Revista de Climatología

Vol. 22 (2022): 130-147 ISSN 1578-8768

Estimación de datos faltantes de temperatura combinando IDW y una serie truncada de Fourier

Zaira Carolina Martínez Vargas[1], Jorge Paredes Tavares [2]; Sergio Ivvan Valdez[2]

[1] *Centro Universitario de Ciencias Exactas e Ingenierías, Universidad de Guadalajara, Guadalajara Jal, México*

[2] *CONACYT-Centro de Investigación en Ciencias de Información Geoespacial A.C., Parque Tecnológico San Fandila, Querétaro, México.*

Autor para correspondencia: Sergio Ivvan Valdez, svaldez@centrogeo.edu.mx

(Recibido: 26-05-2022. Publicado: 01-08-2022.)

Resumen

El estado de Jalisco presenta una alta diversidad climática debido a su ubicación geográfica y a la heterogeneidad de su relieve. Los registros desde 1961 a la fecha presentan diversos problemas, siendo el más importante la ausencia de registros. Para solventar lo anterior, se propone un método de dos etapas para la estimación de datos faltantes; la primera etapa utiliza un algoritmo no supervisado, por lo que no requiere de datos históricos para hacer una primera aproximación, la segunda etapa utiliza una interpolación supervisada para corregir el error de la primera. El algoritmo no supervisado se selecciona evaluando el desempeño entre kriging ordinario (KO), nearest neighbor (NN), inverse distance weighting (IDW) y modified inverse distance weighting (MIDW). Estos métodos utilizan vecinos espaciales que pueden ser afectados por la falta de datos, por lo que se propone un algoritmo que considera variaciones en la cantidad de datos. El número óptimo de vecinos se determina por la mínima mediana del MAE. Las estimaciones de la primera etapa son significativamente mejoradas en una segunda etapa con una serie truncada de Fourier. La combinación de ambos métodos entrega un MAE estimado de 0.5273°C.

Palabras clave: *Interpolación temporal, Interpolación espacial, Series de Tiempo, Normales Climatológicas, Algoritmos no supervisados, Serie truncada de Fourier.*

Abstract

The state of Jalisco presents a high climatic diversity due to its geographical location and its varied landscape. The recorded data, from 1961 to present date, have several problems, the most important is the large amount of missing data. To circumvent this issue, a two-stage method for estimating missing data is proposed; the first uses an unsupervised algorithm; hence it does not require historical data. The second uses supervised interpolation for correcting the error from the first one. The unsupervised algorithm is selected as the best performed among ordinary kriging (KO), nearest neighbor (NN), inverse distance weighting (IDW), and modified inverse distance weighting (MIDW). These methods use spatial neighbors that can be affected by missing data; thus, we propose a methodology for considering variations in the amount of

data. The optimal number of neighbors minimizes the median of the MAE. The estimations, from the first stage, are significantly improved in the second stage, using a truncated Fourier Series. The combined method estimates missing data with a MAE of 0.5273°C.

Keywords: *Temporal interpolation, Spatial interpolation, Time series, Weather normal, Non-supervised algorithms, Truncated Fourier series.*

1. Introducción

El clima se define como el promedio de las condiciones meteorológicas que ocurren en un lugar determinado durante un largo periodo de tiempo que usualmente es de 30 años (Donald Ahrens et al., 2015; Baede et al., 2001). Las principales variables meteorológicas que determinan el clima de una región son la temperatura, precipitación, dirección, velocidad del viento y radiación solar (Shepherd et al., 2016). Contar con datos climáticos precisos y continuos a través del tiempo es fundamental para mejorar los pronósticos a mediano y largo plazo, y para analizar otros fenómenos de alto impacto como el cambio climático, lo que permite diseñar medidas para reducir sus consecuencias negativas (Hartmann et al., 2013; García et al., 2015; Eckstein et al., 2019). Uno de los principales problemas que se enfrentan en el análisis de tendencias y estacionalidades de fenómenos meteorológicos es el acceso a datos suficientes, tanto en la dimensión espacial (cobertura homogénea del territorio), como en la temporal (registros completos a lo largo del tiempo) (Fenta Mekonnen y Disse, 2018; Sørland et al., 2018). Entre las propuestas sugeridas para atender esta problemática, se encuentra la aplicación de métodos de interpolación espacial para estimar datos faltantes a partir de registros de un conjunto de estaciones meteorológicas aledañas (Attorre et al., 2007; Lepot et al., 2017). Los métodos de interpolación espacial se pueden clasificar en métodos de punto y de área, dependiendo si interpolan valores puntuales o sobre una región (Lam, 1983); en supervisados y no supervisados si requieren datos de entrenamiento o no; y en exactos o aproximados dependiendo si la función de interpolación pasa por los datos originales o es ajustada para cumplir condiciones a priori del modelo (Sluiter, 2009). Adicionalmente, estos métodos se pueden clasificar en numéricos, estadísticos y de cómputo suave, dependiendo de la naturaleza de las funciones de interpolación y las bases matemáticas para su desarrollo (Li y Heap, 2008). Entre los métodos numéricos exactos no-supervisados se encuentran: inverse distance weighting (IDW), modified Inverse distance weighting (MIDW), nearest neighbor (NN) y splines. En los métodos aproximados más utilizados están kriging, thin plate smoothing spline (TPSS), series de Fourier y de Potencia con términos truncados para eliminar las frecuencias más altas, entre otros (Li y Heap, 2008). Estos métodos pueden ser aplicados en datos de series de tiempo, en mediciones sobre el espacio, o combinando datos espaciales y temporales.

En México existen diversos estudios que interpolan datos puntuales de estaciones meteorológicas (Boer et al., 2001; Cuervo-Robayo et al., 2014; Cuervo-Robayo et al., 2020; Fernández Eguiarte et al., 2014; Ruiz-Corral et al., 2018; Sáenz-Romero et al., 2010; Téllez et al., 2011; Zhu y Lettenmaier, 2007), también existen coberturas a nivel mundial realizadas mediante interpolaciones que incluyen el territorio nacional (Fick y Hijmans, 2017; Hijmans et al., 2005). Los métodos que se han utilizado para este fin son kriging, IDW^1, IDW^2 y TPSS, siendo este último el más ampliamente utilizado a través del software ANUSPLIN (Hutchinson, 2004). Dichas investigaciones se caracterizan por trabajar con métodos supervisados, ya que requieren registros históricos para hacer estimaciones. A diferencia de la tendencia a nivel mundial (Hartmann et al., 2013; Eckstein et al., 2019), las investigaciones en México enfocadas en comparar diferentes métodos de interpolación resultan escasas. Entre las encontradas en la literatura está el traba-

jo de Fernández Eguiarte et al. (2014), quienes concluyen que los métodos spline e IDW¹ son sensiblemente más efectivos que kriging e IDW²; en cambio, los resultados del trabajo Boer et al. (2001) indican que la regresión trivariada de kriging es ligeramente más precisa que thin plate spline, aunque su implementación requiere mayor experiencia y recursos computacionales; mientras que para Díaz et al. (2008), el método TPSS ofreció mejores resultados al interpolar datos de precipitación en las zonas de barlovento y sotavento del Golfo de México, seguido por los métodos kriging, co-kriging e IDW. En este sentido, la presente investigación estima valores faltantes de temperatura media mensual en el estado de Jalisco con datos publicados por la Comisión Nacional del Agua (CONAGUA, 2020) aplicando cuatro métodos de interpolación sobre la dimensión espacial: kriging ordinario, IDW, MIDW y NN, y una serie de tiempo truncada sobre la dimensión temporal para mejorar la aproximación inicial. La interpolación espacial se puede aplicar como etapa única cuando no se dispone de datos históricos, o se puede hacer una estimación de mayor precisión si se dispone de los mismos. Los resultados obtenidos se compararon utilizando el error absoluto medio (MAE por las siglas en inglés de Mean Absolute Error). Las contribuciones son la selección del mejor método de interpolación no supervisado para la zona de estudio, la determinación del número de vecinos óptimo para dicho método, y la mejora de los resultados con la serie de tiempo de Fourier, donde también se determinó el número adecuado de términos de la serie. El método propuesto se validó con 29 estaciones que tienen datos completos de temperatura media mensual usando validación cruzada.

2. Metodología

2.1 Área de estudio

El estado de Jalisco se ubica en la región centro-occidente de la República Mexicana, sus coordenadas extremas son: 18°58'00z 22°45'00"de latitud norte y 101 °28'15z 105°43'16"de longitud oeste (INEGI, 1988), en esta zona convergen las provincias fisiográficas Sierra Madre Occidental, Eje Neovolcánico, Sierra Madre del Sur, y el Altiplano Mexicano, lo que le otorga a la entidad un relieve complejo formado por planicies costeras, zonas montañosas separadas por valles de vertientes abruptas y secuencias de valles y montañas de origen volcánico. Esta diversidad en el relieve se manifiesta en diferencias altitudinales que van desde los cero hasta los 4260 m.s.n.m (INEGI, 1988), lo que, aunado a las barreras orográficas, la latitud y la continentalidad, ha definido para Jalisco tres grupos de climas de acuerdo con la clasificación de Köppen modificada por García (1981): cálido, seco y templado, que se dividen en 25 subgrupos (Cruz Angón et al., 2017). Esta diversidad geográfica genera condiciones de complejidad durante la estimación de datos de temperatura, debido a que, por un lado, esta variable está influenciada por la altitud y la orografía (Masson y Frei, 2014), y por el otro, algunos métodos tienen un mejor desempeño al aplicarse en ciertas unidades de paisaje..

2.2 Métodos de interpolación espacial

La interpolación tiene como objetivo estimar los valores $\hat{z}$ a partir de una muestra z_i tomadas en la locación i dentro de una vecindad espacial. La fórmula general de interpolación es la siguiente:

$$\hat{z} = \sum_{i=1}^{n} w_i z_i \tag{1}$$

donde wi es el peso para cada interpolador y el valor de n depende de cada método y la cantidad

de datos disponibles.

2.2.1 Interpolador nearest neighbor (NN)

Este método se basa en construir bisectrices perpendiculares para cada punto conocido (Li y Heap, 2008). Se construyen estas bisectrices para formar polígonos de Voronoi (Vi). Para la interpolación de un punto objetivo se localiza el polígono al que pertenece y se le asigna el valor del punto que generó el polígono. De esta manera, el peso se describe por:

$$w_i = 1siz_i\varepsilon V_i, 0 \tag{2}$$

Se utilizó el método nearest neighbor interpolator de Python (Virtanen et al., 2020) y se trataron los datos para remover los valores faltantes en cada interpolación.

2.2.2 Inverse squared distance weighting (IDW)

En este método el valor en la estación de interés está dado por una combinación lineal de los pesos de las estaciones cercanas (Boke, 2017), por lo que, mientras más cerca esté una estación vecina a la estación objetivo, su peso será mayor. El peso está dado por:

$$\frac{\frac{1}{d_i^p}}{\sum_{i=1}^{n}\frac{1}{d_i^p}} \tag{3}$$

donde n es el total de vecinos por estación, di es la distancia de la estación de interés con sus vecinos y p es un parámetro del interpolador que al ser inverso de la distancia cuadrado es igual a 2. Para la distancia se utilizó la fórmula del semiverseno que representa la distancia más corta en una superficie esférica, la cual utiliza los parámetros de longitud y latitud.

2.2.3 Modified inverse distance weighting (MIDW)

Este método permite incluir el efecto de la diferencia de elevación entre estaciones vecinas y la estación objetivo, por lo que, habrá más influencia de una estación vecina si la diferencia de altura es menor (Boke, 2017). El peso está dado por:

$$w_i = \frac{\left(\frac{d_i}{\Delta H_i+\varepsilon}\right)^p}{\sum_{i=1}^{n}\left(\frac{d_i}{\Delta H_i+\varepsilon}\right)^p} \tag{4}$$

donde d_i es la distancia entre la estación de interés y sus vecinos, ΔH_i es la diferencia de altura entre estos, n es la cantidad de vecinos, ε es un término igual a 0.0001 que se añadió para evitar una división entre cero y p es el parámetro del interpolador que en este caso fue 2. Los vecinos para cada estación se encontraron con la distancia mínima.

2.2.4 Kriging ordinario (KO)

Los métodos de kriging se consideran de tipo estimador lineal insesgado óptimo; insesgado porque el error promedio es 0 y óptimo porque su varianza es mínima. En el caso de kriging ordinario, y en contraste con el kriging simple, considera que las medias locales no son necesariamente próximas a las medias poblacionales. De forma general, el valor interpolado se estima mediante la combinación lineal de valores cercanos, donde los pesos y el número de valores utilizados dependen de la correlación entre los valores existentes (Lam, 1994). Se utilizó la implementación PyKrige de Python (Murphy, 2014) y el modelo de variograma gaussiano al ser uno de los más comunes.

$$z\hat{=}\sum_{i=1}^{n}\lambda_i Z_i \tag{5}$$

donde, λ_i es una ponderación para la locación i que considera la distancia espacial, un ajuste debido a la correlación y un modelo de variograma.

2.2.5 Interpolación temporal por series de tiempo

Cuando se disponen de datos históricos, como datos de series de tiempo con coordenadas $[t_i, T_i]$, para una fecha ti, con una temperatura Ti, en una estación i, estos se pueden modelar como una función de tiempo, dado que estos datos tienen tendencias estacionales, es razonable que esa función del tiempo sea una serie periódica (Zahroh et al., 2019; Shea et al., 1992). Utilizamos una versión truncada de la serie de Fourier en la Ecuación 6.

$$f(t) = \frac{a_0}{2} + \sum_{i=1}^{n}(a_n cos(2n\pi t) + b_n sin(2n\pi t)) \tag{6}$$

donde a_0, a_n y b_n son los coeficientes de Fourier, n es el número de armónicos en la serie y N es el término donde se trunca la serie.

2.2.6 Cálculo de los errores de interpolación

Los errores de interpolación y predicción son generalmente medidos usando errores cuadráticos medios o sumas de errores al cuadrado. Como criterio del error de una estación se utiliza el Error Absoluto Medio (MAE), como se muestra en la Ecuación 7.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{z_i} - z_i| \tag{7}$$

donde n es la cantidad de valores en la muestra, $\hat{z_i}$ es el valor interpolado y z_i es el valor real. Note que tanto la media como la suma de los errores cuadráticos pueden ser fuertemente sesgados por datos atípicos debidos a errores de registro o medición, por los que investigadores de la literatura especializada (Armstrong y Collopy, 1992) han sugerido que estas medidas no son adecuadas en todos los casos y que el uso de la mediana u otras medidas pueden ser más apropiadas. En este caso, utilizamos la mediana como criterio del error representativo de todas

las estaciones debido a la gran cantidad de datos faltantes, para así tener una medida robusta del desempeño de los métodos. Finalmente, para tener una medida del error del método de dos etapas, se utiliza validación cruzada. Esta técnica divide la cantidad de datos entre k grupos y los separa entre datos de entrenamiento y datos de validación, de manera que los datos de entrenamiento son $k-1$ y el grupo restante es el de validación, el modelo se entrena k veces, y se obtienen k medidas de desempeño con los datos de validación, por último, se reportan métricas que resumen estas k ejecuciones, para reducir el sesgo y reportar medidas de desempeño del modelo sobre todos los datos posibles.

2.3 Recopilación y procesamiento de datos

Se compiló información publicada por el Servicio Meteorológico Nacional (SMN) para el estado de Jalisco y sus estados colindantes (Nayarit, Zacatecas, Colima, San Luis Potosí, Aguascalientes, Guanajuato y Michoacán). Del "Proyecto de bases de datos climatológicos" se utilizaron los siguientes datos: nombre de estación, número de estación, latitud, longitud, altitud y nombre de la estación. Por otra parte, se consultó el proyecto "Normales climatológicas", el cual contiene los siguientes datos: días con granizo, días con niebla, días con tormenta, evaporación mensual, lluvia máxima 24 horas, lluvia total mensual, temperatura máxima extrema, temperatura máxima promedio, temperatura media mensual, temperatura mínima extrema y temperatura mínima promedio. En esta investigación sólo se utilizó la información de la temperatura media mensual ya que mostró una correlación alta entre estaciones vecinas. Posteriormente, se realizó una limpieza previa a los datos y se convirtieron a un formato tabular considerando la información del año, mes, número de estación y el valor de temperatura para el periodo que comprendió los años de 2002 al 2017. La elección del periodo de tiempo se basó en la disponibilidad de datos, considerando que los registros comienzan en años distintos, y que a partir del año 2002 la mayoría de las estaciones contienen datos. Se buscó un intervalo de tiempo comparando períodos de cinco años, siendo el periodo 2002-2006 el que mejor cumple esta condición.

2.4 Método en dos etapas

Se propuso un método de dos etapas para la estimación de los datos faltantes (Fig. 1). En la primera etapa no se utilizan datos históricos, es decir, basta con los datos espaciales en la misma coordenada de tiempo para estimar el dato faltante; en la segunda etapa, se usan los datos de los 5 años para mejorar la estimación espacial por medio de un ajuste temporal. El propósito del método es doble: por un lado, se estima la temperatura con el método de interpolación espacial seleccionado para tener la menor mediana del MAE; posteriormente, se utilizan datos históricos para la corrección del error de la primera etapa usando la serie de tiempo truncada de Fourier.

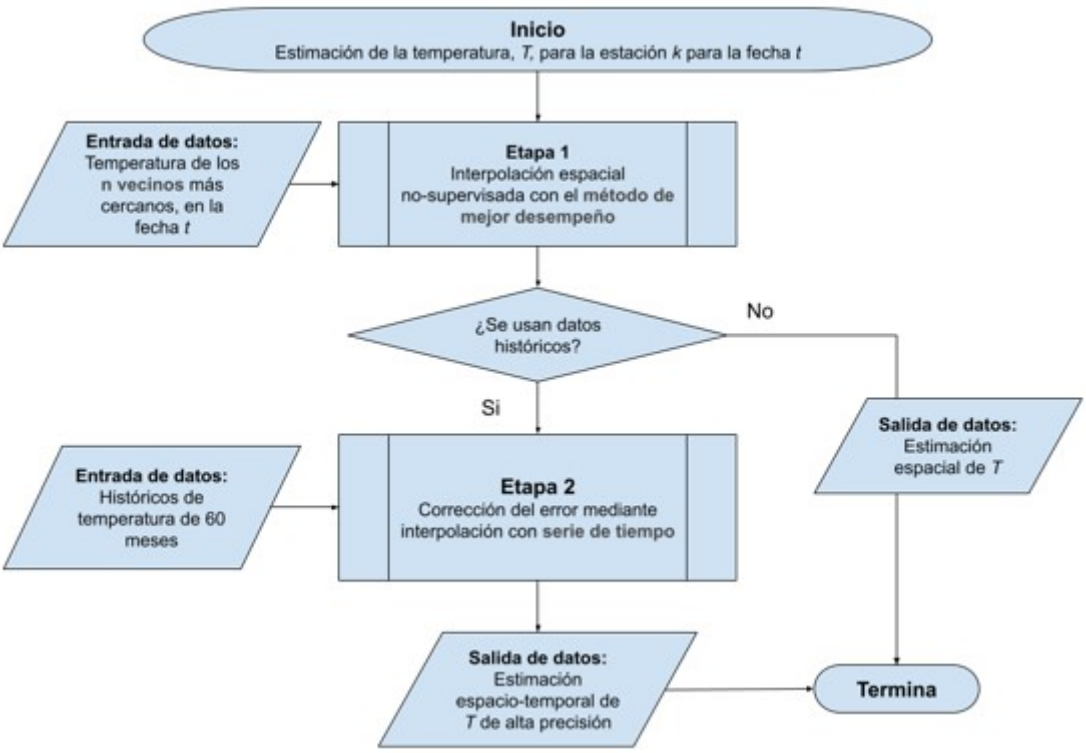

Figura 1: Diagrama de flujo para la estimación de un valor de temperatura faltante. las fuentes en azul indican propuestas de este trabajo.

Fuente: Elaboración propia.

2.5 Primera etapa: selección del método y criterios para la interpolación espacial

Para determinar el método de interpolación espacial no supervisado que arroja mejores resultados, se hace una comparación de la siguiente forma: primero se calcula el MAE de estimación de las temperaturas para los 60 meses del periodo de estudio para cada una de las 29 estaciones con datos completos; posteriormente, se calculan medidas estadísticas de tendencia central y de dispersión (promedio, desviación estándar, mínimo, mediana y máximo). En el caso de los métodos IDW y MIDW, se ajustan modelos para $n = 2$ hasta $n = 50$ vecinos (incluyendo estaciones de los estados colindantes). La búsqueda de las estaciones más cercanas se realiza usando la distancia Haversine. Para solventar el problema de falta de datos entre los vecinos para el periodo estudiado de 60 meses se propone lo siguiente: 1) Para la temperatura i-ésima en el mes j-ésimo, T_{ij}, se remueven los vecinos con datos faltantes, y 2) se aplica el método de interpolación. Por lo tanto, el método usa $N - M_{ij}$ vecinos, donde N es un número óptimo fijo de vecinos, mientras que M_{ij} es el número de datos faltantes, el cual, en general, es diferente para cada estación y mes. Para los métodos NN y KO se utilizan todas las estaciones de Jalisco y de los estados colindantes, se depuran los datos para cada mes interpolado, eliminando las estaciones con datos faltantes.

2.6 Segunda etapa: Corrección del error con serie de tiempo

Para esta etapa se calcula la diferencia entre los valores reales y los estimados con el método de interpolación espacial de menor mediana del MAE. A los errores resultantes se les ajusta el

modelo de la Ecuación 5, utilizando el módulo Statsmodels de Python (Seabold y Perktold, 2010).

3. Resultados

Para validar la metodología propuesta se presentan los resultados en el siguiente orden: 1) Se determinó el número óptimo de vecinos para hacer la interpolación espacial de menor error con el IDW y MIDW. 2) Posteriormente se compararon los 4 métodos de interpolación espacial: IDW, MIDW, con el número de vecinos óptimo, NN y KO, para seleccionar el de menor error. 3) Usando el método de menor error para la etapa de interpolación espacial, se aplica la serie de tiempo truncada para compensar los errores con información histórica, y mejorar la interpolación. Aquí se muestran las evidencias estadísticas del desempeño de la propuesta general de 2 etapas, mediante dos técnicas de estimación del error. 4) Finalmente, se analizan casos de estudio de estaciones particulares, para mostrar el efecto de cada etapa de la interpolación sobre los datos puntuales de estación y mes, para mostrar el efecto de cada etapa por separado.

3.1 Número óptimo de vecinos para los métodos IDW y MIDW

Para determinar este valor se ejecutaron ambos métodos usando desde 2 hasta 50 estaciones más cercanas, las 5 mejores ejecuciones se resumen en las Tablas 1 y 2. Para el método IDW, la menor mediana del MAE se consigue con 25 vecinos, mientras que para el método MIDW este parámetro alcanza con 26.

Tabla 1: Concentrado de los cinco mejores resultados para diferente número de vecinos del error absoluto medio del método IDW.

Vecinos	Promedio	Desviación Estándar	Mínimo	Mediana	Máximo
25	1.7337	1.4929	0.3685	1.0490	6.4265
15	1.7926	1.5293	0.3581	1.0809	6.8183
16	1.7982	1.5517	0.3581	1.0809	6.9242
24	1.7344	1.4852	0.3645	1.0906	6.4265
23	1.7377	1.4946	0.3609	1.1011	6.5064

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

Tabla 2: Concentrado de los cinco mejores resultados para diferente número de vecinos del error absoluto medio del método MIDW.

Vecinos	Promedio	Desviación Estándar	Mínimo	Mediana	Máximo
26	2.0147	1.9280	0.6220	1.0848	9.1178
25	2.0222	1.9506	0.6222	1.0923	9.3154
24	2.0296	1.9950	0.6152	1.1052	9.5866
18	1.9161	1.6823	0.5619	1.1618	6.5112
19	1.9956	1.7035	0.5619	1.1618	6.5112

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

3.2 Desempeño los métodos de interpolación espacial

Las medidas estadísticas de tendencia central y de dispersión de los métodos se muestran en la Tabla 3. Se toma como criterio de selección el valor de mediana más bajo, debido a que sabemos que 50 % de las estaciones tienen menor error que este valor, no es sensible a datos atípicos, y siempre es un valor existente o representativo de los errores.

Tabla 3: Comparación de los métodos de interpolación espacial, el de menor mediana se considera el de mejor desempeño.

Método	No Vecinos	Promedio	Desviación Estándar	Mínimo	Mediana	Máximo
IDW	25	1.7337	1.4929	0.3685	1.0490	6.4265
MIDW	26	2.0147	1.9280	0.6220	1.0848	9.1178
KO	-	2.1768	0.7849	0.7751	2.0915	4.5132
NN	-	2.0755	1.0548	0.4246	1.9582	5.4762

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

Los resultados muestran que el método con la mínima mediana del MAE es IDW, seguido por los métodos MIDW, NN y KO; por lo que el primero es seleccionado para la segunda etapa, en la cual se le calcula la serie de tiempo truncada de Fourier para mejorar la estimación usando los históricos temporales.

3.3 Estimación de la temperatura y errores de la interpolación espacio-temporal

El método con la mínima mediana del error absoluto medio fue el IDW con 25 vecinos, por ello, sus resultados se utilizan para ajustar la serie tiempo, con el fin de mejorar la estimación.

Se encontró que truncar la serie de la Ecuación 5 en $N = 7$ armónicos, muestra los mejores resultados junto con una normalización del periodo de 1/36, es decir, t de la Ecuación 5 está en el intervalo [1/36,60/36]. Con el fin de tener una perspectiva del desempeño del interpolador sobre todos los datos, se aplicó una validación cruzada con k=10 sobre las 29 estaciones. Es decir, se aleatorizan los índices de los datos, se realizan 10 particiones, se estiman las temperaturas de la primera partición usando las 9 restantes para estimar los coeficientes de la serie de tiempo, se calcula el MAE de la primera partición, se repite el procedimiento para la segunda partición usando el resto para ajustar la serie, y así sucesivamente. Con lo anterior se obtienen 10 valores de MAE para cada estación, que permiten analizar el error en centralidad y dispersión con los boxplots de las Figuras 2a y 2b. Observe que la mayoría de las estaciones tienen una mediana del MAE menor a 0.5, con distancia intercuartil menor a 0.2°C, lo que habla de un buen desempeño no solo en promedio, sino que es robusto por la baja dispersión que presenta. Solo las estaciones 14195 y 14324 presentan alto error y dispersión, que se puede adjudicar a características especiales a las condiciones geográficas de estas estaciones o de sus datos. Aún en estos casos los máximos MAE no superan los 1.5°C, y la mediana es de alrededor de 1°C. Sin embargo, se observa una reducción del MAE promedio de todas las estaciones de 1.2064°C al aplicar la serie de Fourier. Este es el desempeño esperado para casos reales, en los cuales no se puede medir el error. Note que la única suposición de esta estimación es en el uso del 90 % de los datos temporales (las 9 particiones de entrenamiento) de una estación, pero no se hace ninguna suposición sobre los datos espaciales, ya que se usan las estaciones cercanas con sus datos faltantes como se presentan en la realidad, y se utilizan para este análisis las 29 estaciones con datos completos del periodo, ya que de otra forma sería imposible conocer el error de estimación.

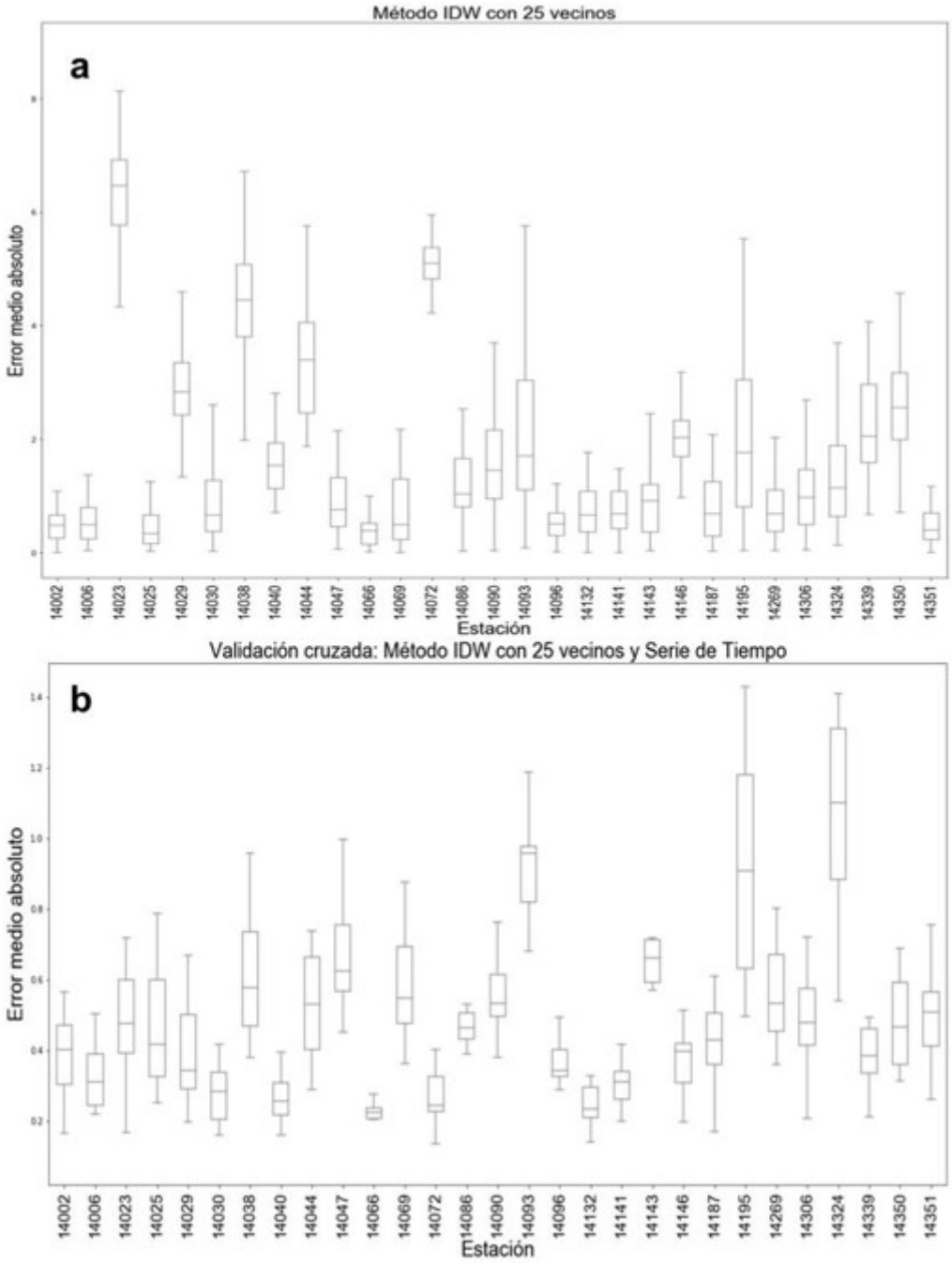

Figura 2: Box Plot resultado de la validación cruzada, cada estación tuvo diez resultados ($k = 10$). a) MAE de la validación cruzada con 10 particiones de la interpolación espacial del IDW con 25 vecinos. b) mae de la validación cruzada con 10 particiones de la interpolación espacio-temporal (IDW + serie de Fourier).

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

Con el fin de obtener una estimación realista del error se propone el siguiente procedimiento:

1) Se calculó el porcentaje de datos faltantes en todas las estaciones de Jalisco para el periodo de estudio, que tiene un valor redondeado de 49 %. 2) Se tomaron las 29 estaciones de las cuales se tienen datos completos para el mismo periodo. 3) Se removió el 49 % de los datos de

las 29 estaciones de manera aleatoria, para simular la falta de datos. 4) Se aplicó el método de dos etapas propuesto y se calcularon estadísticas. Las estadísticas del desempeño del método, bajo las condiciones reales de los datos reportados para todo el estado de Jalisco, muestran que el promedio (0.5273) y la mediana (0.4447) son altamente competitivos con otros métodos reportados en la literatura especializada, los cuales son más complejos, menos intuitivos y más costosos computacionalmente (Fick y Hijmans (2017). El MAE mínimo fue de 0.1985, el máximo de 1.6843 y la desviación estándar fue de 0.2982.

3.4 Resultados de la interpolación espacio-temporal en estaciones específicas

La Figura 3a muestra que, aunque el IDW logra reproducir el patrón real de la temperatura, hay una sobreestimación sistemática, que indica que esta estación usualmente presenta una temperatura menor que la de sus vecinas más cercanas. Este error recurrente se corrige con la serie truncada Fourier, observe en la Figura 3b que los triángulos de la estimación corregida ya están a la misma altura que los valores reales, y el error ha disminuido considerablemente. Estas correcciones no solo se dan en estaciones con estas temperaturas evidentemente trasladadas, como se aprecia en la estación 14351 (Fig. 4a), donde la estimación con IDW es muy cercana a los valores reales. Note en la Figura 4b, como aún en este caso los triángulos están más cerca de los valores reales y logran mejorar la estimación en la mayoría de los casos.

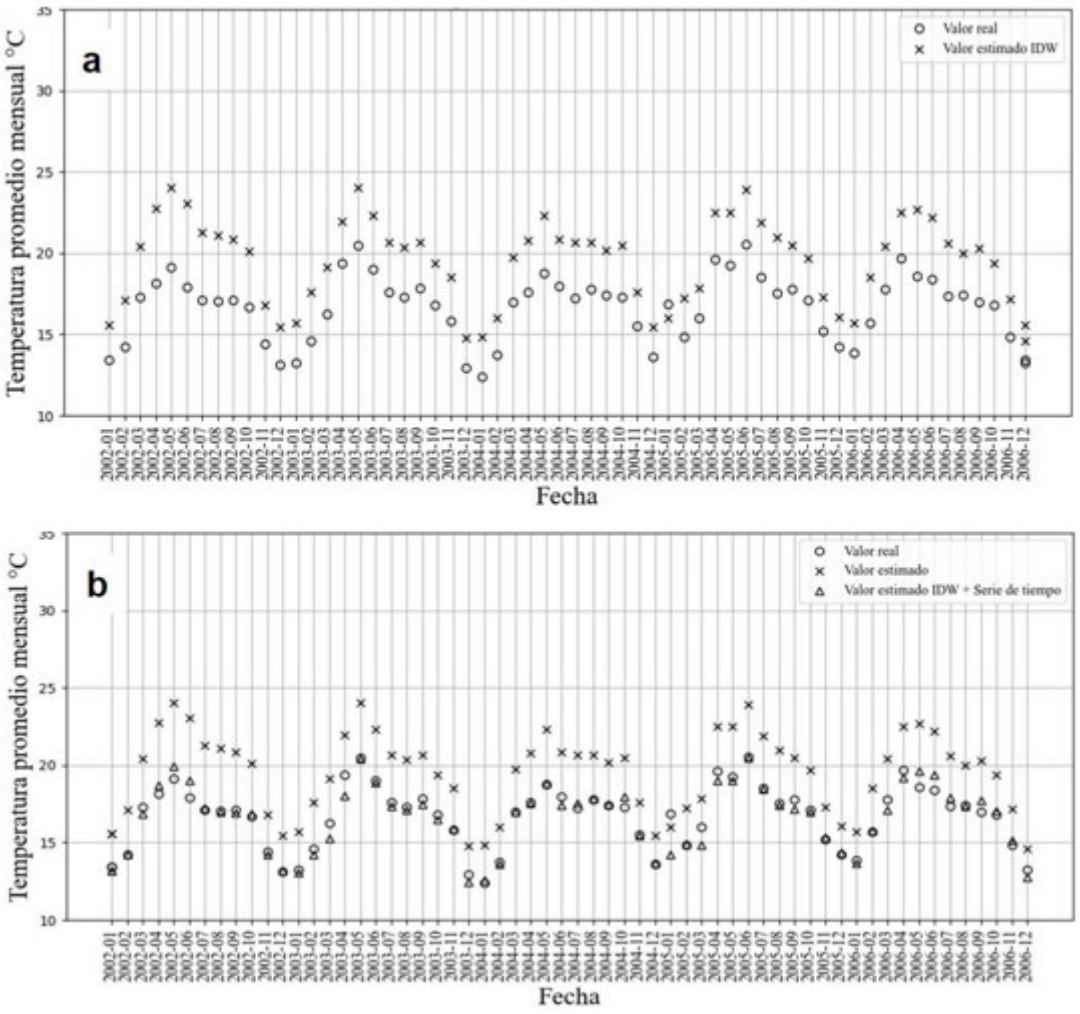

Figura 3: a) Estimación con interpolación espacial IDW para la estación 14029 con 25 vecinos. b) Resultado de la interpolación IDW más la serie truncada de Fourier.

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

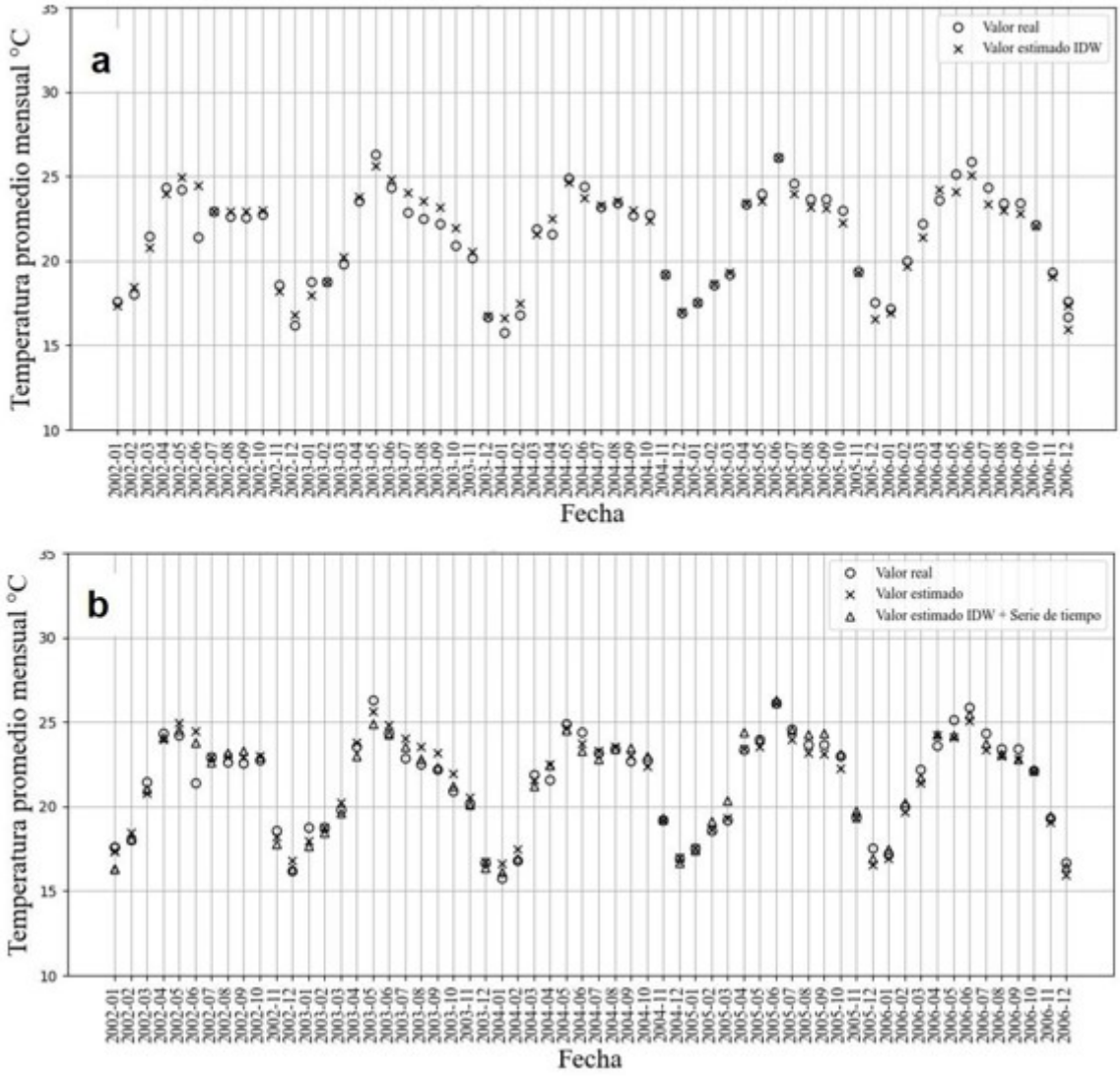

Figura 4: a) Estimación con interpolación espacial IDW para la estación 14351 con 25 vecinos. b) Resultado de la interpolación IDW más la serie truncada de Fourier.

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

3.5 Discusión

Dos de las ventajas de los métodos IDW y MIDW son su facilidad de implementación y de modificación. Las Tablas 1 y 2 muestran que los resultados consecutivos del número de vecinos, 15-16 (para el método IDW) y 18-19 (para el método MIDW), son muy similares. En el primer caso, tanto la mediana como el mínimo son iguales, pero hay diferencias en el promedio y desviación estándar, una posible razón del aumento de promedio y de máximo es la existencia de un valor atípico cuando se añade otro vecino. De manera similar, en el segundo caso, los resultados son iguales excepto en la desviación estándar y el promedio, que aumenta ligeramente cuando se añade otro vecino. Se realizaron pruebas de hipótesis, pero no se encontraron diferencias estadísticas entre un número de vecinos y otro, por lo cual se decidió tomar como criterio la menor mediana. En la Figura 2 se reportan los resultados de la validación cruzada para el ensamble de la serie de tiempo y el método IDW; se muestra que tanto la estación 14324 y la estación 14195 tienen un mayor error. La Figura 5 muestra la estación 14324 y sus vecinos más cercanos; de

estos, la estación más próxima es la 14164, sin embargo, no tiene registros en el periodo 2002-2006 (60 meses), la segunda estación más cercana (14153) tiene 54 registros de 60, la estación 14026 tiene 41, la estación 14331 tiene 22 (menos del 50 % de los datos), y la siguiente estación (32124) se encuentra a 22.25 km de distancia. Es decir, el error podría ser explicado por falta de datos y lejanía de estaciones cercanas; otra posible explicación es que las estaciones usadas para un subperiodo no son las mismas que para otro subperiodo, lo que hace difícil de corregir los errores con la serie temporal, sin embargo, al analizar otras estaciones que presentan estas condiciones no se observan errores tan elevados. Estos casos particulares requieren un mayor análisis (que está fuera del alcance de esta investigación) para determinar las causas del error, incluyendo las condiciones geográficas como topografía, la orientación de laderas, las zonas de barlovento y sotavento, las características del entorno (rural o urbano), la cercanía a cuerpos de agua, las condiciones de instalación de la estación etc., aunado a los problemas de la falta de datos y cercanía de las estaciones.

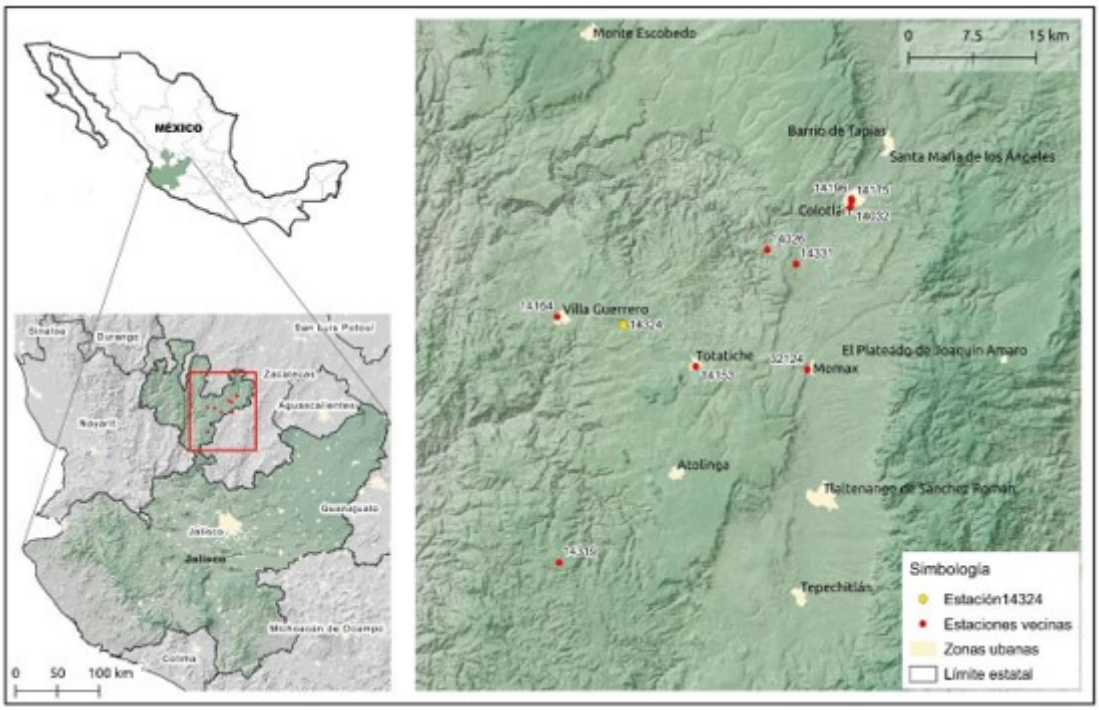

Figura 5: Ubicación de la estación 14324 y las estaciones vecinas usadas en la interpolación.

Fuente: Elaboración propia a partir de datos de CONAGUA 2020.

4. Conclusiones

En este artículo se introduce un método de dos etapas para la estimación de datos faltantes de temperatura. Las ventajas del método son varias: una estimación con muy bajo error (MAE), menor a 0.6°C en la mayoría de las estaciones, un método computacionalmente simple, por su baja complejidad computacional, y robusto, dado que las interpolaciones se realizan en estaciones que tienen en su conjunto 49 % de datos faltantes. El error es competitivo aún con métodos altamente complejos de software comercial ampliamente usado en la literatura especializada (Hutchinson, M. F. ANUSPLIN Version 4.3). Aunque el método de kriging es muy usado en la literatura, este requiere mayor análisis o información, ya que, al ser geoestadístico, se tiene que conocer la zona de estudio para seleccionar el modelo de variograma adecuado, lo cual no es

necesario para los métodos NN, IDW y MIDW. En el caso de las estaciones con mayor error, que los datos de temperatura por sí solos difícilmente explican los errores de estimación, que muy probablemente requieren de conocimiento de campo y argumentos geográficos más complejos o que tomen en cuenta más variables. Finalmente, consideramos que la investigación futura puede dirigirse hacia la generación de otros métodos simples, de baja complejidad computacional usando interpolación espacio-temporal. En la parte temporal, la serie de truncada de Fourier entrega excelentes resultados, porque es la que permite reducir el MAE promedio de todas las estaciones de 0.5273°C, sin embargo, requiere de los datos históricos de un periodo considerable de tiempo, en otras palabras, se combinan un método espacial no-supervisado con un método temporal supervisado. La investigación futura podría centrarse en un método diferente a los considerados en este estudio que realice una mejor estimación espacial, combinado con un componente temporal que requiera una menor cantidad de datos de entrenamiento para la estimación temporal.

5. Reconocimientos

S. Ivvan Valdez es financiado por la Cátedra CONACYT 7795. Jorge Paredes-Tavares es financiado por la Cátedra CONACYT 7308.

Referencias bibliográficas

Armstrong JS, Collopy F (992): Error measures for generalizing about forecasting methods: Empirical comparisons. *Int. journal forecasting.*, (8)1:69-80.

Attorre F, Alfo' M, De Sanctis M, Francesconi F, Bruno F (2007): Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. *Int. J. Climatol.*, (27)13:1825-1843.

Baede A, Ahlonsou E, Ding Y, Schimel D. (2001): The climate system: An overview in climate change 2001: Impacts, adaptation and vulnerability. Cambridge University Press.

Boer EP, De Beurs KM, Hartkamp AD (2001): Kriging and thin plate splines for mapping climate variables. *Int. J. Appl. Earth Obs. Geoinformation.*, (3)2:146-154.

Boke AS (2017): Comparative evaluation of spatial interpolation methods for estimation of missing meteorological variables over Ethiopia. *J. Water Resource Prot.*, (9)8:945-959.

Cruz Angón A, Melgarejo ED, Ordorica Hermosillo, Valero Padilla J (2017): La biodiversidad en Jalisco estudio de estado. México: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. CONABIO. Recuperado de: https://www.biodiversitylibrary.org/bibliography/169260

Cuervo-Robayo AP, Téllez-Valdés O, Gómez-Albores MA, Venegas-Barrera CS, Manjarrez J, Martínez-Meyer E (2014): An update of high-resolution monthly climate surfaces for mexico. *Int. J. Climatol.*, (34)7:2427-2437.

Díaz Padilla G, Sánchez Cohen I, Quiroz R, Garatuza Payán J, Watts Thorp C, Cruz Medina

IR (2008): Interpolación espacial de la precipitación pluvial en la zona de barlovento y sotavento del golfo de México. *Agricultura técnica en México.*,(34)3:279-287.

Donald Ahrens C, Henson R (2015): Meteorology today: an introduction to weather, climate and the environment.

Fenta Mekonnen D, Disse M (2018): Analyzing the future climate change of upper blue Nile river basin using statistical downscaling techniques. *Hydrol. Earth System Sci.*, (22)4:2391-2408.

Fernández Eguiarte A, Romero Centeno R, Zavala Hidalgo J (2014): Metodologías empleadas en el atlas climático digital de México para la generación de mapas de alta resolución. *GeoActa.*, (39)1:165-173.

Fick SE, Hijmans RJ (2017): Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol*, (37)12:4302-4315.

García E (1981). Modificaciones al sistema de clasificación climática de Köppen: para adaptarlos a las condiciones de la República Mexicana. Tech. Rep.

Hartmann Dl, Amg Klein Tank, Rusticucci M, Alexander L, Brönnimann S, Charabi Y, Dentener FJ, Dlugokencky EJ, Easterling DR, Kaplan A (2013): Climate change: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Observations: Atmosphere and Surface. Cambridge University Press.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis (2005): Very high-resolution interpolated climate surfaces for global land areas. Int. J. Climatol. *A Journal of the Royal Meteorological Society.* (25)15:1965-1978.

Hutchinson M (2004): Anusplin version 4.3. Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia.

INEGI (1988): Jalisco en síntesis. ISBN: 968-892-086-X. Recuperado de: http://internet.contenidos.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/historicos/920/702825920330/702825920330_1.pdf

Lam N (1994): An Introduction to Applied Geostatistics. Ohio State University Pres.

Lam N (1983): Spatial interpolation methods: a review. *The Am. Cartogr*, (10)2:129-150.

Lepot M, Aubin JB, Clemens FH (2017): Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, (9)10:796.

Li J, Heap AD (2008): A review of spatial interpolation methods for environmental scientists, Geoscience Australia Canberra.

Masson D, Frei C (2014): Spatial analysis of precipitation in a high-mountain region: exploring methods with multi-scale topographic predictors and circulation types. *Hydrol. earth system sciences*, (18)11:4543-4563.

Murphy BS (2014): Pykrige: development of a kriging toolkit for Python. *AGU fall meeting abstracts*, Vol. 2014, pp. H51K–0753.

Ruiz-Corral JA, Medina-García G, García Romero GE (2018): Sistema de información agroclimático para México-Centroamérica. *Rev. mexicana de ciencias agrícolas*, (9)1:1-10.

Sáenz-Romero C, Rehfeldt GE, Crookston N L, Duval P, Stamant R, Beaulieu J, Richardson BA (2010): Spline models of contemporary, 2030, 2060 and 2090 climates for Mexico and their use in understanding climate-change impacts on the vegetation. *Climatic change*, (102)3:595-623.

Seabold S, Perktold J (2010): Statsmodels: Econometric and statistical modeling with Python. In 9th Python in Science Conference. Shea DJ, Trenberth KE, Reynolds RW (1992): A global monthly sea surface temperature climatology. *J. of Clim.*, (5)9:987-1001.

Shepherd J, Shindell D, O'carroll CM (2016): What's the difference between weather and climate? Retrieved from NASA 6.

Sluiter R (2009): Interpolation methods for climate data: literature review. KNMI intern rapport, Royal Netherlands Meteorological Institute, De Bilt.

Sørland SL, Schär C, Lüthi D, Kjellström E (2018): Bias patterns and climate change signals in gcm-rcm model chains. Environ. *Research Lett.*, (13)7:2-10.

Téllez O, Hutchinson MA, Nix HA, Jones P (2011): Desarrollo de coberturas digitales climáticas para México. Cambio Climático. Aproximaciones para el Estudio de su Efecto sobre la Biodiversidad., pp. 15-23.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, Van Der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson AR, Jones E, Kern R, Larson E, Carey CJ, Polat I˙, Feng Y, Moore EW, Vanderplas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, Van Mulbregt P (2020): SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, (17)261-272.

Zahroh S, Hidayat Y, Pontoh RS, Santoso A, Bon AT (2019): Modeling and forecasting daily temperature in bandung. In Proceedings of the International Conference on Industrial Engineering and Operations Management Riyadh, Saudi Arabia, pp. 406-412.

Zhu C, Lettenmaier DP (2007). Long-term climate and derived surface hydrology and energy flux data for Mexico: 1925–2004. *J. of Clim.*, (20)9:1936-1946.