

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

José Alberto Benavides Vázquez

MAESTRÍA EN INGENIERÍA CON ESPECIALIDAD EN INGENIERÍA DE SISTEMAS
FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

18 de septiembre de 2019

5 Conclusiones

5 Conclusiones

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.

- Proponer metodología para extraer y preparar datos
- Establecer procedimiento de preprocesamiento de datos
- Obtener características temporales representativas
- Describir los grupos generados

1 Introducción

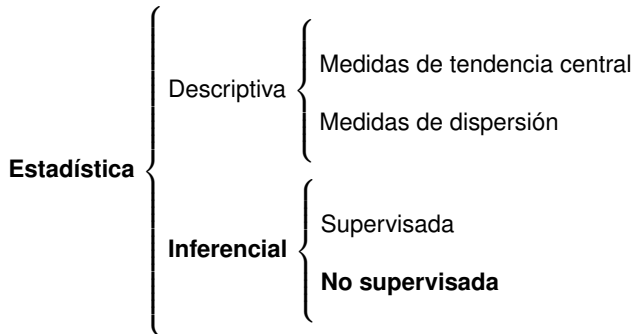
2 Antecedentes

3 Metodología

- Obtención de los datos
- Series de tiempo
- Caracterización
- Agrupamiento

4 Resultados

5 Conclusiones



Algoritmos de agrupamiento para series de tiempo (1979 a 2017)

		Artículos
Métodos	k-medias	[3, 6, 11, 15, 20, 22, 23, 27, 28, 34, 39]
	DTW	[17, 21, 26, 50]
	ARMA	[3, 46, 47]
	ARIMA	[7, 19]
	Jerárquico	[35, 50]
	Red compleja	[12]
Entrada	ACF	[10, 11, 39, 44]
	Wavelets	[23, 42, 49]

5 Conclusiones

Origen

Archivos en línea

- Secretaría de Salud de México
- Boletines epidemiológicos
- Medio oficial del Sistema Nacional de Vigilancia Epidemiológica [38]
- Publicación semanal (51 o 52 por año)
- Contienen entre 20 y 70 páginas

ENTIDAD FEDERATIVA	Neumonías y Bronconeumonías CIE-10 ^a REV. J12-J18 excepto J18.2				Influenza (A H1N1) CIE-10 ^a REV. J09			Influenza Estacional CIE-10 ^a REV. J10-J11		
	2012			2011	2012			2012		2011
	Sem.	Acum.		Acum.	Sem.	Acum.		Acum.		Acum.
		M	F			M	F	M	F	
Aguaascalientes	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	88	210	184	417	9	20	31	2	1	-
Hidalgo	100	250	244	677	16	25	21	1	1	-
Jalisco	100	250	244	677	16	25	21	1	1	-
México	1 000	2 500	2 444	6 777	166	250	282	13	21	65
Morelos	50	125	122	339	10	12	10	0	0	-
Nayarit	20	50	48	86	4	11	16	-	3	-
Oaxaca	60	150	144	417	12	15	10	0	0	-
Puebla	70	175	168	477	14	17	13	0	0	-
Quintana Roo	30	75	72	201	6	7	5	0	0	-
Querétaro	40	100	96	267	8	10	8	0	0	-
Sinaloa	50	125	122	339	10	12	10	0	0	-
Sonora	60	150	144	417	12	15	10	0	0	-
Tamaulipas	70	175	168	477	14	17	13	0	0	-
Tlaxcala	30	75	72	201	6	7	5	0	0	-
Veracruz	80	200	192	537	16	20	16	0	0	-
Yucatán	40	100	96	267	8	10	8	0	0	-
Zacatecas	50	125	122	339	10	12	10	0	0	-

- URL cambia con el tiempo
- Distintos extensiones de archivos (PDF, ZIP, JPG)
- Diferentes estilos de cuadros
- Formatos de números no estandarizados

Extracción

```

para cada directorio en año hacer
    para cada archivo en directorio hacer
        si termina con .pdf entonces
            leer páginas con PyPDF2 [30];
            para cada página en archivo hacer
                si contiene cuadro de interés entonces
                    extraer contenido con tabulapy [2];
                    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
    
```

Resultado

2006|02|Tabasco| 'Paratifoidea\r y otras
salmonelosis\rCIE-10a REV.\rA01.1-A02'| '121'

784 660 registros

```
for i in range(len(lines)):
    lines[i][1] = lines[i][1][-2:]

    lines[i][4] = str(lines[i][4])
    lines[i][4] = lines[i][4].replace("'", "")
    lines[i][4] = re.sub("^s+", "", lines[i][4])
    if lines[i][4][0] is "-":
        lines[i][4] = "0"
    if " " in lines[i][4]:
        lines[i][4] = lines[i][4][:lines[i][4].find(" ")]
    lines[i][4] = lines[i][4].replace("n.e", "NA")
    lines[i][4] = lines[i][4].replace("n.d", "NA")
    lines[i][4] = lines[i][4].replace(".", "")
    lines[i][4] = lines[i][4].replace(" ", "")

    lines[i][3] = re.sub("\s+", " ", lines[i][3])
    lines[i][3] = re.sub("^s+", "", lines[i][3])
    lines[i][3] = lines[i][3].replace("'", "")
    lines[i][3] = lines[i][3].replace("\\r", " ")
```


Simplificación y resultado

Resultados nacionales (23 722 registros)

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

Conversión

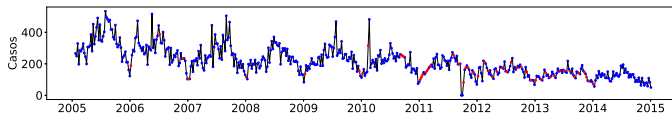
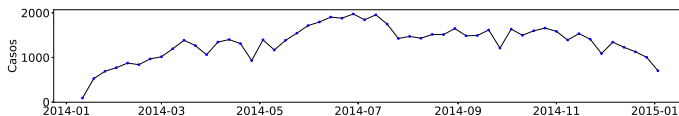
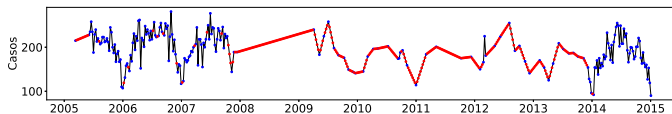
Software para series de tiempo

- Python v3 [31]: Lenguaje de programación de alto nivel.
- Pandas [25]: *Python Data Analysis Library*.
- datetime [33]: Manipulación de fechas y horarios.
- interpolate [24]: Interpolación por frecuencia semanal.
- detrend [40]: Eliminación de la tendencia.

Procedimiento (28 049 registros)

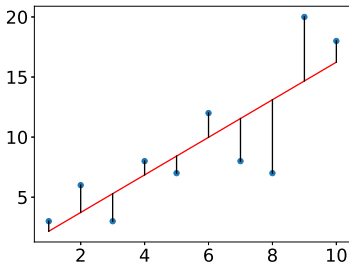
- 1 Convertir años y semanas epidemiológicas a fechas
- 2 Separar por CIE
- 3 Aplicar frecuencias semanales
- 4 Interpolar datos faltantes

Series de tiempo

**Figura:** A03: Shigelosis.**Figura:** A02: Otras salmonelosis.**Figura:** W55: Mordeduras por otros mamíferos.

Regresión lineal

Recurso: `linregress` de la librería `SciPy` [41]



$$\hat{y}(t, \omega) = \omega_0 + \omega_1 x_1 + \dots + \omega_t x_t. \quad (1)$$

$$\min \left(\sum_t (y_t - \hat{y}_t)^2 \right). \quad (2)$$

Series de tiempo

Normalización por derechohabientes (INEGI)

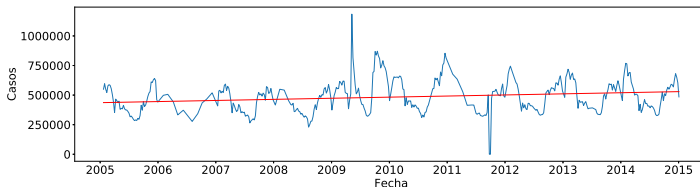


Figura: J00–J06: Infecciones respiratorias agudas sin normalizar.

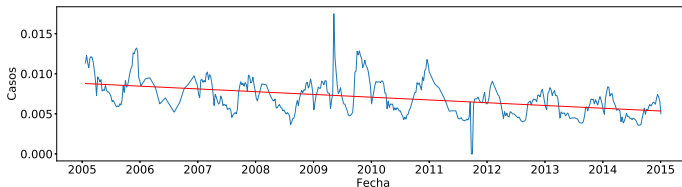


Figura: J00–J06: Infecciones respiratorias agudas normalizadas.

Series de tiempo

Enfermedades con pendiente positiva significativa

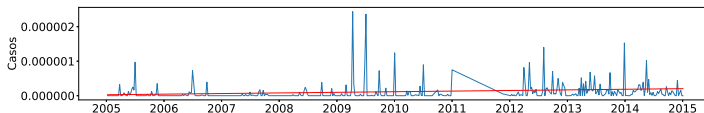


Figura: A37: Tos ferina.

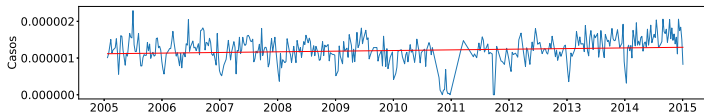


Figura: Z21: Infección asintomática por VIH.

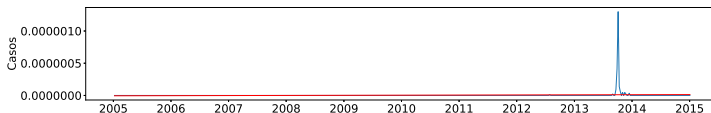


Figura: A00: Cólera.

Caso del cólera

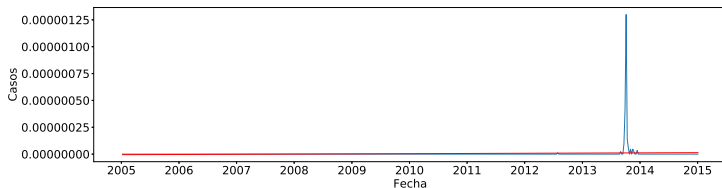


Figura: A00: Cólera.

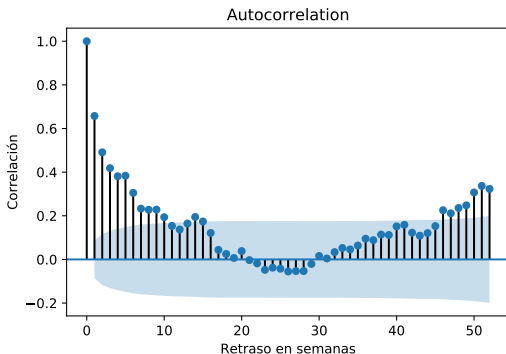
- Estado de Hidalgo 2013
- Posible causa: Huracanes Ingrid y Manuel
- 95 % tenían cepa igual a países caribeños
- 5 % tenían cepa igual al brote 1991–2001 en México

Autocorrelaciones

Recurso: `acf` de la librería `StatsModels` [29]

$$\gamma_x(h) = \text{CoV}(X_{t+h}, X_t) \quad (3)$$

$$\hat{p}_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(X_{t+h}, X_t). \quad (4)$$



k-medias

Recurso: `k-means` de `scikit-learn` [8]

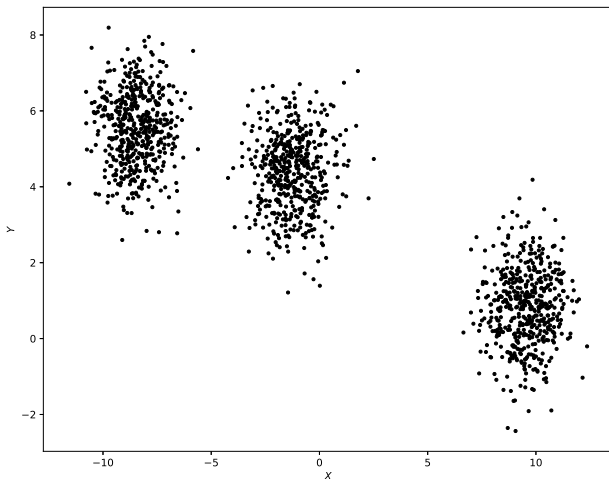
- $X = x_i, i = 1, \dots, n$ puntos d -dimensionales
- $C = c_j, j = 1, \dots, k$ centroides
- μ_j : media de los puntos $x_i \in c_j$

$$J(C) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (5)$$

$$\text{mín}(J(C)) \quad (6)$$

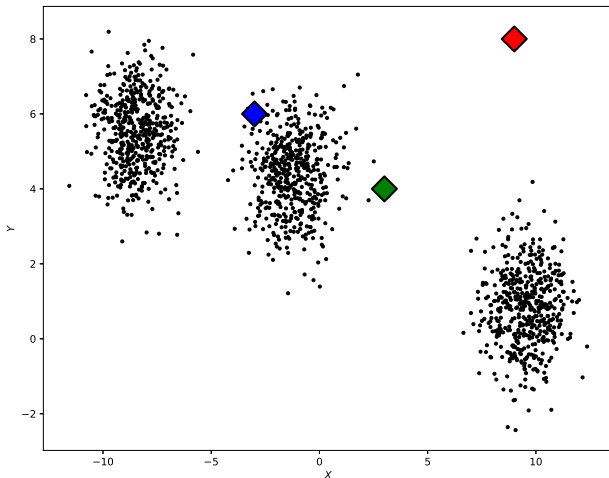
Algoritmo

- 1 Seleccionar un número k de grupos.



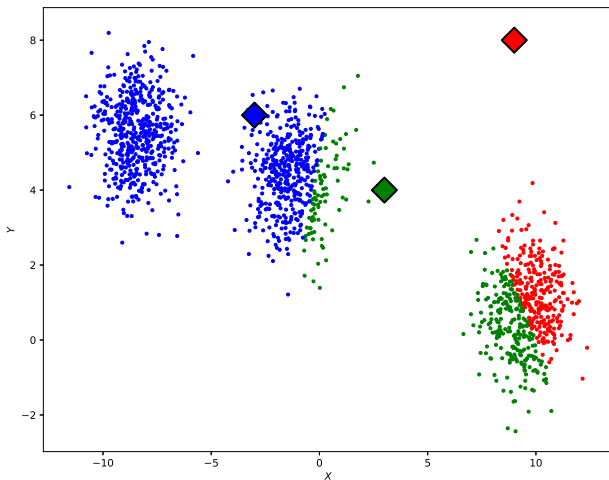
Algoritmo

- 2 Posicionar centros uniformemente al azar.



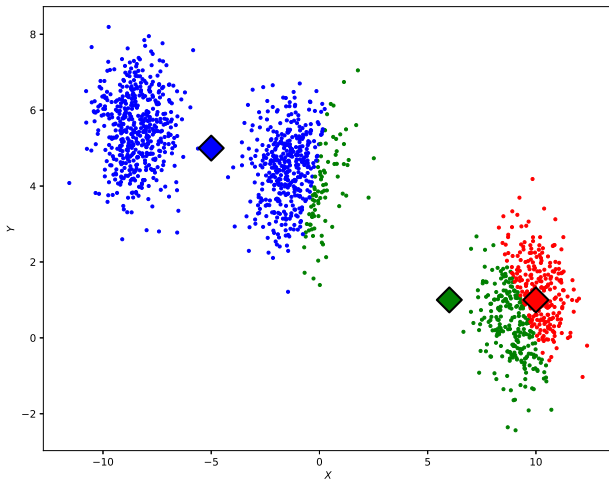
Algoritmo

- 3 Asociar cada punto con el centroide C_j más cercano.



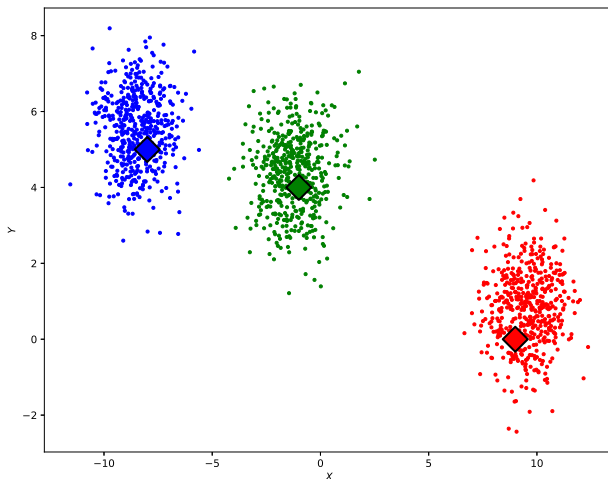
Algoritmo

- 4 Mover cada centroide C_j a la media μ_j de cada grupo.



Algoritmo

- 5 Medir $J(C)$ y si es menor que el anterior, repetir desde el paso 4 [4, 18].



Error y método del codo

Recurso: kneedle [37]

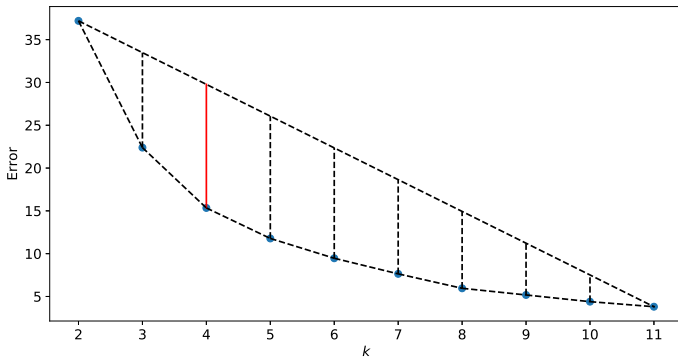


Figura: Selección de la cantidad de centros.

1 Introducción

2 Antecedentes

3 Metodología

- Obtención de los datos
- Series de tiempo
- Caracterización
- Agrupamiento

4 Resultados

5 Conclusiones

Frecuencia y estadística descriptiva

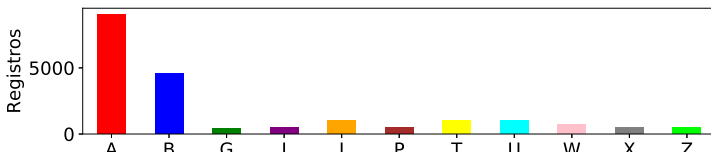


Figura: Registros por CIE general.

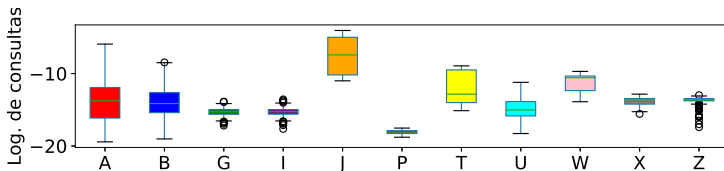


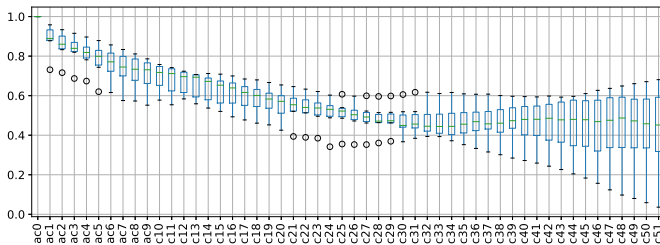
Figura: Logaritmo de número de consultas por CIE general.

Figura: Matriz de correlación de características.

Figura: PCA de dos componentes principales.

- Correlación significativa durante el primer bimestre
- Frecuencia anual
- Tras ajuste, se puede pronosticar por modelos autorregresivos
- Contiene: Infecciosas intestinales, Fiebre tifoidea, Intoxicación alimentaria bacteriana, Tuberculosis respiratoria, Brucelosis, Escarlatina, Hepatitis vírica A, Hepatitis vírica B, Paludismo por *p. vivax*, Fiebre reumática aguda, Infecciones respiratorias agudas, Infección asintomática por VIH

Grupo 1



- Correlaciones significativas durante el periodo
- Frecuencia anual y semestral
- Difícil de pronosticar
- Contiene: Paratifoidea, Shigelosis, Rubeola, Parotiditis infecciosa, Mordeduras por perro, Mordeduras por otros mamíferos

- Correlaciones significativas para el primer y último bimestre y semestre
- Frecuencia anual inversa a la semestral
- Contiene: Dengue clásico, Varicela, Neumonías y bronconeumonías, Intoxicación por plaguicidas, Intoxicación por veneno de escorpión, mordeduras por serpiente

- Serie estacionaria fácilmente pronosticable
- Se puede pronosticar por modelos autorregresivos
- Contiene: Cólera, Meningitis tuberculosa, Leptospirosis, Tétanos neonatal, Tétanos, Tos ferina, Meningitis meningocócica, Sífilis congénita, Sífilis adquirida, Infección gonocócica genitourinaria, Hepatitis vírica C, Sida, Paludismo por *p. falciparum*, Meningitis, Rubeola congénita, Enfermedad febril exantemática, Parálisis flácida aguda

1 Introducción

2 Antecedentes

3 Metodología

- Obtención de los datos
- Series de tiempo
- Caracterización
- Agrupamiento

4 Resultados

5 Conclusiones

Conclusiones

- Normalización por derechohabencia elimina el crecimiento poblacional de los casos registrados de las enfermedades
- Tres enfermedades con tendencia positiva tras la normalización: la enfermedad asintomática del VIH, la tos ferina y el cólera
- Las autocorrelaciones como características de k -medias permiten agrupar series de tiempo con base en sus características temporales
- Las enfermedades contienen grupos estacionales y estacionarios

Contribuciones

- Metodología de preprocesamiento
- Limpieza por `awk` y expresiones regulares
- Definición de columnas por pixeles para extraer datos de cuadros
- Autocorrelaciones como características relevantes
- Agrupamiento permite visualización general de comportamientos temporales para que especialistas analicen sus relaciones y una intuición de patrones de comportamiento temporal

Trabajo a futuro

- Comparación con otros algoritmos de agrupamiento
- Usar otras características y medidas de distancias
- Comparar con registros diarios de consultas
- Usar grupos resultantes para mejorar algoritmos de clasificación
- Comprobar si los síntomas de las enfermedades contenidas en los grupos guardan entre sí relaciones estadísticamente significativas
- Conocer relación entre número de consultas respecto a calidad del aire

Gracias por su atención.



jose.benavidesvz@uanl.edu.mx

- A: Enfermedades infecciosas y parasitarias

- Cólera
- Enfermedades infecciosas intestinales
- Fiebre tifoidea
- Paratifoidea
- Fiebre paratifoidea
- Shigelosis
- Intoxicación alimentaria bacteriana
- Tuberculosis respiratoria
- Meningitis tuberculosa
- Brucelosis
- Leptospirosis
- Tétanos neonatal
- Tétanos
- Tos ferina
- Escarlatina
- Meningitis meningocócica
- Sífilis congénita
- Sífilis adquirida
- Infección gonocócica genitourinaria
- Dengue clásico

- B: Enfermedades infecciosas y parasitarias

- Varicela
- Rubeola

- Hepatitis vírica A
- Hepatitis vírica B
- Hepatitis vírica C
- Sida
- Parotiditis infecciosa
- Paludismo por *p. falciparum*
- Paludismo por *p. vivax*
- G: Enfermedades del sistema nervioso
 - Meningitis
- I: Enfermedades del sistema circulatorio
 - Fiebre reumática aguda
- J: Enfermedades del sistema respiratorio
 - Infecciones respiratorias agudas
 - Neumonías y bronconeumonías
- P: Enfermedades originadas en el periodo perinatal
 - Rubeola congénita
- T: Heridas y envenenamientos
 - Intoxicación por plaguicidas
 - Intoxicación por veneno de escorpión
 - Intoxicación por picadura de alacrán
- U:
 - Enfermedad febril exantemática

- Parálisis flácida aguda
- W: Causas externas de morbilidad y mortalidad
 - Mordeduras por perro
 - Mordeduras por otros mamíferos
- X: Causas externas de morbilidad y mortalidad
 - Mordeduras por serpiente
- Z: Factores que influyen en el estado de salud y el contacto con servicios de salud
 - Infección asintomática por VIH

- [1] Arias, J. R. (2006). What is an epidemiological week and why do we use them? *The Seeker*, 6(1):7.
- [2] 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.
- [3] Bagnall, A. and Janacek, G. (2005). Clustering time series with clipped data. *Machine Learning*, 58(2):151–178.
- [4] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Singapore.
- [5] Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer, Switzerland.
- [6] Chen, J. R. (2005). Making subsequence time series clustering meaningful. In *Fifth IEEE International Conference on Data Mining*.
- [7] Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, 52(4):1860–1872.
- [8] Desarrolladores de scikit-learn (2019a). 2.3.2. *k*-means.
<https://scikit-learn.org/stable/modules/clustering.html#k-means> Accedido: 2019-03-12.

- [9] Desarrolladores de scikit-learn (2019b).
 sklearn.feature_selection.variancethreshold.
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html#sklearn.feature_selection.VarianceThreshold
 Accedido: 2019-03-22.
- [10] D'Urso, P. and 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.
- [11] Ernst, J., J. Nau, G., and 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.
- [12] Ferreira, L. N. and Zhao, L. (2015). Time Series Clustering via Community Detection in Networks. *arXiv e-prints*, 1:1–23.
- [13] Free Software Foundation (2011). GawK–GNU Project–Free Software Foundation (FSF).
<https://www.gnu.org/software/gawk/gawk.html> Accedido: 02-02-2019.

- [14] Fulcher, B. D. and Jones, N. S. (2014). Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037.
- [15] Hartigan, J. A. and 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.
- [16] Instituto nacional de estadística y geografía (2018). Población. <http://www.beta.inegi.org.mx/temas/estructura/> Accedido: 2018-09-28.
- [17] Izakian, H., Pedrycz, W., and Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244.
- [18] Jain, A. (2010). Data clustering: 50 years beyond k -means. *Pattern Recognition Letters*, 31:651–666.
- [19] Kalpalis, K., Gada, D., and 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.

- [20] Keogh, E. and Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and Information Systems*, 8(2):154–177.
- [21] Keogh, E. J. and 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.
- [22] Lai, R. K., Fan, C.-Y., Huang, W.-H., and 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.
- [23] Lin, J., Vlachos, M., Keogh, E., and Gunopulos, D. (2004). Iterative incremental clustering of time series. In *Advances in Database Technology*, pages 106–122, Berlin. Springer.
- [24] NumFOCUS (2019a). pandas.series.interpolate.
<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html> Accedido: 2019-03-22.
- [25] NumFOCUS (2019b). Python data analysis library.
<https://pandas.pydata.org/> Accedido: 2019-04-07.

- [26] 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.
- [27] Paparrizos, J. and 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.
- [28] Paparrizos, J. and Gravano, L. (2017). Fast and accurate time-series clustering. *ACM Transactions Database Systems*, 42(2):8:1–8:49.
- [29] Perktold, J., Seabold, S., and Taylor, J. (2019). statsmodels.tsa.stattools.acf. <https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.acf.html> Accedido: 2019-04-07.
- [30] Phaseit Inc. and Mathieu Fenniak (2016). PyPDF2 Documentation. https://pythonhosted.org/PyPDF2/?utm_source=recordnotfound.com Accedido: 02-07-2018.
- [31] Python Software Foundation (2018a). Python 3.7.0. <https://www.python.org/downloads/release/python-370/> Accedido: 2018-08-13.

- [32] Python Software Foundation (2018b). re – Regular expression operations. <https://docs.python.org/3.7/library/re.html>
Accedido: 2018-08-13.
- [33] Python Software Foundation (2019). datetime – basic date and time types. <https://docs.python.org/3.8/library/datetime.html>
Accedido: 2019-04-07.
- [34] Ratanamahatana, C., Keogh, E., Bagnall, A. J., and 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.
- [35] Rodrigues, P. P., Gama, J., and Pedroso, J. P. (2008). Hierarchical clustering of time-series data streams. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):615–627.
- [36] Salvador, S. and Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 576–584.

- [37] Satopää, V., Albrecht, J., Irwin, D., and 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.
- [38] 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.
- [39] Singhal, A. and Seborg, D. (2002). Clustering of multivariate time-series data. In *Proceedings of the 2002 American Control Conference*, pages 273–280, Arkansas. IEEE.
- [40] The SciPy community (2019a). `scipy.signal.detrend`. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.detrend.html> Accedido: 2019-04-07.
- [41] 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.

- [42] Vlachos, M., Lin, J., Keogh, E., and 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.
- [43] w69b (2018). PDF Mergy – WebApp to merge PDF files. <https://pdfmerge.w69b.com/> Accedido: 2018-11-23.
- [44] Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3):335–364.
- [45] World Health Organization (2018). WHO — International Classification of Diseases, 11th Revision (ICD-11). <http://www.who.int/classifications/icd/en/> Accedido: 2018-09-30.
- [46] Xiong, Y. and 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.
- [47] Xiong, Y. and Yeung, D.-Y. (2004). Time series clustering with ARMA mixtures. *Pattern Recognition*, 37(8):1675–1689.

- [48] Yildiz, B., Kaiser, K., and Miksch, S. (2005). pdf2table: A method to extract table information from pdf files. In *Indian International Conference on Artificial Intelligence*.
- [49] Zhang, H., Ho, T., Zhang, Y., and Lin, S. (2006). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica (Slovenia)*, 30:305–319.
- [50] Zhang, X., Liu, J., Du, Y., and Lv, T. (2011). A novel clustering method on time series data. *Expert Systems with Applications*, 38(9):11891–11900.