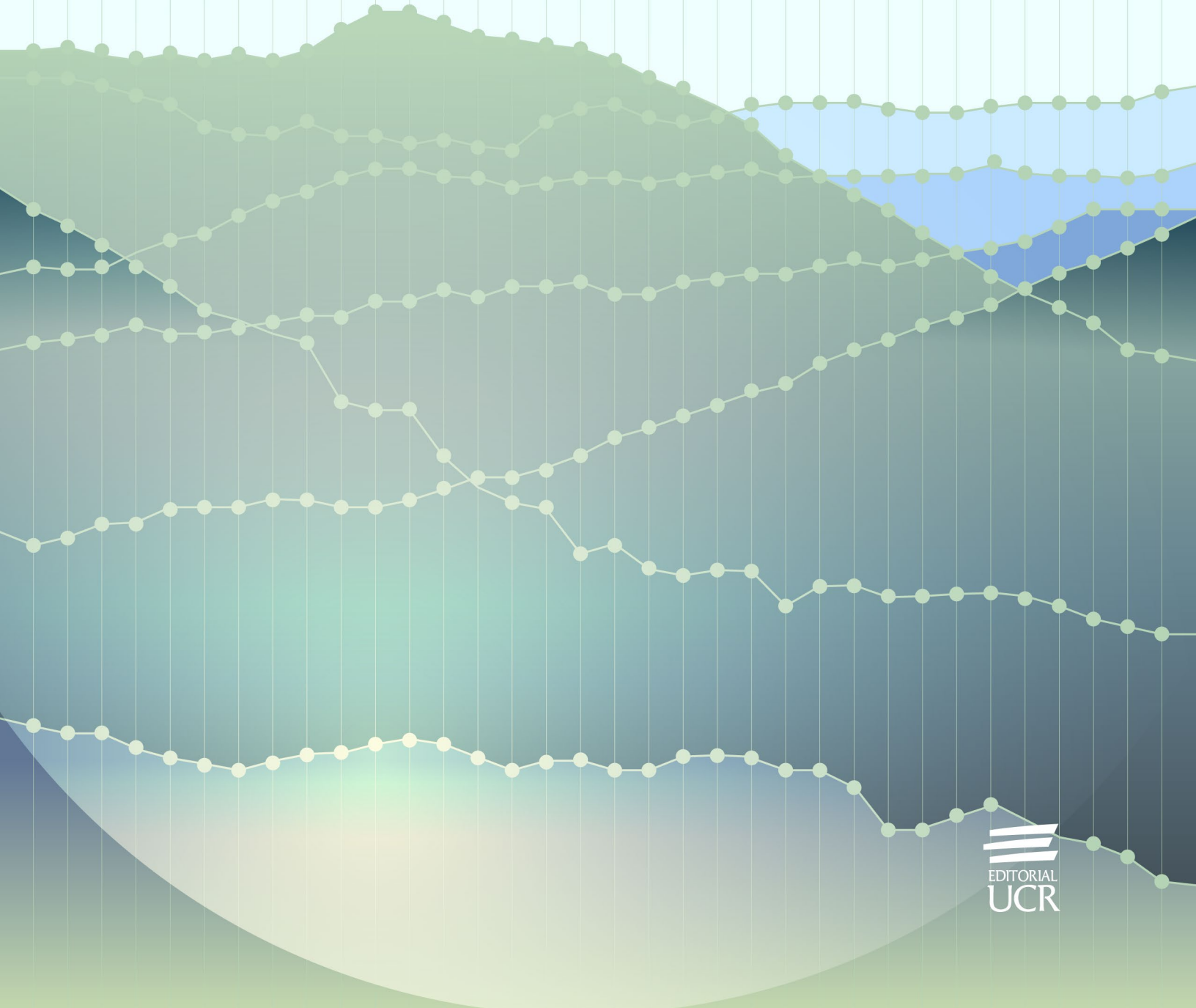
https://revistas.ucr.ac.cr/index.php/ingenieria/index
www.ucr.ac.cr / ISSN: 2215-2652

Ingeniería

Revista de la Universidad de Costa Rica

JULIO/DICIEMBRE 2025 - VOLUMEN 35 (2)



EDITORIAL UCR

UNIVERSIDAD DE COSTA RICA

Ingeniería. Revista de la Universidad de Costa Rica

Vol. 35. No. 2: 29-38, Julio-Diciembre, 2025. ISSN: 2215-2652. San José, Costa Rica

Ingeniería Topográfica

Interpolación espacial de enfermedades foliares en viveros de palma aceitera: una aproximación metodológica

Spatial Interpolation of Foliar Diseases in Oil Palm Nurseries: a Methodological Approach

Jaime Garbanzo-León [1], *Gabriel Garbanzo León* [2]

[1] *Docente e investigador, Escuela de Ingeniería Topográfica, Universidad de Costa Rica, San José, Costa Rica.*
correo: jaime.garbanzoleon@ucr.ac.cr

[2] *Laboratorio de Suelos y Foliares, Centro de Investigaciones Agronómicas, Universidad de Costa Rica, San José, Costa Rica.*
correo: juan.garbanzo@ucr.ac.cr

Palabras clave:

Análisis Espacial, Distancia Interna Ponderada (IDW), Enfermedades Foliares, Severidad, Validación.

Resumen

El uso de interpoladores simples para el diagnóstico de enfermedades foliares puede mejorar el manejo agronómico oportuno en cultivos. Actualmente, existe una carencia de herramientas calibradas y validadas para aplicar de manera precisa los interpoladores en entornos agrícolas, de acuerdo con la distribución espacial de enfermedades foliares en cultivos.

Este trabajo tiene el propósito de definir el algoritmo más apropiado para interpolar enfermedades foliares y determinar el porcentaje adecuado de muestras para realizar la validación cruzada. Para ello, se realizó un estudio de caso utilizando el porcentaje de severidad de la hoja en viveros de palma aceitera. Este se enfocó en analizar los siguientes interpoladores: triangulación (Triangulation), Distancia Inversa Ponderada (IDW), vecinos naturales (Natural Neigbor), spline cúbica (Cubic Spline) y el geoestadístico Kriging Ordinario.

Asimismo, se analizaron distintos porcentajes de muestras para la validación cruzada, que correspondió desde 2.5 % hasta 30 %, utilizando el I-Moran y el método del análisis de poder T ("Power Analysis"). Se encontró que, en la separación muestral a partir de 10 % de la totalidad de los datos, hay menos autocorrelación espacial y, por tanto, el desempeño del interpolador se hace más evidente.

Se concluye que el interpolador IDW presentó la mayor eficacia ($\alpha = 0.05$) para predecir la distribución espacial de una enfermedad foliar.

Recibido: *06/09/2024*
Aceptado: *05/05/2025*

Keywords:

Foliar Diseases, Inverse Distance Weighted (IDW), Severity, Spatial Analysis, Validation.

Abstract

The use of simple interpolation techniques for the diagnosis of foliar disease can improve timely agronomic management in crop systems. Currently, there is a lack of calibrated and validated tools to accurately apply interpolation methods in agricultural environments, according to the spatial distribution of foliar diseases.

This study aims to identify the most appropriate algorithm for interpolating foliar diseases and determining the optimal sample size for cross-validation. A case study was conducted using the leaf severity percentage in oil palm nurseries. This focused on analyzing the performance of the following interpolators: Triangulation, Inverse Distance Weighted (IDW), Natural Neighbor, Cubic Spline, and Ordinary Kriging.

Additionally, various sample sizes for cross-validation were examined, with ranges from 2.5 % to 30 %, using the I-Moran and Power analysis as metrics. It was found that, when the sample size reached or exceeded 10 % of the total dataset, spatial autocorrelation decreased, and thus the performance of the interpolation method became more critical for prediction.

The study concluded that IDW was the most effective interpolation method ($\alpha = 0.05$) for predicting the spatial distribution of a foliar disease, outperforming the other evaluated algorithms.

DOI: 10.15517/ri.v35i2.61815

I. INTRODUCCIÓN

El diagnóstico de las enfermedades foliares es una necesidad en el entorno agrícola, puesto que ayuda en la prevención de la diseminación y de las pérdidas en productividad, que se traducen en aspectos económicos de impacto. El modelado de la distribución espacial es importante para el entendimiento de la propagación de un patógeno; además, uno de los desafíos claves en la patología es describir los patrones de enfermedad con un número limitado de muestras [1]. En términos de modelación, la falta de continuidad espacial en las observaciones de campo hace que los procesos de interpolación sean importantes para estimar o predecir información en sitios donde no son muestreados.

Los algoritmos de interpolación se han desarrollado con distintas finalidades. Por ejemplo, el método de la Distancia Inversa Ponderada (IDW por sus siglas en inglés) se creó para la interpolación espacial de puntos irregulares [2]. Por otra parte, el Kriging se ideó para predecir la espacialidad y probabilidad de encontrar betas de oro [3]. Esta diversidad de orígenes indica que el desempeño de diversas técnicas de interpolación debe ser validado en otras áreas para determinar si representa en forma precisa una distribución espacial. Asimismo, la distribución espacial permite obtener información para el manejo agronómico de enfermedades.

En múltiples trabajos, se han utilizado técnicas de interpolación para diagnosticar enfermedades foliares; sin embargo, estos carecen de validaciones apropiadas. Algunos se han desarrollado para estudiar enfermedades en maíz [4], cacao [5], aguacate [6], algodón [7], frijol [8], [9] y café [10]. Estos estudios han generado planes que permiten manejar y estudiar la espacialidad de enfermedades foliares. No obstante, existe la necesidad de obtener parámetros epidemiológicos temporales y espaciales que ayuden a comprender la distribución asociada a una enfermedad dentro de un agroecosistema [4], [11], [12], [13]. Esto hace necesario estudiar el desempeño de los distintos métodos de interpolación en este campo.

El objetivo de este estudio es evaluar el desempeño de cinco métodos de interpolación espacial para determinar su efectividad y precisión en la predicción de la distribución espacial de enfermedades foliares, utilizando, como caso de estudio, enfermedades en viveros de palma aceitera. Además, se investiga el porcentaje adecuado de la muestra que se debe utilizar para generar las validaciones y optimizar los resultados de interpolación.

II. MATERIALES Y MÉTODOS

Se utilizó una base de datos de vivero de palma aceitera evaluado en el periodo correspondiente a junio de 2014 hasta abril de 2015. El vivero se ubicó en la zona de Corredores, Costa Rica, en la compañía Palma Tica S.A., como parte de un experimento a pequeña escala (6480 m^2). La plantación tenía una densidad de siembra de 8000 plantas ha^{-1}, con un sistema de siembra distribuido en 1.05 cm entre planta y 1.20 cm entre filas.

El manejo del experimento fue igual al desarrollado en la empresa, a excepción de la fertilización, ya que se evaluaron dosis crecientes de fertilizantes y el efecto en la severidad de enfermedades foliares [14]. El experimento presentó seis fechas de muestreos, realizados a los 85, 127, 176, 219, 261 y 304 días después de siembra (dds). En cada fecha, se cuantificó el porcentaje de afectación por las enfermedades en la lámina (% severidad) y se desarrolló una base de datos ubicando sistemáticamente cada planta evaluada, con el fin de realizar un análisis espacial de la distribución de la enfermedad en las láminas foliares. El porcentaje de severidad fue utilizado para todas las interpolaciones.

A. Métodos de interpolación

Para este estudio, las cinco técnicas de estimación utilizadas fueron triangulación (*Triangulation*), Distancia Inversa Ponderada (IDW), vecinos naturales (*Natural Neighbor*), spline cúbica (*Cubic Spline*), y *Kriging* Ordinario. Las primeras cuatro técnicas son clasificadas como interpoladores simples, mientras que Kriging es clasificado como geoestadístico [15]. Para este trabajo, se implementó una metodología de cálculo basada en los principios teóricos de cada interpolador.

En primer lugar, la triangulación es la técnica de interpolación más intuitiva (ver (1)). Esta técnica utiliza los criterios de la triangulación de Delaunay, donde se establece que, para cualquier conjunto S de puntos en un espacio euclidiano, no existe ningún círculo que, al pasar por tres puntos distintos de S, incluya algún otro punto del mismo conjunto S en su interior [15]. Una vez establecida la triangulación, se puede emplear la ecuación canónica del plano para calcular el valor de Z en cualquier punto (X, Y) ubicado dentro del triángulo [16].

$$1 = \frac{x}{a} + \frac{y}{b} + \frac{z}{a}, \tag{1}$$

donde

x, y, z	coordenadas cartesianas;
a, b, c	constantes que se deben determinar para cada triángulo.

En segundo lugar, la técnica de interpolación de vecinos naturales no solo se limita a realizar una triangulación de Delaunay, sino que también toma en cuenta las áreas de influencia de cada punto que contribuya a la interpolación (ver (2)). Para realizar este procedimiento, se calculan los polígonos de Thiessen con el conjunto S de puntos. Luego, se inserta el punto X, lo que conlleva al cálculo de un nuevo polígono de Thiessen con sus vecinos naturales. Estos vecinos están determinados por la intersección de las áreas de los polígonos calculados para S y el polígono resultante de la intersección en el punto. Esta intersección de áreas se utiliza para calcular las coordenadas de los vecinos naturales, la cual es un peso ($\omega(x)$). Estos pesos van de 0 a 1, donde 0 significa la no intersección de área y 1 la completa sobreposición [17].

$$Z = \sum_{i=1}^{n} w_i(x) z_i, \tag{2}$$

donde

Z	valor interpolado;
$\omega_i(x)$	coordenada del vecino natural;
Z_i	atributo de la entidad i.

En tercer lugar, la IDW se basa en el hecho de que los objetos más cercanos están más relacionados entre sí, como se observa en (3). Por lo tanto, este interpolador calcula los pesos como función inversa proporcional a las distancias [15]. Es decir, la influencia de cada punto conocido en la estimación del valor desconocido está ponderada por la distancia entre ellos.

$$Z(\mu) = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}}, \tag{3}$$

donde

$Z(\mu)$	valor interpolado en el punto desconocido μ;
d	distancia al elemento desconocido μ con el atributo conocido i;
Z_i	valor conocido en el puto i;
n	número de entidades utilizadas para interpolar Z(μ);
p	potencia que pondera el peso de la distancia.

En cuarto lugar, el método de interpolación mediante spline cúbica implica la construcción de una función continua y suave, compuesta por polinomios de tercer grado (ver (4)) [18]. En cada segmento de S, se determina un polinomio cúbico que interpola los puntos de dicho segmento, asegurando una transición sin discontinuidades al conectarse con el polinomio del segmento subsecuente.

Estos puntos donde los polinomios se unen se llama nodos[15]. Aunque existen soluciones paramétricas, este trabajo se realizó utilizando una técnica de aproximación [19], programada en el set de herramientas de SAGA-GIS.

$$S_i(x) \quad a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \tag{4}$$

donde

$S_i(x)$	la aproximación del atributo requerido;
a_i, b_i, c_i, d_i	coeficientes calculados según condiciones de la interpolación;
x	es el punto donde el atributo es requerido;
x_i	nodos.

Por último, el método de *Kriging* es ampliamente reconocido como un interpolador geoestadístico, debido a su fundamento en un modelo que incorpora la distribución y el comportamiento específico espacial de las variables requeridas, como lo describe (5) [15].

Este método es reconocido muchas veces como un método óptimo, pues los pesos de la interpolación son escogidos basándose en el mejor estimador lineal no sesgado (BLUE por sus siglas en inglés) [20].

$$Z(S_0) = \sum_{i=1}^{n} \lambda_i \cdot Z(S_i), \tag{5}$$

donde

$\hat{Z}(S_0)$	valor interpolado;
$\hat{Z}(Si)$	valor puntual medido en una ubicación (x_i, y_i);
λi	peso ponderado dependiendo de la distribución de los datos, que va de 0 a 1;
S0	ubicación de la predicción;
n	cantidad puntos de la muestra.

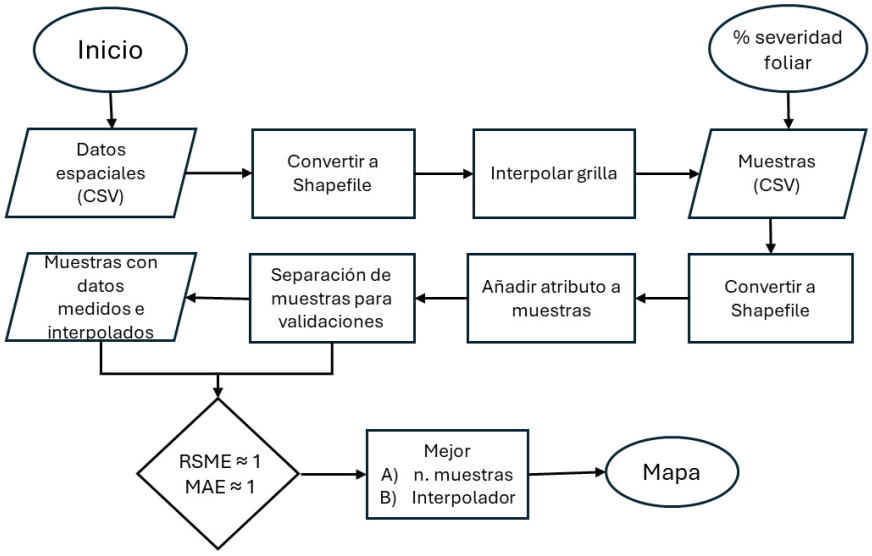


Fig. 1. Diagrama de flujo de proceso de interpolación.

B. *Análisis de datos*

Como lo muestra la Fig. 1, para procesar los datos sistemáticamente, se procedió a utilizar los softwares de sistemas de información geográfica Arcgis, QGIS y el lenguaje de programación Python. Los datos fueron procesados por modelos programados en Python para la validación cruzada. Además, se generaron modelos digitales de "grilla" para analizar la distribución espacial de las enfermedades foliares. Asimismo, se generaron interpolaciones mediante los cinco algoritmos descritos en la sección anterior.. Para cada uno, se generó una interpolación de los datos de severidad (%), se extrajo aleatoriamente y de manera sistemática una muestra creciente (2.5 %, 5 %, 10 %, 15 %, 20 %, 25 % y 30 %) de los datos interpolados y datos originales. Estos con su respectiva ubicación en la grilla, mediante un código programado con el software Matlab versión R2017 (ver (6)).

$$T \subseteq P \quad \text{and} \quad M \subseteq P \quad M \cup T = P \quad \text{and} \quad M \cap T = \emptyset \tag{6}$$

$$\mathrm{M} := \{m_k \in P \mid k \in \mathbb{N} \;\; k \leq N_P\}$$

donde

M y T	conjunto de muestra y conjunto por interpolar, respectivamente, no se repiten y su intersección es vacía;
m_k	elementos que componen a M ($m_k \in M$);
P	totalidad de elementos disponibles de donde se extraen m_k;
k	número aleatorio perteneciente a los enteros naturales ($\mathbb{N}$) y no se repite;
N_P	cantidad de elementos en P.

A partir de (6), M obtendrá el número de elementos correspondiente al porcentaje de elementos extraídos para ser evaluados.

C. *Análisis por el índice de Moran*

Por otra parte, se realizó un análisis geoestadístico mediante el índice de Moran, conocido como I-Moran (ver (7)), con el fin de medir el efecto de aleatorización de los modelos generados en las interpolaciones. Este índice fue desarrollado para evaluar si la autocorrelación espacial es susceptible a variaciones en la densidad de muestreo dentro del conjunto de datos. Asimismo, analiza la autocorrelación espacial en datos, considerando la probabilidad de encontrar muestras similares de manera aleatoria. Un valor p inferior a 0.05 indica una autocorrelación espacial significativa en la población [21]. Este índice se calculó mediante una herramienta "Spatial Autocorrelation Stat" ubicado en el software ArcGIS que viene en el módulo de "ArcPy". Esto permitió un procesado de datos confiables y eficientes (CUADRO I).

$$\text{I-Moran} = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{7}$$

$$\mathrm{S}_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}$$

donde

S_0	a sumatoria de todos los pesos ($\omega_{i,j}$);
Z_i	residuo del atributo de la entidad i con respecto a la media ($x_i - \hat{X}$);
n	número de entidades.

D. *Validaciones de los modelos*

Para determinar el tamaño de muestra adecuado para la validación cruzada de los datos, se procedió a utilizar el método de "Power analysis" [22], [23] (ver (8)). Así, se procedió a realizar bases de datos para cada porcentaje de muestra extraída y se calculó el poder de las muestras.

$$\mathrm{d} = \frac{\bar{y}_B - \bar{y}_A}{S_d} \tag{8}$$

donde

d	tamaño del efecto;
$\bar{y}_A$	media del tratamiento A;
$\bar{y}_B$	media del tratamiento B;
S_d	desviación estándar de los tratamientos B y A.

De igual forma, se realizó un análisis de validación cruzada para los distintos interpoladores. Para esto, se procedió a utilizar la metodología desarrollada por [24]. Asimismo, se calcularon los índices del promedio absoluto del error (MAE), el promedio del cuadrado del error (RSME), así como una correlación de Pearson (ρ) entre los datos observados y los interpolados (predichos).

Una vez obtenido el algoritmo de interpolación más preciso, se procedió a efectuar un análisis temporal a los datos, los cuales fueron analizados por la prueba de Shapiro Wilk para verificar normalidad y, posteriormente, fueron comparados mediante la prueba de Kruskal Wallis ($\alpha = 0.05$).

Esta última prueba se realizó para comparar el efecto de la severidad en las distintas fechas según el algoritmo y validar su trazabilidad en la variabilidad espacial y temporal mediante la interpolación.

Por último, se calculó un valor absoluto, realizando una resta del valor real evaluado en campo y el valor interpolado en el mismo punto ubicado de la grilla. A esta nueva variable se le llamó "valor de diferencia". Con el valor de diferencia, se procedió a realizar un análisis comparativo entre los distintos interpoladores. Todas las variables fueron calculadas mediante el software estadístico RStudio versión 2024.4.0.735 [25].

III. RESULTADOS

Al analizar el efecto de extracción de muestras sobre los datos, se observa la sensibilidad de la variable de respuesta (% severidad), pues si se extrae un porcentaje elevado de muestras, el fenómeno estudiado pierde los patrones espaciales por insuficiencia de información. Esto es debido a que los errores dependen del número y la proximidad de los puntos de información usados para la interpolación de cada celda [26]. Al comparar el efecto del tamaño de las muestras con el I-Moran, se encontró que la menor extracción de datos presentó, como es esperado, una alta autocorrelación espacial (ver CUADRO I). Al aplicar el I-Moran sobre el tamaño de la muestra extraída de los datos originales, se determinó que los grupos superiores a 10 % de la muestra (n = 600) comienzan a mostrar evidencia estadística de ser producto de un evento aleatorio. Ese resultado se puede analizar desde otro ángulo, sugiriendo que las muestras menores a 10 % de los datos generan resultados que están altamente correlacionados

CUADRO I
ANÁLISIS DEL ÍNDICE DE MORAN (I-MORAN α = 0.05), EN MUESTRAS PARA VALIDACIÓN DE 2.5 % HASTA 30 % EN DIFERENTES FECHAS DE MUESTREO DE ENFERMEDADES FOLIARES

Factores	**85 dds**	**127 dds**	**176 dds**	**219 dds**	**261 dds**	**304 dds**
			2.5 % : n = 150			
I-Moran	0.1257	0.1414	0.0714	0.2364	0.0754	0.0549
z *score*	8.5975	6.1571	3.2009	9.5033	3.1449	2.2987
p *value*	< 0.001	< 0.001	0.003	< 0.001	0.002	0.022
% PEA	**< 1 %**	**< 1 %**	**< 1%**	**< 1 %**	**< 1 %**	**< 5 %**
			5 % : n = 300			
I-Moran	0.1598	0.0945	0.0664	0.2380	0.0919	0.0637
z *score*	6.4375	5.3772	2.7425	9.3639	3.7250	2.5919
p *value*	< 0.001	< 0.001	0.006	< 0.001	< 0.001	0.01
% PEA	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**
			10 % : n = 600			
I-Moran	0.1526	0.1634	0.0671	0.1564	0.0537	0.0062
z *score*	6.0672	6.5460	2.6287	9.6985	2.0823	2.5919
p *value*	< 0.001	< 0.001	0.009	< 0.001	< 0.001	0.648
% PEA	**< 1%**	**< 1%**	**< 1%**	**< 1%**	**< 5%**	ns
			15 % : n = 906			
I-Moran	0.1145	0.1850	0.0993	0.1209	0.0276	0.0188
z *score*	5.4558	8.1266	3.6698	7.6166	1.3698	1.1872
p *value*	< 0.001	< 0.001	< 0.001	< 0.001	0.171	0.235
%PEA	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**	**ns**	**ns**
			20 % : n = 1206			
I-Moran	0.1592	0.0925	0.0696	0.2143	0.0269	0.0037
z *score*	7.3150	5.9097	4.0225	8.7573	1.5786	0.4524
p *value*	< 0.001	< 0.001	< 0.001	< 0.001	0.114	0.651
% PEA	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**	**ns**	**ns**
			25 % : n = 1512			

I-Moran	0.1688	0.0363	0.0340	0.1400	0.0340	-0.0055
z *score*	7.8132	2.9523	2.6926	7.4355	1.8397	-0.1782
p *value*	< 0.001	0.003	0.007	< 0.001	0.066	0.859
% PEA	**< 1 %**	**< 1 %**	**< 1 %**	**< 1 %**	**< 10 %**	**Ns**
			30 %: n = 1812			
I-Moran	0.1117	0.0466	0.0453	0.1761	0.0622	0.0116
z *score*	5.8238	2.2436	2.4450	7.5514	2.5068	0.8887
p *value*	< 0.001	0.025	0.014	< 0.001	0.012	0.374
% PEA	**< 1 %**	**< 5 %**	**< 5 %**	**< 1 %**	**< 5 %**	**ns**

Notas: n = tamaño de muestra porcentualmente (%); PEA = probabilidad de que el patrón sea producto de un evento aleatorio en término porcentual (%); ns = el resultado no es estadísticamente diferente de un patrón aleatorio.

La sensibilidad en los patrones de cambio es coherente, pero estos son menos definidos con bajas densidades de datos, lo que influye en las estimaciones para validación. En la Fig. 2, se muestran los efectos visuales en la interpolación de la variable de respuesta (% severidad) con extracciones del 2.5 %y 30 % de los datos, utilizando el método del IDW, cuyo patrón exhibe una notable consistencia visual, sin mostrar alteraciones visuales significativas. No obstante, en las fechas 5 (261 dds) y 6 (304 dds), la estructura espacial se torna menos definida, lo que evidencia una mayor sensibilidad a variaciones en la densidad muestral.

En síntesis, como se esperaba, se observaron mayores discrepancias entre las interpolaciones generadas con muestras de 2.5 % y 30 %. Los resultados acentúan la importancia de considerar la densidad muestral para evaluar la autocorrelación espacial y la estabilidad de los patrones observados en la muestra de datos para validaciones. En efecto, el mejor modelo espacial que se puede obtener es el que tenga la mayor cantidad de datos posibles y, si se desea una validación, se pueden emplear los métodos de validación cruzada tal como leave-one-out [27]

Al analizar el poder estadístico en los diferentes grupos de datos para validación, se encontró que un 15 % de los datos refleja una significancia del 99 % (ver CUADRO II). Los diferentes grupos de muestras exhibieron que, por arriba de 15 % de datos (n = 906), se encuentra un tamaño de muestra suficiente para detectar efectos estadísticamente existentes a un nivel de significancia del 5 %. Esto establece que muestras superiores a 15 % de datos son suficientes para encontrar efectos verdaderos, mientras que un menor porcentaje insinúa una mayor probabilidad de no detectar estos efectos, inclusive si realmente existen.

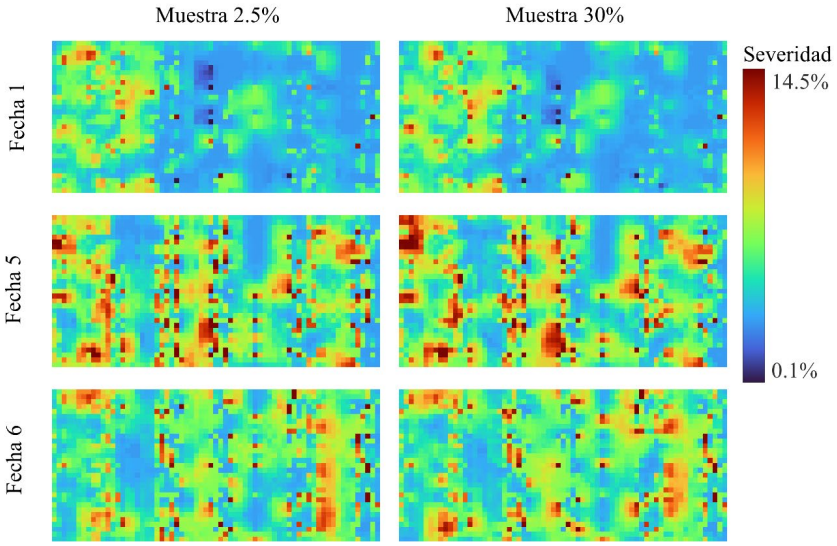


Fig. 2. Modelos digitales de grilla elaborados con el interpolador IDW. Se contrastan las grillas para las fechas 1 (85 dds), 5 (261 dds) y 6 (304 dds), interpoladas con el 2.5 % (n = 150) y el 30 % (n = 1812) de los datos extraídos para validación.

CUADRO II
ANÁLISIS DE PODER ESTADÍSTICO EN DIFERENTES PORCENTAJES DE DATOS PARA VALIDACIÓN EN INTERPOLACIONES FOLIARES DE DATOS

Parámetros	Porcentaje de muestras						
	2.5 %	**5 %**	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**
n	150	300	600	906	1206	1512	1812
k	5	5	5	5	5	5	5
Datos totales	750	1500	3000	4530	6030	7560	9060
N Sig.	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Datos							
f	0.0789	0.0357	0.0673	0.0726	0.0812	0.0726	0.0858
Poder T.	0.37	0.16	0.85	0.99	1.0	1.0	1.0
Datos transformados (log)							
f	0.0879	0.0789	0.0771	0.0776	0.0823	0.0675	0.0661
Poder T.	0.45	0.68	0.94	0.99	1.0	1.0	1.0

Notas: n = tamaño de muestra porcentualmente; k = número de grupos; N Sig.= nivel de significancia.; f = tamaño del efecto; Poder T = prueba de Poder 1 – β.

El promedio absoluto del error (MAE), el promedio del cuadrado del error (RSME) y una correlación de Pearson (ρ) mostraron el mejor ajuste para los datos superiores a 15 % y el interpolador IDW (ver CUADRO III). Se encontró que los mejores ajustes RMSE (más cercanos a 0) se encontraron con kriging, con las muestras de 5 %, 10 %, 15 % y 30 %, seguido de IDW en las muestras 2.5 %, 20 %, 25 %.

Por otro lado, el mejor ajuste de MAE fue para el interpolador IDW en todos los porcentajes de muestras para validación. Solo en la muestra de 10 % el interpolador de vecinos naturales mostró el mismo ajuste que el IDW. La mayor ρ fue con la muestra de 2.5 % (0.43) para el interpolador IDW, seguido de la muestra 25 % (0.37) y 20 % (0.34).

CUADRO III
VALIDACIONES SEGÚN EL PORCENTAJE DE MUESTRAS E INTERPOLADOR EN LA EXPLORACIÓN ESPACIAL DE ENFERMEDADES FOLIARES

Parámetros	IDW	Kriging	*Natural neig.*	*Spline cubic*	*Triangulation*
2.5 % : n = 150					
RMSE	3.92	4.39	4.21	4.96	4.13
MAE	2.05	2.55	2.20	2.78	2.07
ρ	0.43	0.23	0.36	0.27	0.38
5 % : n = 300					
RMSE	3.94	3.76	4.19	4.43	4.47
MAE	2.59	2.64	2.69	2.92	2.75
ρ	0.18	0.24	0.16	0.23	0.09
10 % : n = 600					

RMSE	4.18	4.12	4.32	4.95	4.85
MAE	2.54	2.67	2.54	3.17	2.57
ρ	0.30	0.26	0.30	0.26	0.24
			15 % : n = 906		
RMSE	4.93	4.89	5.09	6.12	5.22
MAE	2.68	2.88	2.82	3.57	2.81
ρ	0.30	0.30	0.27	0.18	0.25
			20 % : n = 1206		
RMSE	3.97	4.09	4.12	5.17	4.37
MAE	2.55	2.79	2.61	3.29	2.59
ρ	0.34	0.26	0.31	0.21	0.29
			25 % : n = 1512		
RMSE	4.30	4.46	4.51	5.88	4.66
MAE	2.62	2.84	2.70	3.39	2.68
ρ	0.37	0.30	0.33	0.25	0.32
			30 % : n = 1812		
RMSE	4.06	4.01	4.29	5.46	4.54
MAE	2.61	2.73	2.73	3.44	2.73
ρ	0.32	0.28	0.29	0.22	0.28

Notas: El promedio absoluto del error (MAE), el promedio del cuadrado del error (RSME) y una correlación de Pearson (ρ).

Al comparar el efecto en una forma no factorizada, se encontró que la muestra de 20 % y IDW es el tamaño de muestra y algoritmo que mostró menores errores de predicción (ver Fig. 3). El RMSE y MAE mostraron una menor tendencia en las muestras 2.5 %, 5 % y 20 %. Esta tendencia se encontró similar con los algoritmos IDW y vecinos naturales. Esto que determina que, para este trabajo, una muestra para validación de interpolación es eficiente con un 20 % de la muestra utilizando el interpolador IDW.

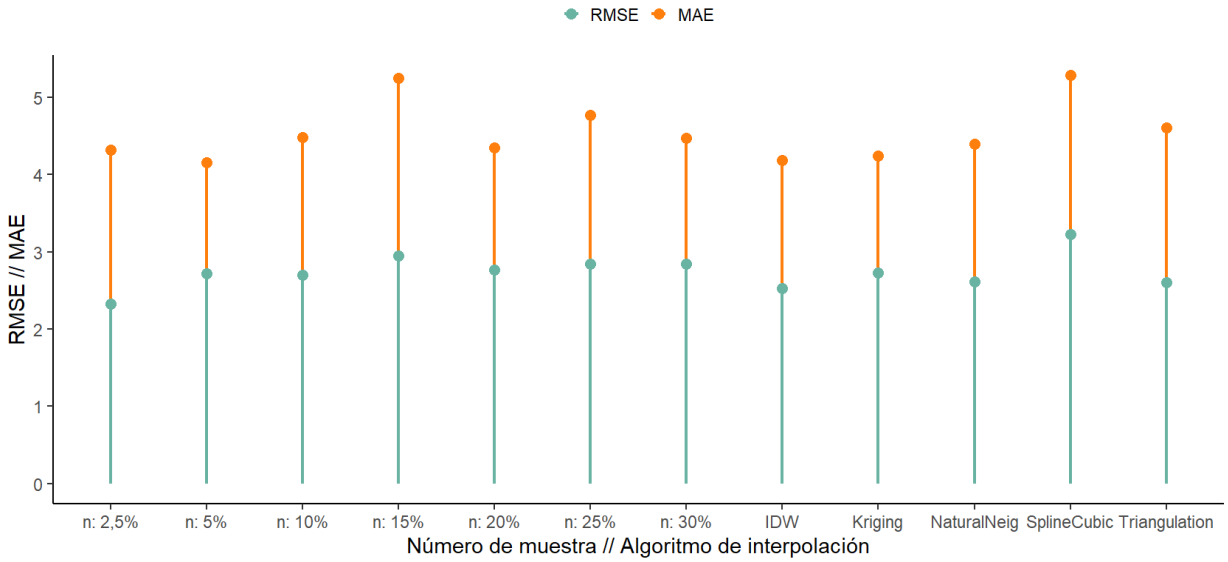


Fig. 3. Determinación no factorial del promedio absoluto del error (MAE, %) y el promedio del cuadrado del error (RSME, %) en siete diferentes porcentajes de muestras para validación y cinco algoritmos de interpolación.

Al comparar las predicciones realizadas por el interpolador, se encontró que el porcentaje de predicción sobreestimó de manera no significativa ($\alpha = 0.05$) un 0.2 % en comparación a lo observado. Asimismo, al analizar los porcentajes de severidad (% severidad) en las distintas fechas, se encontró diferencias significativas ($\alpha = 0.05$) entre las primeras tres fechas (127, 176 dds) y últimas fechas (361, 304 dds). Estas mismas diferencias significativas se cuantificaron con los datos predichos por el IDW. Los valores de diferencia absolutos se encontraron entre 0-0.3 unidades por arriba en los valores predichos de las severidades (ver CUADRO IV). En promedio, el porcentaje sobreestimado es de 0.15 %, lo que podría ser casi nulo para una escala visual de severidad de enfermedades foliares.

CUADRO IV
COMPARACIONES MÚLTIPLES EN DATOS OBSERVADOS Y DATOS PREDICHOS MEDIANTE IDW CON UNA ESTIMACIÓN DE VALOR DE DIFERENCIA ABSOLUTO EN DISTINTAS FECHAS DE MUESTREO

	IDW		
dds	**Observaciones originales**	**Datos predichos**	**Valor diferencia**
85	3266 (4.7) b*	2982 (4.7) c	0
127	2651 (4.0) d	1995 (4.1) e	0.1
176	2811 (4.1) d	2305 (4.2) d	0.1
219	3073 (5.3) c	3298 (5.6) b	0.3
261	3809 (6.6) a	4492 (6.8) a	0.2
304	3851 (6.2) a	4391 (6.4) a	0.2

Notas: *Medias de igual letra no representa diferencias significativas ($\alpha = 0.05$) según prueba Kruskal Wallis. Valores fuera de paréntesis representa rangos de cálculo en la prueba estadística, mientras que valores en paréntesis representan % de severidad en las hojas.

V. DISCUSIÓN

La distribución espacial de una enfermedad foliar está vinculada con la cercanía al foco de inóculos, los cuales pueden ser transportados a hospederos cercanos e incrementar su patrón de distribución sistemáticamente [28], [29], [30], [31]. En este estudio, el método de interpolación IDW demostró ser el algoritmo más eficaz para representar e interpolar la distribución espacial de un inóculo. Esto se debe principalmente a su modelo determinístico de predicción, el cual utiliza las observaciones más cercanas (ver (3)) para estimar el valor desconocido, en función de la distancia de influencia de cada observación. Sin embargo, aunque los otros algoritmos presentaron también errores y probabilidades similares a las del IDW, este último fue el más consistente a través de las métricas y las fechas.

Dicho resultado también es apoyado por el hecho de que el interpolador IDW es un algoritmo establecido en el campo de los sistemas de información geográfica desde su introducción en 1968 por [2]. Esto, pues ha demostrado ser comparable con otros interpoladores cuando ha sido aplicado en datos regulares [32], como los de este experimento. Además, este algoritmo ha superado al Kriging ordinario y Kriging por indicadores en predicción de severidad asociada a enfermedades provocadas por hongos para la nuez de areca [33].

La escogencia de la partición de porcentaje para muestra de validación siempre es una pregunta abierta. Sin embargo, según los análisis basados en la autocorrelación espacial (I-Moran), los datos que tienen una separación de muestras después del 10 % comienzan a percibir más efectos producto de la aleatoriedad, ocasionados por la falta de datos. Por ello, a partir de este porcentaje, el desempeño del interpolador se vuelve más importante.

Por tanto, si el objetivo de los estudios es evaluar las predicciones, a partir de un corte de 90 % - 10 % (%datos - %validación), se podría evaluar la exactitud de los modelos en la predicción de la variable de respuesta. Como es esperado, el I-Moran muestra una mayor autocorrelación cuando la muestra tomada de los datos es del 2.5 % y del 5 %. Sin embargo, este resultado es útil, pues indica que, cuando se tienen datos escasos, se debe buscar otras técnicas para medir la exactitud de los modelos; por ejemplo, leave-one-out.

Asimismo, al utilizar el método de "Power analysis", se determinó que muestras superiores a 15 % tienen suficiente poder estadístico para calcular una mayor probabilidad de efectos. Por consiguiente, basados en el estudio empírico de este experimento, estas métricas sugieren que el porcentaje de datos utilizado para muestras de validación esté en el rango entre 15 % y 30 %. Este resultado es consistente con la regla general empírica de partición de datos para entrenamiento y prueba (80/20) [34].

Por último, es importante destacar que este estudio se llevó a cabo utilizando datos organizados en cuadrícula regular (grilla), donde los métodos de interpolación suelen ofrecer buenos resultados. Por lo tanto, se sugiere realizar más pruebas para evaluar y validar el rendimiento de los interpoladores con datos de distribución espacial irregular, con el fin de obtener una evaluación más completa y precisa de su optimización.

V. CONCLUSIONES

Este estudio evaluó el desempeño de cinco métodos de interpolación y los porcentajes de partición de datos para entrenamiento y validación, aplicados a la variable de respuesta que mide el porcentaje de severidad de enfermedades foliares en las hojas (% severidad). De todos los interpoladores evaluados, la Distancia Inversa Ponderada (IDW) mostró un mejor desempeño, seguido por el kriging ordinario, a través de los indicadores de ajuste de los modelos. Por otro lado, se determinó empíricamente que un rango de entre 10 % y 30 % de los datos para muestras de validación es el más recomendable. De esta manera, la autocorrelación espacial comienza a disminuir cuando se separa

el 10 % de los datos para muestras, lo que permite una evaluación más precisa de la eficiencia del interpolador. Asimismo, este estudio demostró que muestras superiores al 15 % para validación, proporcionan suficiente poder estadístico para calcular con mayor precisión la probabilidad de los efectos. Para futuros trabajos en el tema de interpolación y enfermedades foliares, se recomienda repetir este estudio con datos con distribución espacial irregular y en otros tipos de cultivos, para generar una visión más completa del desempeño de los interpoladores y con diferentes enfermedades foliares.

DECLARACIÓN DE CONFLICTO DE INTERESES

Los autores declaran que no tienen intereses financieros ni relaciones personales conocidos que hayan podido influir en el trabajo presentado en este artículo.

AGRADECIMIENTOS

Se extiende un agradecimiento especial a Palma Tica S.A. por la información brindada a lo largo de varios años de trabajo. Asimismo, agradecemos a los M.Sc. Floria Ramírez y M.Sc. Jesús Céspedes por sus valiosos comentarios y aportes que contribuyeron a mejorar esta investigación.

ROLES DE LOS AUTORES

Jaime Garbanzo León: Conceptualización, Análisis formal, Metodología, Software, Validación, Redacción – borrador original, Redacción – revisión y edición.

Gabriel Garbanzo León: Curación de datos, Conceptualización, Análisis formal, Software, Metodología, Validación, Redacción – borrador original, Redacción – revisión y edición.

REFERENCIAS

[1] W. Luo et al., "An improved regulatory sampling method for mapping and representing plant disease from a limited number of samples", Epidemics, vol. 4, no. 2, pp. 68-77, 2012, doi: 10.1016/j.epidem.2012.02.001.

[2] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data", en Proceedings of the 1968 23rd ACM national conference, ene. 1968, pp. 517-524.

[3] D. G. Krige, "A statistical approach to some basic mine valuation problems on the Witwatersrand", J. South Afr. Inst. Min. Metall., vol. 52, no. 6, pp. 119-139, dic. 1951.

[4] M. V. Micca, N. R. Andrada y A. S. Larrusse, "Análisis exploratorio espacial de tizón común exserohilum turcicum (Leonard and Suggs) en estratos foliares de maíz en Villa Mercedes, San Luis", FAVE Sección Ciencias Agrarias, vol. 14, no. 2, pp. 111-122, jun. 2016, doi: 10.14409/fa.v14i2.5724.

[5] N. J. Cárdenas Pardo, A. E. Darghan Contreras, M. D. Sosa Rico y A. Rodríguez, "Análisis espacial de la incidencia de enfermedades en diferentes genotipos de cacao (Theobroma cacao L.) en El Yopal (Casanare), Colombia", Acta Biolo. Colomb., vol. 22, no. 2, pp. 209-220, may. 2017, doi: 10.15446/abc.v22n2.61161.

[6] A. Tapia-Rodríguez, J. F. Ramírez-Dávila, D. K. Figueroa-Figueroa, M. L. Salgado-Siclan y R. Serrato-Cuevas, "Spatial analysis of anthracnose in avocado cultivation in the State of Mexico", Rev. Mexic. Fitopatol., vol. 38, no. 1, pp. 132-145, dic. 2019, doi: 10.18781/R.MEX.FIT.1911-1.

[7] J. A. Pizzato et al., "Geostatistics as a Methodology for Studying the Spatiotemporal Dynamics of Ramularia areola in Cotton Crops", Am. J. Plant Sci., vol. 05, no. 15, pp. 2472-2479, jul. 2014, doi: 10.4236/ajps.2014.515262.

[8] M. de Carvalho Alves y E. A. Pozza, "Indicator kriging modeling epidemiology of common bean anthracnose", Appl. Geomatics, vol. 2, no. 2, pp. 65-72, jun. 2010, doi: 10.1007/s12518-010-0021-1.

[9] M. C. Alves, E. A. Pozza, J. C. Machado, D. V. Araújo, V. Talamini y M. S. Oliveira, "Geoestatística como metodologia para estudar a dinâmica espaço-temporal de doenças associadas a Colletotrichum spp. transmitidos por sementes", Fitopatol. Bras., vol. 31, no. 6, pp. 557-563, dic. 2006, doi: 10.1590/S0100-41582006000600004.

[10] M. de Carvalho Alves, F. M. da Silva, E. A. Pozza y M. S. de Oliveira, "Modeling spatial variability and pattern of rust and brown eye spot in coffee agroecosystem", J. Pest. Sci., vol. 82, no. 2, pp. 137-148, may. 2009, doi: 10.1007/s10340-008-0232-y.

[11] M. R. Nelson, T. V. Orum, R. Jaime-Garcia y A. Nadeem, "Applications of Geographic Information Systems and Geostatistics in Plant Disease Epidemiology and Management", Plant. Dis., vol. 83, no. 4, pp. 308-319, abr. 1999, doi: 10.1094/PDIS.1999.83.4.308.

[12] L. V. Madden y G. Hughes, "Plant Disease Incidence: Distributions, Heterogeneity, and Temporal Analysis", Annu. Rev. Phytopathol., vol. 33, no. 1, pp. 529-564, sep. 1995, doi: 10.1146/annurev.py.33.090195.002525.

[13] R. Jaime-Garcia, T. V. Orum, R. Felix-Gastelum, R. Trinidad-Correa, H. D. VanEtten y M. R. Nelson, "Spatial Analysis of Phytophthora infestans Genotypes and Late Blight Severity on Tomato and Potato in the Del Fuerte Valley Using Geostatistics and Geographic Information Systems", Phytopathology, vol. 91, no. 12, pp. 1156-1165, dic. 2001, doi: 10.1094/PHYTO.2001.91.12.1156.

[14] G. Garbanzo, E. Molina, G. Cabalceta, and F. Ramírez, "Evaluación de Si y Ca foliar en el crecimiento y tolerancia de complejo de necrosis foliar en palma aceitera", Agronomía Costarricense, vol. 42, no. 2, jun. 2018, doi: 10.15517/rac.v42i2.33777.

[15] N. Panigrahi, "Spatial Interpolation Techniques", en Computing in Geographic Information Systems. Londres, Reino Unido: Taylor & Francis Group, 2014, pp. 155-167.

[16] E. Stefanakis, Geographic Databases and Information Systems. Los Ángeles, CA, Estados Unidos: CreateSpace Independent Publishing Platform, 2014.

[17] H. Ledoux yC. Gold, "An Efficient Natural Neighbour Interpolation Algorithm for Geoscientific Modelling", en Developments in Spatial Data Handling, 2005, pp. 97-108. doi: 10.1007/3-540-26772-7_8.

[18] M. A. Ramadan, I. F. Lashien y W. K. Zahra, "Polynomial and nonpolynomial spline approaches to the numerical solution of second order boundary value problems", Appl. Math. Comput., vol. 184, no. 2, pp. 476-484, ene. 2007, doi: 10.1016/j.amc.2006.06.053.

[19] J. Haber, F. Zeilfelder, O. Davydov y H.-P. Seidel, "Smooth approximation and rendering of large scattered data sets", en Proceedings Visualization, 2001, pp. 341-571. doi: 10.1109/VISUAL.2001.964530.

[20] R. Sunila y K. Kollo, "Kriging and Fuzzy Approaches for DEM", en Quality aspects in spatial data mining, A. Stein, W. Shi y W. Bijker, Eds., Londres, Reino Unido: Taylor & Francis Group, 2009, pp. 102-114.

[21] M. Goodchild, Spatial autocorrelation. Norwich, Reino Unido: Geo Book, 1986.

[22] R. B. Bausell y Y.-F. Li, Power analysis for experimental research : a practical guide for the biological, medical, and social sciences. Cambridge, Reino Unido: Cambridge University Press, 2002.

[23] J. Cohen, Statistical Power Analysis for the Behavioral Sciences Second Edition, vol. 2. Mahwah, NJ, Estados Unidos: Lawrence Erlbaum Associates, 1988.

[24] C. A. Schloeder, N. E. Zimmerman y M. J. Jacobs, "Division S-8—nutrient management & soil & plant analysis: comparison of methods for interpolating soil properties using limited data", Soil Sci. Soc. Am. J., vol. 65, no. 2, pp. 470-479, 2001, doi: 10.2136/sssaj2001.652470x.

[25] R: A Language and Environment for Statistical Computing (2024), R Foundation for Statistical Computing. Accesado: mar. 30, 2024. [En línea]. Disponible en: https://www.R-project.org

[26] D. L. Phillips y D. G. Marks, "Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models", Ecol. Modell., vol. 91, no. 1, pp. 213-229, nov. 1996, doi: 10.1016/0304-3800(95)00191-3.

[27] A. M. Molinaro, R. Simon y R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods", Bioinformatics, vol. 21, no. 15, pp. 3301-3307, ago. 2005, doi: 10.1093/bioinformatics/bti499.

[28] R. K. Meentemeyer, B. L. Anacker, W. Mark y D. M. Rizzo, "Early detection of emerging forest disease using dispersal estimation and ecological niche modeling", Ecological Applications, vol. 18, no. 2, pp. 377-390, mar. 2008, doi: 10.1890/07-1150.1.

[29] L. Willocquet, L. Lebreton, A. Sarniguet y P. Lucas, "Quantification of within-season focal spread of wheat take-all in relation to pathogen genotype and host spatial distribution", Plant Pathol., vol. 57, no. 5, pp. 906-915, oct. 2008, doi: 10.1111/j.1365-3059.2008.01834.x.

[30] M. M. Ndoungué Djeumekop et al., "Spatial and Temporal Analysis of Phytophthora megakarya Epidemic in Newly Established Cacao Plantations", Plant Dis., vol. 105, no. 5, pp. 1448-1460, may. 2021, doi: 10.1094/PDIS-09-19-2024-RE.

[31] F. Hay, D. W. Heck, A. Klein, S. Sharma, C. Hoepting y S. J. Pethybridge, "Spatiotemporal Dynamics of Stemphylium Leaf Blight and Potential Inoculum Sources in New York Onion Fields", Plant Dis., vol. 106, no. 5, pp. 1381-1391, may. 2022, doi: 10.1094/PDIS-07-21-1587-RE.

[32] G. Garnero y D. Godone, "Comparisons between different interpolation techniques", Int. Arch. of the Photogramm. Remote Sens. Spatial Inf. Sci., vol. 40, pp. 139-144, ene. 2014, doi: 10.5194/isprsarchives-XL-5-W3-139-2013.

[33] P. Balanagouda et al., "Assessment of the spatial distribution and risk associated with fruit rot disease in Areca catechu L.", J. Fungi, vol. 7, no. 10, p. 797, sep. 2021, doi: 10.3390/jof7100797.

[34] V. R. Joseph, "Optimal ratio for data splitting", Stat. Anal. Data Min.: ASA Data Sci. J., vol. 15, no. 4, pp. 531-538, ago. 2022, doi: 10.1002/sam.11583.