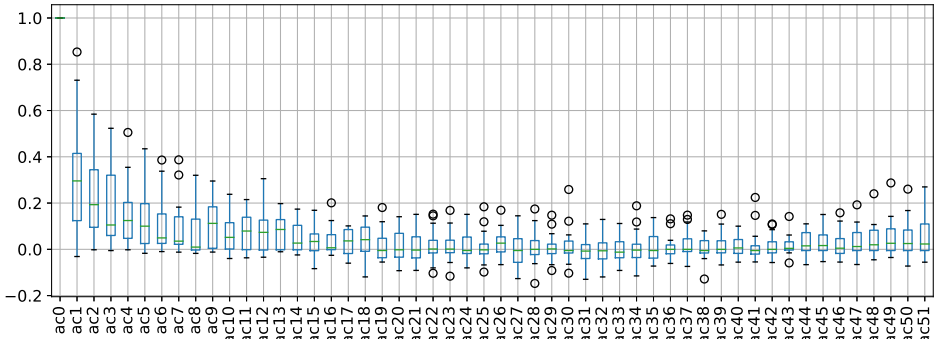
(a) Grupo 0



(b) Grupo 1



(c) Grupo 2



(d) Grupo 3

Figura 5.9: Las figuras 5.9a y 5.9d evidencian enfermedades con un componente estacional semestral fuertemente marcado; la figura 5.9b contiene enfermedades cuyas series de tiempo podrían ser estacionarias o impredecibles; la figura 5.9c agrupa series de tiempo de enfermedades con un componente estacionario anual fuertemente marcado.

Capítulo 6

Conclusiones

Este trabajo representa una novedosa y primera aproximación a los datos epidemiológicos reportados durante 2005 a 2015 por la Secretaría de Salud de México en documentos públicos compartidos en formato `PDF` cuyo contenido no había sido extraído para su estudio, motivo por el cual su extracción ofrece potencialmente una riqueza de resultados que podrían ayudar a comprender estos datos y proponer tomadores de decisiones a partir de los resultados que se puedan obtener de las mismas. Además, el preprocesamiento, caracterización de las series de tiempo y agrupamiento por k-medias implica un nuevo conocimiento de estos datos en que se comprende su forma de aparición y las relaciones meramente temporales entre las series de tiempo implicadas.

6.1 Contribuciones

La contribución principal respecto a la hipótesis planteada es que las enfermedades de los grupos generados por k-medias no guardan relación estadísticamente significativa con los grupos más generales de la CIE establecidos por la OMS, aunque puede concluirse que a partir de cinco grupos se tienen agrupamientos de enferme-

dades un error aceptable. De los cinco grupos arrojados por el algoritmo se observan autocorrelaciones que permitirían agrupar otras enfermedades de las que se conozca la frecuencia de consultas generadas a lo largo de cinco años.

Sobresale el descubrimiento de tres enfermedades cuya tendencia es positiva para el periodo estudiado, las cuales son la enfermedad asintomática del VIH, la tos ferina y la cólera, en orden de mayor a menor tendencia.

En cuanto a la selección de características, resalta el hecho de que la pendiente (tendencia) y la ordenada en el origen de las regresiones lineales de las series de tiempo fueron ambas descartadas por el algoritmo de umbral de varianza, así como las autocorrelaciones con retrasos de 5 a 43 semanas, o 2 a 10 meses, quedando las autocorrelaciones de semanas con retraso de un mes y de 11 y 12 meses. En cuanto a las autocorrelaciones con más correlación entre sí, se encuentran la de retraso de 3 y 4 semanas, y las de 51 y 52 semanas.

Con relación al origen de los datos, cabe destacar que el uso combinado de la información de posiciones por pixeles de un cuadro de un `PDF` ayuda a mejorar la precisión para definir el ancho de columnas en cuadros que puedan prescindir de dibujar las líneas que las delimiten. Así, se comparte un procedimiento efectivo de extracción de información de cuadros contenida en `PDF`s cuya labor resultaba difícil y que consiste en extraer información de encabezados de columnas de interés con las posiciones y dimensiones en pixeles de la página del rectángulo que las enmarca para con ello especificar los anchos de columna que se leerán por las herramientas propuestas.

Finalmente, este trabajo ofrece la primera cota de referencia respecto a futuros trabajos de agrupamiento de series de semanas epidemiológicas a partir de datos publicados por la Secretaría de Salud de México.

6.2 Trabajo a futuro

Puesto que este es el primer trabajo que agrupa estas series de tiempo, existen muchas ramificaciones de trabajo a futuro que se pueden explorar. En primer lugar podrían compararse otros algoritmos de agrupamiento contra k-medias, así como utilizar otras características para cada dato y otras medidas de distancias.

Existen, además, registros diarios de consultas a lo largo de la república mexicana con los que podrían cotejarse estos resultados, mejorar la precisión que, también, tienen características valiosas como el género del paciente, la CIE por la que se fue a consultar y con la que fue diagnosticado tras la consulta, entre otros. De igual manera, existe información georreferenciada, social y médica que puede ser asociada a estos datos. Entre esta información, despunta la labor de asociar a los grupos generados por el algoritmo de k-medias en este estudio, los síntomas que presentan las enfermedades contenidas en los mismos, para conocer si existe alguna relación entre ambos.

Otra de las áreas de interés a partir de los resultados obtenidos consiste en comparar los grupos generados respecto a la CIE 11, actual clasificación de enfermedades, versión que toma en cuenta la frecuencia de registros de enfermedades para su clasificación, a diferencia de la CIE 10 sobre la que se realizaron las comparaciones en este estudio por ser la que coincidía temporalmente con el periodo estudiado.

Finalmente, estos resultados y los que se podrían obtener de otros algoritmos de agrupamiento pueden ser utilizados para mejorar algoritmos de clasificación y pronóstico.

Bibliografía

[1] Adobe (2018). Lector de PDF, visor de PDF — Adobe Acrobat Reader DC. `https://acrobat.adobe.com/mx/es/acrobat/pdf-reader.html` [Accedido: 2018-11-26].

[2] Arias, J. R. (2006). What is an epidemiological week and why do we use them? *The Seeker*, 6(1):7.

[3] Ariga, A. (2018). chezou/tabula-py: Simple wrapper of tabula-java: extract table from pdf into pandas dataframe. `https://github.com/chezou/tabula-py` Accedido: 2018-07-01.

[4] Bagnall, A. y Janacek, G. (2005). Clustering time series with clipped data. *Machine Learning*, 58(2):151–178.

[5] Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications.* Charles Griffin and Company Ltd, High Wycombe, UK.

[6] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, Singapore.

[7] Brockwell, P. J. y Davis, R. A. (2016). *Introduction to Time Series and Forecasting.* Springer, Switzerland.

[8] Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., y Swerdlow, D. (2011). Role of social networks in shaping disease

transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830.

[9] Chen, J. R. (2005). Making subsequence time series clustering meaningful. In *Fifth IEEE International Conference on Data Mining*.

[10] Corduas, M. y Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, 52(4):1860–1872.

[11] Darlington, R. B. y Hayes, A. F. (2017). *Regression Analysis and Linear Models. Concepts, Applications, and Implementation*. The Guilford Press, London, UK.

[12] Desarrolladores de scikit-learn (2019a). 2.3.2. k-means. `https://scikit-learn.org/stable/modules/clustering.html#k-means` Accedido: 2019-03-12.

[13] Desarrolladores de scikit-learn (2019b). sklearn.feature_selection.variancethreshold. `https://scikit-learn.org/stable/modules/generatedsklearn.feature_selection.VarianceThreshold.html#sklearn.feature_selection.VarianceThreshold` Accedido: 2019-03-22.

[14] Desarrolladores de scikit-learn (2019c). sklearn.preprocessing.minmaxscaler. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html` Accedido: 2019-03-22.

[15] D'Urso, P. y Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24):3565–3589. Theme: Non-Linear Systems and Fuzzy Clustering.

[16] Ernst, J., J. Nau, G., y Bar-Joseph, Z. (2005). Clustering short time series gene expression data. In *Proceedings of the Sixth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, pages 285–289, New York. ACM.

[17] Ferreira, L. N. y Zhao, L. (2015). Time Series Clustering via Community Detection in Networks. *arXiv e-prints*, 1:1–23.

[18] Focardi, S. M. y Fabozzi, F. J. (2004). A methodology for index tracking based on time-series clustering. *Quantitative Finance*, 4(4):417–425.

[19] Free Software Foundation (2011). Gawk–GNU Project–Free Software Foundation (FSF). `https://www.gnu.org/software/gawk/gawk.html` Accedido: 02-02-2019.

[20] Frühwirth-Schnatter, S. y Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, 26(1):78–89.

[21] Fulcher, B. D. y Jones, N. S. (2014). Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037.

[22] Hartigan, J. A. y Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[23] Hautamäki, V., Nykänen, P., y Fränti, P. (2008). Time-series clustering by approximate prototypes. In *2008 19th International Conference on Pattern Recognition*, pages 1–4.

[24] Instituto nacional de estadística y geografía (2016). Estadísticas por tema. `http://www3.inegi.org.mx/sistemas/sisept/default.aspx?t=msoc01` Accedido: 2018-10-29.

[25] Instituto nacional de estadística y geografía (2018). Población. `http://www.beta.inegi.org.mx/temas/estructura/` Accedido: 2018-09-28.

[26] Izakian, H., Pedrycz, W., y Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244.

[27] Jain, A. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666.

[28] Kalpalis, K., Gada, D., y Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 273–280, California. IEEE.

[29] Kavitha, V. y Punithavalli, M. (2010). Clustering time series data stream – a literature survey. *International Journal of Computer Science and Information Security*, 8.

[30] Keogh, E. y Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and Information Systems*, 8(2):154–177.

[31] Keogh, E. J. y Pazzani, M. J. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 239–243, New York. Association for Computing Machinery.

[32] Keogh, E. J. y Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 285–289, New York. ACM.

[33] Klovdahl, A., Graviss, E., Yaganehdoost, A., Ross, M., Wanger, A., Adams, G., y Musser, J. (2001). Networks and tuberculosis: an undetected community outbreak involving public places. *Social Science and Medicine*, 52(5):681–694.

[34] Lai, R. K., Fan, C.-Y., Huang, W.-H., y Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, 36(2, Part 2):3761–3773.

[35] Layton, R., Watters, P., y Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, pages 1–8.

[36] Li, L. y Prakash, A. (2011). Time series clustering: Complex is simpler! *Proceedings of the 28th International Conference on Machine learning*, pages 185–192.

[37] Lin, J., Vlachos, M., Keogh, E., y Gunopulos, D. (2004). Iterative incremental clustering of time series. In *Advances in Database Technology*, pages 106–122, Berlin. Springer.

[38] Möller-Levet, C. S., Klawonn, F., Cho, K.-H., y Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *Advances in Intelligent Data Analysis V*, pages 330–340, Berlin. Springer.

[39] Ng, A. Y.-T. (2018). Machine learning yearning. `https://www.deeplearning.ai/machine-learning-yearning/`.

[40] NumFOCUS (2019a). pandas.series.interpolate. `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html` Accedido: 2019-03-22.

[41] NumFOCUS (2019b). Python data analysis library. `https://pandas.pydata.org/` Accedido: 2019-04-07.

[42] Oates, T. (1999). Identifying distinctive subsequences in multivariate time series by clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 322–326, New York, NY, USA. ACM.

[43] Oates, T., Firoiu, L., y Cohen, P. R. (1999). Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pages 17–21.

[44] Organización Mundial de la Salud (2013). Cólera en México – Actualización. `https://www.who.int/csr/don/2013_11_13/es/` Accedido: 2019-06-19.

[45] Organización Mundial de la Salud (2018). La Organización Mundial de la Salud (OMS) publica hoy su nueva Clasificación Internacional de Enfermedades (CIE-11). `https://www.who.int/es/news-room/detail/17-06-2018-who-releases-new-international-classification-of-diseases(icd-11` Accedido: 2019-03-20.

[46] Organization, W. H. (2016). *International statistical classification of diseases and related health problems—10th revision*. WHO Library Cataloguing, France.

[47] Paparrizos, J. y Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1855–1870, New York, NY, USA. ACM.

[48] Paparrizos, J. y Gravano, L. (2017). Fast and accurate time-series clustering. *ACM Transactions Database Systems*, 42(2):8:1–8:49.

[49] Perktold, J., Seabold, S., y Taylor, J. (2019). statsmodels.tsa.stattools.acf.

https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.acf.html Accedido: 2019-04-07.

[50] Phaseit Inc. and Mathieu Fenniak (2016). PyPDF2 Documentation. https://pythonhosted.org/PyPDF2/?utm_source=recordnotfound.com Accedido: 02-07-2018.

[51] Phython Software Foundation (2018). Python 3.7.0. https://www.python.org/downloads/release/python-370/ Accedido: 2018-08-13.

[52] Python Software Foundation (2019). datetime – basic date and time types. https://docs.python.org/3.8/library/datetime.html Accedido: 2019-04-07.

[53] Rakthanmanon, T., Keogh, E. J., Lonardi, S., y Evans, S. (2011). Time series epenthesis: Clustering time series streams requires ignoring some data. In *2011 IEEE 11th International Conference on Data Mining*, pages 547–556.

[54] Rakthanmanon, T., Keogh, E. J., Lonardi, S., y Evans, S. (2012). MDL-based time series clustering. *Knowledge and Information Systems*, 33(2):371–399.

[55] Ratanamahatana, C., Keogh, E., Bagnall, A. J., y Lonardi, S. (2005). A novel bit level time series representation with implication of similarity search and clustering. In *Advances in Knowledge Discovery and Data Mining*, pages 771–777, Berlin. Springer.

[56] Rodrigues, P. P., Gama, J., y Pedroso, J. P. (2008). Hierarchical clustering of time-series data streams. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):615–627.

[57] Satopää, V., Albrecht, J., Irwin, D., y Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

[58] Secretaría de Salud (2016). Boletín Epidemiológico Sistema Nacional de Vigilancia Epidemiológica Sistema Único de Información. https://www.gob.mx/salud/acciones-y-programas/direccion-general-de-epidemiologia-boletin-epidemiologico Accedido: 2019-05-23.

[59] Singhal, A. y Seborg, D. (2002). Clustering of multivariate time-series data. In *Proceedings of the 2002 American Control Conference*, pages 273–280, Arkansas. IEEE.

[60] The SciPy community (2019a). scipy.signal.detrend. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.detrend.html Accedido: 2019-04-07.

[61] The SciPy community (2019b). scipy.stats.linregress – scipy v1.2.1 reference guide. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html Accedido: 04-07-2019.

[62] The SciPy community (2019c). scipy.stats.wilcoxon. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html Accedido: 2019-03-22.

[63] Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Community ACM*, 11(6):419–422.

[64] Vlachos, M., Lin, J., Keogh, E., y Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. *Proceedings Workshop on Clustering High Dimensionality Data and its Applications*, pages 1–12.

[65] w69b (2018). PDF Mergy – WebApp to merge PDF files. https://pdfmerge.w69b.com/ Accedido: 2018-11-23.

[66] Wang, X., Smith, K., y Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3):335–364.

[67] Wang, X., Wirth, A., y Wang, L. (2007). Structure-based statistical features and multivariate time series clustering. In *Seventh IEEE International Conference on Data Mining*, pages 351–360.

[68] World Health Organization (2016). International Statistical Classification of Diseases and Related Health Problems 10th Revision. `https://icd.who.int/browse10/2016/en` Accedido: 2018-09-30.

[69] World Health Organization (2018). WHO — International Classification of Diseases, 11th Revision (ICD-11). `http://www.who.int/classifications/icd/en/` Accedido: 2018-09-30.

[70] Xiong, Y. y Yeung, D.-Y. (2002). Mixtures of ARMA Models for Model-Based Time Series Clustering. In *2002 IEEE International Conference on Data Mining*, pages 717–720, Maebashi. IEEE.

[71] Xiong, Y. y Yeung, D.-Y. (2004). Time series clustering with ARMA mixtures. *Pattern Recognition*, 37(8):1675–1689.

[72] Yildiz, B., Kaiser, K., y Miksch, S. (2005). pdf2table: A method to extract table information from pdf files. In *Indian International Conference on Artificial Intelligence*.

[73] Zakaria, J., Mueen, A., y Keogh, E. (2012). Clustering time series using unsupervised-shapelets. In *2012 IEEE 12th International Conference on Data Mining*, pages 785–794.

[74] Zhang, H., Ho, T., Zhang, Y., y Lin, S. (2006). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica (Slovenia)*, 30:305–319.

[75] Zhang, X., Liu, J., Du, Y., y Lv, T. (2011). A novel clustering method on time series data. *Expert Systems with Applications*, 38(9):11891–11900.

Apéndice A

CIEs y sus nombres de enfermedades

Cuadro A.1: CIEs y el nombre de la enfermedad correspondiente presentes en la población de 23 721 registros tomados de los datos obtenidos a nivel nacional.

CIE	Enfermedad
a00	Cólera
a01.0	Fiebre tifoidea
a01.1-a02	Fiebre paratifoidea
a01.1-a02	Paratifoidea
a01-a03	Enfermedades infecciosas intestinales
a02	Otras salmonelosis
a03	Shigelosis
a04 a08-a09	Infección intestinal debida a virus y otros organismos
a04	Infecciones intestinales por otros organismos

a05	Intoxicación alimentaria bacteriana
a06.0-a06.3	Amebiasis intestinal
a06.4	Absceso hepático amebiano
a07.0	Otras infecciones intestinales debidas a protozoarios
a07.1	Giardiasis
a08.0	Enteritis debida a rotavirus
a15-a16	Tuberculosis respiratoria
a17.0	Meningitis tuberculosa
a17.1	Tuberculosis otras formas
a23	Brucelosis
a27	Leptospirosis
a30	Lepra
a33	Tétanos neonatal
a34	Tétanos
a37	Tos ferina
a38	Escarlatina
a39.0	Meningitis meningocócica
a40.3	Enfermedad invasiva por neumococo
a41.3	Infecciones invasivas por *haemophilus influenzae*
a46	Erisipela
a50	Sífilis congénita

a51-a53	Sífilis adquirida
a54.0-a54.2	Infección gonocócica genitourinaria
a55	Linfogranuloma venéreo por clamidias
a57	Chancro blando
a59.0	Tricomoniasis urogenital
a60.0	Herpes genital
a67	Mal del pinto
a71	Tracoma
a75.0	Tifo epidémico
a75.2	Tifo murino
a75.9	Otras rickettsiosis
a76.0	Vulvovaginitis inespecífica
a77.0	Fiebre manchada
a90	Dengue clásico
a91	Fiebre hemorrágica por dengue
a92.0	Enfermedad por virus chikungunya
a92.3	Fiebre del oeste del Nilo
b01	Varicela
b06	Rubeola
b15	Hepatitis vírica A
b16	Hepatitis vírica B

b17.1	Hepatitis vírica C
b17-b19	Otras hepatitis víricas
b20-b24	SIDA
b26	Parotiditis infecciosa
b30	Conjuntivitis
b30.3	Conjuntivitis epidémica aguda hemorrágica
b37.3-b37.4	Candidiasis urogenital
b50	Paludismo por *p. falciparum*
b51	Paludismo por *p. vivax*
b55.0	Leishmaniasis visceral
b55.1	Leishmaniasis cutánea
b57	Tripanosomiasis americana
b58	Toxoplasmosis
b60.2	Meningoencefalitis amebiana primaria
b65-b67	Otras helmintiasis
b68	Teniasis
b69	Cisticercosis
b73	Oncocercosis
b75	Triquinosis
b77	Ascariasis
b80	Enterobiasis

b86	Escabiosis
b97.7	Infección por virus de papiloma humano
c50	tumor maligno de la mama
c53	tumor maligno del cuello del útero
e01	Bocio
e10	Diabetes mellitus insulinodependiente (tipo i)
e11-e14	Diabetes mellitus no insulinodependiente (tipo ii)
e40-e43	Desnutrición severa
e44.0	Desnutrición moderada
e44.1	Desnutrición leve
e66	Obesidad
f10.0	Intoxicación aguda por alcohol
f10-f19	Adicciones
f32	Depresión
f50	Anorexia y bulimia
g00-g03	Meningitis
g20	Enfermedad de Parkinson
g30	Enfermedad de Alzheimer
h10	Conjuntivitis
h65.0-h65.1	Otitis media aguda
i00-i02	Fiebre reumática aguda

i10-i15	Hipertensión arterial
i20	Enfermedad isquémica del corazón
i60-i67	Enfermedad cerebrovascular
i87.2	Insuficiencia venosa periférica
j00-j06	Infecciones respiratorias agudas
j02.0	Faringitis y amigdalitis estreptocócicas
j09	Influenza a H1N1
j09-j11	Influenza
j12	Neumonías y bronconeumonías
j45	Asma
k05	Gingivitis y enfermedad periodontal
k25-k29	Úlceras, gastritis y duodenitis
k70	Enfermedad alcohólica del hígado
k70.3	Cirrosis hepática
n30	Infección de vías urinarias
n40	Hiperplasia de próstata
n87.0-n87.1	Displasia cervical leve y moderada
n87.2	Displasia cervical severa y cacu in situ
o24.4	Diabetes mellitus en el embarazo
p35.0	Rubeola congénita
q00	Anencefalia

q01	Encefalocele
q05	Espina bífida
q35-q37	Labio y paladar hendido
r50	Síndrome febril
t20-t32	Quemaduras
t58	Intoxicación por monóxido de carbono
t60	Intoxicación por plaguicidas
t63 excepto t63.2	Intoxicación por animales venenosos
t63.2	Intoxicación por picadura de alacrán
t63.2	Intoxicación por veneno de escorpión
t63x21	Intoxicación por ponzoña de animales
t67	Efectos del calor y de la luz
t68	Hipotermia
u97	Enfermedad febril exantemática
u98	Parálisis flácida aguda
u99	Síndrome coqueluchoide
v01-v09	Peatón lesionado en accidente de transporte
v20-v29 v40-v79	Accidente de transporte en vehículos con motor
w32-w34	Herida por arma de fuego y punzocortantes
w54	Mordeduras por perro
w55	Mordeduras por otros mamíferos

x20	Mordeduras por serpiente
y07.0-y07.2	Lesiones por violencia intrafamiliar
y58	Efectos adversos temporalmente asociados a vacunas
y95	Afección nosocomial
z21	Infección asintomática por VIH

Resumen autobiográfico

José Alberto Benavides Vázquez

Candidato para obtener el grado de

Maestría en Ciencias

en Ingeniería de Sistemas

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Tesis:

Agrupamiento no supervisado de series de tiempo epidemiológicas de México entre 2005 y 2015

Nací el 9 de agosto de 1987 en la ciudad de Monterrey, México; mis padres son José Loreto Benavides Ruíz y Bertha Alicia Vázquez Méndez. En 2012 egresé como Licenciado en Filosofía y Humanidades en la Facultad de Filosofía y Letras de la Universidad Autónoma de Nuevo León (UANL). En 2017 concluí mis estudios en la Licenciatura de Multimedia y Animación Digital en la Facultad de Ciencias Físico Matemáticas de la misma Universidad.