UNIVERSIDAD DE SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA
DOCTORADO EN ESTADÍSTICA MULTIVARIANTE APLICADA

TESIS DOCTORAL

"BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB): Una aplicación a datos de mortalidad por cáncer de mama en el Ecuador"

AUTOR
Leyda Elizabeth Jaramillo Feijoo

DIRECTOR
Dr. Jhony Joe Real Cotto

TUTORA
Dra. María Purificación Galindo Villardón

2023

“BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB): Una aplicación a datos de mortalidad por cáncer de mama en el Ecuador”

UNIVERSIDAD DE SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA

Memoria que, para optar al Grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Leyda Elizabeth Jaramillo Feijoo

Salamanca
2023

UNIVERSIDAD DE SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA

Dr. Jhony Joe Real Cotto. Director
Jefe del Departamento Bioestadística Hospital SOLCA Guayaquil

Dra. M. Purificación Galindo Villardón. Tutora
Catedrática de la Universidad de Salamanca. Departamento de Estadística.

CERTIFICAN que Doña Leyda Elizabeth Jaramillo Feijoo ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que, para optar título de Grado de Doctor, presenta con el título "BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB): Una aplicación a datos de mortalidad por cáncer de mama en el Ecuador", autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca al 30 de mayo 2023.

Dr. Jhony Joe Real Cotto

Dra. M. Purificación Galindo Villardón

AGRADECIMIENTOS

Agradezco a Dios por darme la fortaleza en mi vida diaria

Mi más sincero agradecimiento a mi Director Jhony Real Cotto por su apoyo incondicional, por sus conocimientos invaluables y consejos que me brindó durante el transcurso de este programa de doctorado. ¡Gracias¡

Un agradecimiento muy especial a mi querida Purificación Galindo Villardón, por darme esta oportunidad, por la confianza que depositó en mí, por sus conocomientos y por transmitir esa pasión por la estadística con la cual me identifico.

Agradezco infinitamente a los Directivos de SOLCA Guayaquil, que me brindaron su apoyo económico y me dieron todas las facilidades.

Esto no hubiera sido posible sin el apoyo de mi querida familia, José Luis, Linda, Luis y Leyda, mis padres y hermanas que me brindaron su apoyo y comprensión, les quedo agradecida por siempre.

Finalmente, agradezco a los amigos del doctorado por su apoyo, por su amistad y compartir sus conocimientos.

MUCHAS GRACIAS

DEDICATORIA

A Dios

A José Luis, Linda, Luis y Leyda

A mis padres y hermanas

A los guerreros que luchan contra el cáncer

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURA

Pág.

RESUMEN

El objetivo del estudio es combinar las técnicas GWPCA y la prueba estadística no paramétrica Mann-Kendall que son ampliamente usadas para analizar la componente espacial y temporal. Se aplican de forma individual y no hay una representación simultánea. En este artículo se propone una técnica multivariante que la hemos denominado Biplot Logístico ponderado espacio temporal (TSWLB) combina las componentes espacial y temporal para representarlos en un gráfico facilitando la interpretación de las relaciones entre los sitios geográficos y las variables, siendo de interés su aplicación en distintas áreas. Nosotros aplicamos la técnica propuesta en datos de mortalidad por cáncer de mama en el Ecuador. Se utilizó el paquete GWModel, la librería Kendall ambos del lenguaje R y el programa MultiBiplot. Se observó un incremento sostenido de las tasas de mortalidad por cáncer de mama en el Ecuador con una mayor variabilidad de las muertes por esta enfermedad al norte y sur del país. La técnica TSWLB representó simultáneamente las características espacio temporales dando un ordenamiento a los sitios geográficos e identificando cuatro clústeres, siendo el clúster dos, conformada por las provincias: Guayas, El Oro, Santo Domingo de los Tsáchilas y Chimborazo, el más prioritario por presentar una tendencia creciente estadísticamente significativa de la tasa de mortalidad por cáncer de mama y con presencia de altas tasas en años recientes, información que permite orientar las intervenciones en salud por esta enfermedad.

Palabras clave: Geográficamente ponderado, tendencia temporal, análisis espacial, mortalidad, cáncer de mama

ABSTRACT

The objective of the study is to combine the GWPCA techniques and the Mann-Kendall non-parametric statistical test that are widely used to analyze the spatial and temporal component. They are applied individually and there is no simultaneous representation. This article proposes a multivariate technique that we have called Time-Space Weighted Logistic Biplot (TSWLB) combines the spatial and temporal components to represent them in a graph, facilitating the interpretation of the relationships between geographic sites and variables, its application being of interest. in different areas. We apply the proposed technique to breast cancer mortality data in Ecuador. The GWModel package, the Kendall library both from the R language and the MultiBiplot program were used. A sustained increase in mortality rates from breast cancer was observed in Ecuador with a greater variability of deaths from this disease in the north and south of the country. The TSWLB technique simultaneously represented the spatio-temporal characteristics, ordering the geographic sites and identifying four clusters, with cluster two, made up of the provinces: Guayas, El Oro, Santo Domingo de los Tsáchilas and Chimborazo, the highest priority for presenting a statistically significant increasing trend in the mortality rate from breast cancer and with the presence of high rates in recent years, information that allows guiding health interventions for this disease.

Keywords: Geographically weighted, temporal trend, spatial analysis, mortality, breast cancer.

INTRODUCCIÓN

Según la Agencia Internacional para investigación de Cáncer (IARC por sus siglas en inglés), una de cada cinco personas en todo el mundo desarrolla cáncer durante su vida. La prevención y la detección oportuna del diagnóstico de cáncer son retos de salud pública en el siglo XXI. Según la evidencia científica actual, al menos el 40% de todos los casos de cáncer pueden prevenirse mediante estrategias de prevención primaria, con la finalidad de reducir la mortalidad. (IARC, Cáncer)

El impacto del cáncer en la comunidad mundial es inmediato por el número de casos nuevos y la mortalidad. Los registros de cáncer de todos los países generan información que permiten evidenciar cambios en el tiempo (tendencias) en los diferentes tipos de tumores. Una vez reconocidos, estos cambios tanto en la incidencia y mortalidad por cáncer a menudo se pueden atribuir a patrones de desarrollo humano y a su vez, tales relaciones brindan claras oportunidades para la prevención del cáncer. (IARC, Cáncer)

Actualmente, uno de los tipos de cáncer con mayor incidencia en las mujeres es el cáncer de mama. Es una enfermedad que se forma en las células de las mamas y las mismas se multiplica sin control; es decir, consiste en una proliferación rápida e incontrolable de células del epitelio glandular, que han aumentado su capacidad reproductiva, pudiendo diseminarse a través de la sangre o de los vasos linfáticos y llegar a otras partes del cuerpo.

Este cáncer, según la Organización Mundial de la Salud (OMS) es muy frecuente con más de 2,2 millones de casos en el año 2020, aproximadamente 1 de cada 212 mujeres se enfermarán de cáncer de mama a lo largo de su existencia; siendo esta patología la causa más común de muerte a nivel

mundial, falleciendo alrededor de 685.000 mujeres en el año 2020, como consecuencia de esta enfermedad. (OMS, Cáncer de mama)

En países como España, es el cáncer más frecuente en las mujeres y se calcula que para el año 2022 se diagnosticarían 34.750 casos, es decir que 1 de cada 8 mujeres tendrán un cáncer de mama en algún momento de su vida; así mismo, la mortalidad representa la primera causa de muerte, en el año 2020 fallecieron 6.572 mujeres, es de menciona que existen mejoras en los tratamientos, sin embargo, no es suficiente.

En Ecuador, según GLOBOCAN 2020, el cáncer de mama es la mayor incidencia, con 3.563 casos que representó el 12,2% de los casos de cánceres en mujeres y en la mortalidad ocupa el cuarto lugar con 1.056 fallecimientos por esta enfermedad representando un 7%. (SEOM: Cáncer de mama.) (GLOBOCAN - IARC, 2020)

Además, se han registrado mayormente el número de casos y de muertes en países de ingresos bajos y medianos, observándose disparidades del cáncer de mama con los países de ingresos altos en forma considerable, en estos últimos países la supervivencia del cáncer de mama es superior a los 5 años.

Los países con el mayor porcentaje de defunciones por cáncer de mama son África y Polinesia. La mitad de las muertes en África subsahariana suceden en mujeres menores de 50 años de edad.

Cabe indicar, que desde el año 1980 ha habido avances importantes en el tratamiento del cáncer de mama y entre 1980 al 2020, los países de ingresos altos lograron reducir en un 40% la mortalidad; mientras que, en los países de ingresos bajos y medianos, todavía no se logra este objetivo; siendo la

combinación de la detección precoz y terapias eficaces, que se basan en cirugía, farmacoterapia y radioterapia, la estrategia para obtener la mejora en los resultados.

Hay que considerar que el cáncer de mama se origina en el epitelio que son células de revestimiento de los conductos en el 85% o de los lóbulos del tejido glandular de los senos del 15%; a su inicio el tumor maligno está confinado en el conducto o lóbulo, etapa in situ, donde de manera general no causa síntomas.

Al paso del tiempo este cáncer in situ (estadio 0) puede progresar e invadir el tejido mamario adyacente volviéndose un cáncer de mama invasivo; luego este puede propagarse a los ganglios linfáticos contiguos presentando metástasis regional o a otros órganos del organismo humano y producir una metástasis a distancia; y se conoce que como consecuencia de la metástasis generalizada es cuando una mujer muere por cáncer de mama.

Según la OMS, en el año 2020 se diagnosticaron millones de mujeres con cáncer de mama y defunciones por esa enfermedad anteriormente descrito, y a fines de ese año, aproximadamente 7,8 millones de mujeres a las que en los últimos cinco años se les había diagnosticado seguían con vida, por lo que este cáncer es de mayor prevalencia en el mundo. Además, los años de vida perdidos ajustados superan a los otros tipos de cánceres; considerando que esta enfermedad afecta a mujeres en cualquier edad posterior a la pubertad en todos los países, pero estos indicadores aumentan en su vida adulta.

La mortalidad de mama en los países de ingresos altos entre los años de 1980 al 2020 se redujo en un 40%, esto es resultado de los países que han tenido éxito con sus estrategias y a sus esfuerzos por disminuir las muertes por cáncer de mama siendo un logro del 2% al 4% de reducción anual; por lo

tanto, si este porcentaje se consiguiera reducir anualmente en un 2,5% entre el 2020 al 2040 se podría evitar 2,5 millones de muertes por esta enfermedad, evitándose el 25% de muertes para el año 2030 en mujeres menores de 70 años. Es de anotar, que los pilares fundamentales son la detección precoz, diagnóstico oportuno y la gestión integral del cáncer de mama.

La importancia de mejorar los resultados relativos al cáncer de mama depende de las estrategias de fortalecimiento de los sistemas de salud para proveer los tratamientos que sean eficaces a las pacientes y ofrecer una atención primaria de salud enfocada en la detección temprana para que, a su vez, derive a las pacientes a las unidades de salud oncológicas especializadas y de mayor complejidad, dando una atención eficiente y de calidad, debido a que el cáncer de mama es una enfermedad de referencia para su tratamiento.

Según la OMS en el mundo se mueren por esta enfermedad unas 685.000 mujeres, por lo que resulta importante realizarse las siguientes preguntas ¿qué localizaciones tienen mayor predominio estas muertes? ¿las localizaciones se relacionan con factores de riesgo en la población? ¿o tal vez, se desea conocer el patrón de distribución espacial y de tendencia que tiene cada localización?, el factor común es la localización o sitio geográfico, el cual es un componente esencial en los análisis espacio temporales, por lo que juegan un rol importante en las investigaciones en salud, resultando análisis muy enriquecedores para los tomadores de decisiones de políticas públicas en salud.

Existen varias técnicas multivariantes que se utilizan para realizar los análisis espaciotemporales. Las técnicas con enfoque geográficamente ponderado GW han sido ampliamente aplicadas y desarrolladas, como el GWPCA, por otro lado, la prueba de Mann-Kendall ha sido utilizada para el análisis de la tendencia, sin embargo, dichas técnicas son aplicadas de manera individual y no teniendo una comprensión holística de la estructura de los datos

Siendo el cáncer un problema de salud a nivel mundial y evidenciando la necesidad de aportar con nuevas técnicas innovadoras e integradoras con enfoque espacio temporal, he tenido la motivación de desarrollar una técnica integradora que combina información del GWPCA y la prueba de Mann-Kendall para representarlo de manera simultánea en un Biplot logístico externo, con fácil interpretación y proponiéndolo como una metodología para priorizar sitios geográficos en función de los indicadores o características.

Se ha organizado el trabajo en cinco capítulos de la siguiente manera:

El primer capítulo comprende el planteamiento del problema, la justificación del mismo y los objetivos generales y específicos de la presente investigación.

En el segundo capítulo, se realiza una revisión de los métodos multivariantes con enfoque geográficamente ponderado, las técnicas clásicas de Biplot y las alternativas de dichas técnicas para cuando se tienen matrices binarias, a su vez, se presentan las aplicaciones realizadas en distintos ámbitos hasta la fecha; además, la utilidad de la prueba estadística no paramétrica de Mann-Kendall para examinar la tendencia, finalmente, se presentan análisis espacio-temporales que se han realizado en el Ecuador y las técnicas utilizadas.

En el tercer capítulo, se desarrolla el método propuesto Biplot logístico ponderado espacio temporal (TSWLB) para clasificar las unidades espaciales utilizando las componentes espacial y temporal codificados como variables binarias, se muestra su geometría e interpretación, y la calidad de representación, sirviendo como herramienta integradora entre el GWPCA, la prueba de Mann-Kendall y el biplot logístico externo generando una comprensión holística de la estructura de datos.

En el cuarto capítulo se estudian las muertes por cáncer de mama en el Ecuador en un periodo del 2007 al 2021, mediante la técnica de Biplot Logístico ponderado espacio temporal TSWLB propuesta en esta investigación.

En el capítulo cinco se presentan las conclusiones del análisis realizado mediante el método propuesto y se presentan las bondades del mismo.

Los cálculos y representación gráfica se utilizó el lenguaje R y el software MultiBiplot (Vicente-Villardón, 2021).

CAPÍTULO I

1. PLANTEAMIENTO DEL PROBLEMA

1.1 DETERMINACIÓN DEL PROBLEMA

En la actualidad, existen varias técnicas de análisis espacio temporal y sistemas de información geográfico que son ampliamente usados en distintas áreas, como el área de la salud donde existe un campo que es la epidemiología espacial, sin embargo, las técnicas actuales son un poco complejas tanto la aplicación e interpretación. Por otro lado, el cáncer es un problema de salud a nivel mundial, en las mujeres la primera causa de muerte es el cáncer de mama, siendo necesario realizar investigaciones que capturen los componentes espacial y temporal y que orienten e identifiquen áreas prioritarias de atención, por lo que se propone una técnica multivariante que representen simultáneamente lo espacial y temporal.

1.2 PREGUNTA DE INVESTIGACIÓN

¿Existe una técnica multivariante del Biplot Logístico que combine los componentes espacial y temporal para representarlas en un gráfico que permita una fácil interpretación de las relaciones entre los sitios geográficos y las variables, aplicado en la mortalidad por cáncer de mama en el Ecuador?

1.3 JUSTIFICACIÓN

Esta investigación es para contribuir con una técnica multivariante donde se pueda capturar y representar los componentes espacio temporal, además, de dar un ordenamiento y priorización a los sitios geográficos. Dicha técnica fue

aplicada a datos de mortalidad por cáncer de mama en el Ecuador, aportando con información a los tomadores de decisión para la orientación en la formulación de políticas de prevención y control del cáncer.

1.4 FORMULACIÓN DE OBJETIVOS

1.4.1 Objetivo general

Proponer un Biplot Logístico ponderado espacio temporal (TSWLB) técnica multivariante que combina los componentes espacial y temporal para representarlos en un gráfico que permita una fácil interpretación de las relaciones entre los sitios geográficos y las variables, es de interés su aplicación en distintas áreas.

1.4.2 Objetivos específicos

- Realizar una exhaustiva revisión bibliográfica sobre el estado del arte de la técnica multivariante espacio temporal y técnicas Biplot.
- Aplicar las técnicas espacio temporales GWPCA y prueba de Mann-Kendall con datos de mortalidad por cáncer de mama.
- Integrar las técnicas de GWPCA y Prueba de Mann-Kendall en una matriz binaria para ser representados en un Biplot logístico externo.
- Elaborar el Biplot logístico ponderado espacio temporal para identificar las provincias prioritarias en la mortalidad por cáncer de mama del Ecuador.

CAPÍTULO II

2. MARCO TEÓRICO

En este capítulo, se presenta la matriz de datos espaciales, una revisión de las técnicas multivariantes con enfoque geográficamente ponderado que se utilizan para realizar análisis espacio temporal. Se detallan los paquetes en lenguaje R que han sido desarrollados para su aplicación en distintas áreas, además, se abordan los métodos Biplot, la prueba estadística de Mann-Kendall y la importancia de dichas técnicas en el campo de la salud.

2.1 MATRIZ DE DATOS ESPACIALES

Los datos espaciales, contienen tanto información de atributos como de ubicación geográfica y los datos no espaciales contienen solo información de atributos, por ejemplo, las personas que padecen un determinado tipo de enfermedad en diferentes partes de un país son datos espaciales, es importante diferenciarlos porque existen técnicas estadísticas desarrolladas para datos no espaciales que no son válidas aplicarlas en datos espaciales. (Fotheringham et al., 2003)

Los datos u objetos espaciales pueden ser de tipo vector o ráster. Los datos espaciales tipo vector se clasifican a su vez en: puntos, líneas o áreas, todos se caracterizan por describir la referencia espacial de la entidad a ser analizada, además, de mediciones de los atributos o características de dicha entidad.

Para usar un tipo de objeto espacial dependerá de la naturaleza de lo que se está observando. Dichos objetos espaciales tienen propiedades que requieren ser analizadas con técnicas estadísticas multivariantes, como las

desarrolladas ampliamente con enfoque geográficamente ponderado. (Fotheringham et al., 2000) (Harris et al., 2011)

Las técnicas multivariantes geográficamente ponderadas requieren del cálculo de la distancia entre objetos espaciales, los mismos, deben ser descritos sobre la superficie terrestre en términos de latitud y longitud, es de mencionar que, las unidades de medida de latitud y longitud son en grados, minutos y segundos, resultando el cálculo de la distancia algo complejo, por lo que es conveniente convertirlas a un sistema de coordenadas plano, mediante la proyección de las coordenadas desde la esfera hacia el plano, existen diferentes proyecciones que varían sus propiedades y modos de construcción. (Fotheringham et al., 2000)

El siguiente esquema resume los tipos de datos espaciales.

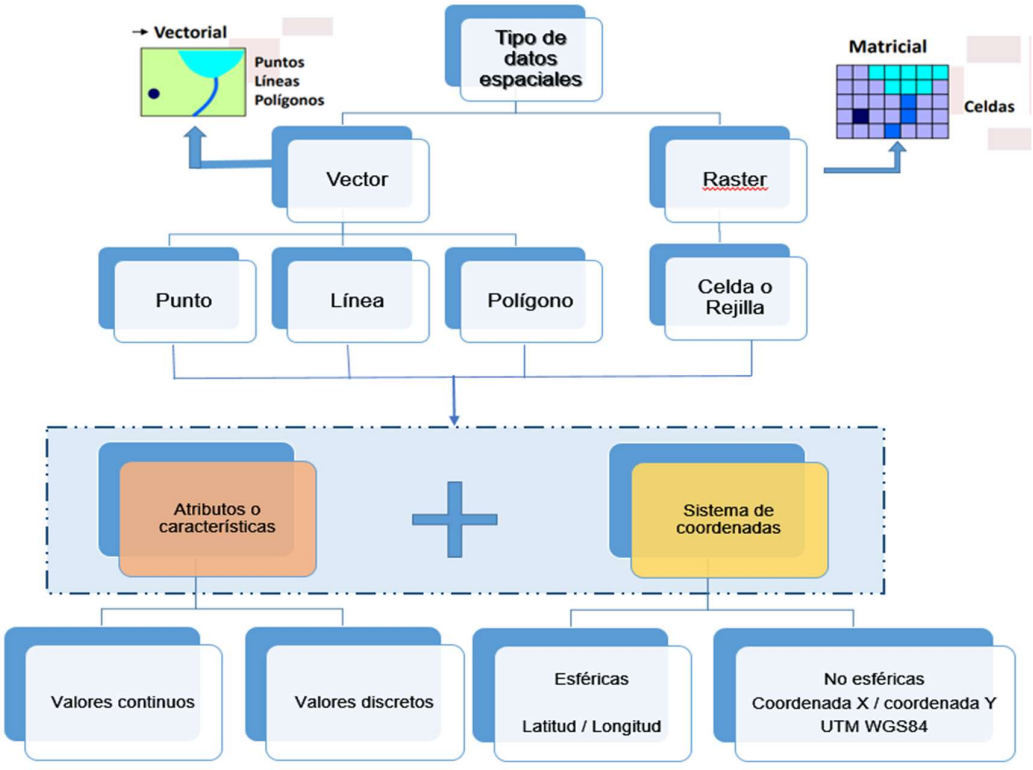


Figura 1. Esquema del tipo de datos espaciales

Para obtener la matriz de datos espaciales, se debe tener en consideración:

1. La elección de la representación del espacio geográfico y de los atributos que se incluirán y como se medirán.

2. La precisión de las mediciones tanto en coordenadas geográficas como en valores de atributo.

La figura 2 muestra esta relación y los términos que se usan para caracterizarla. (Haining, 2003)

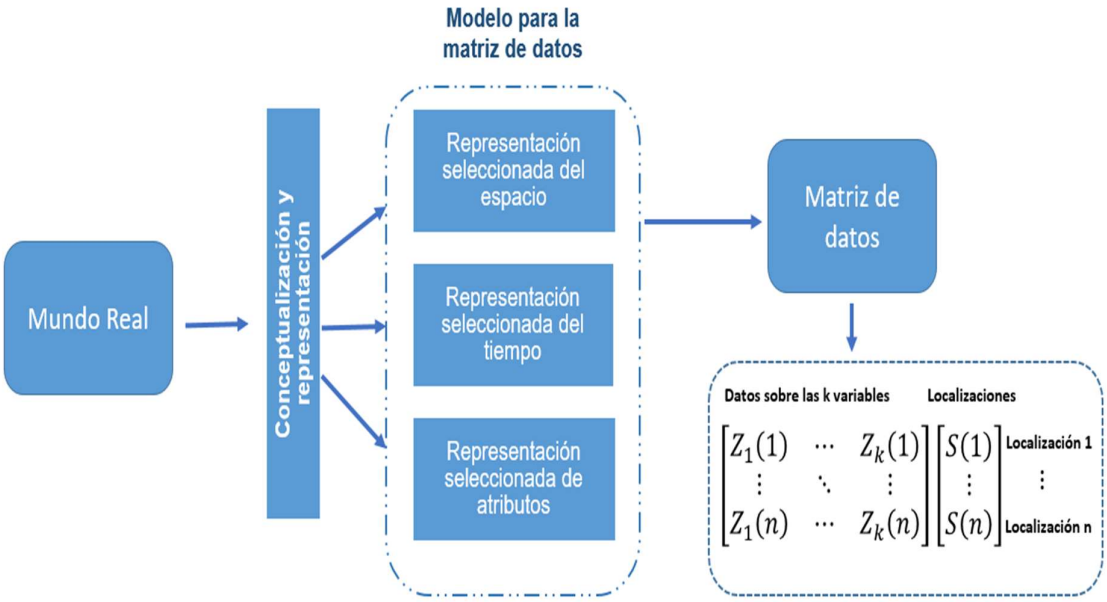


Figura 2. Del contexto geográfico a la matriz de datos

Por otro lado, es de mencionar que, en el contexto geográfico los datos faltantes son un problema, que pueden generar resultados poco confiables.

Hay algunos métodos disponibles para manejar conjuntos de datos con valores faltantes, como el de (Delchambre, 2015) que propone un método de

PCA basado en la diagonalización de la matriz de varianza – covarianza para problemas con datos ponderados o faltantes.

2.2 REVISIÓN DE LAS TÉCNICAS MULTIVARIANTES CON ENFOQUE GW

Las técnicas multivariantes con enfoque geográficamente ponderado GW han sido desarrolladas varias extensiones, siendo concebidas de técnicas estadísticas que son ampliamente usadas, tales como el análisis de componentes principales y la regresión, las mismas que se detallan a continuación:

Regresión ponderada geográficamente GWR propuesto por (Brunsdon et al., 1996) que permite conocer las distintas relaciones espaciales en diferentes puntos del espacio geográfico y sugiere que cualquier modelo que puede ser ponderado puede ser ponderado geográficamente; además, se tiene las estadísticas de resumen de GW (Brunsdon et al., 2002) el cual utiliza la estimación de la función de Kernel para ponderar geográficamente los puntos y obtener un resumen de las estadísticas y relaciones espaciales.

Así mismo, el análisis discriminante ponderado geográficamente GWDA (Brunsdon et al., 2007) esta técnica adapta el enfoque del modelo GWR permitiendo el modelado y la predicción de variables de respuesta categóricas.

De igual manera el modelo lineal generalizado ponderado geográficamente GWGLM, (Nakaya et al., 2009) que es una ampliación de GWR para la predicción de variables no continuas, permitiendo el uso de distintos modelos de regresión, tales como regresión logística para datos binarios, regresión de

Poisson para datos de conteo, que a través del predictor lineal estos modelos de regresión se integran como modelos lineales generalizados.

Finalmente, el GWPCA (Harris et al., 2011) evalúa la heterogeneidad y autocorrelación espacial para conocer la estructura espacial subyacente del conjunto de datos.

Para la aplicación de las técnicas antes descritas, fue desarrollado un paquete en el software estadístico R GWmodel. (Comber et al., 2020) Luego, se realizaron dos adaptaciones al modelo GWR, una con respecto a la temporalidad, que se denominó regresión ponderada geográfica y temporal GTWR (Fotheringham et al., 2015a) la cual analiza los efectos locales tanto en el espacio como en el tiempo, destacando la importancia de la temporalidad.

La otra con respecto a las escalas, es la regresión ponderada geográficamente multiescala MGWR que parte de la suposición que diferentes procesos operan a distintas escalas espaciales, para lo cual propone un vector de ancho de banda óptimo en el que cada elemento indica la escala espacial en la que tiene lugar un proceso particular. (Fotheringham et al., 2017)

Otras extensiones en los modelos de regresión ponderada geográficamente, son los que agregan el componente tiempo, es decir, además, de calcular los pesos para lo espacial, calcula los pesos temporales de un intervalo de tiempo, como lo propuesto por (Que et al., 2020) que calcula la tasa de variación entre los intervalos de tiempo.

El PCA es una de las técnicas multivariantes más utilizadas en la reducción de la dimensionalidad, por lo que se han generado varias alternativas para datos no espaciales y espaciales, como se muestran a continuación:

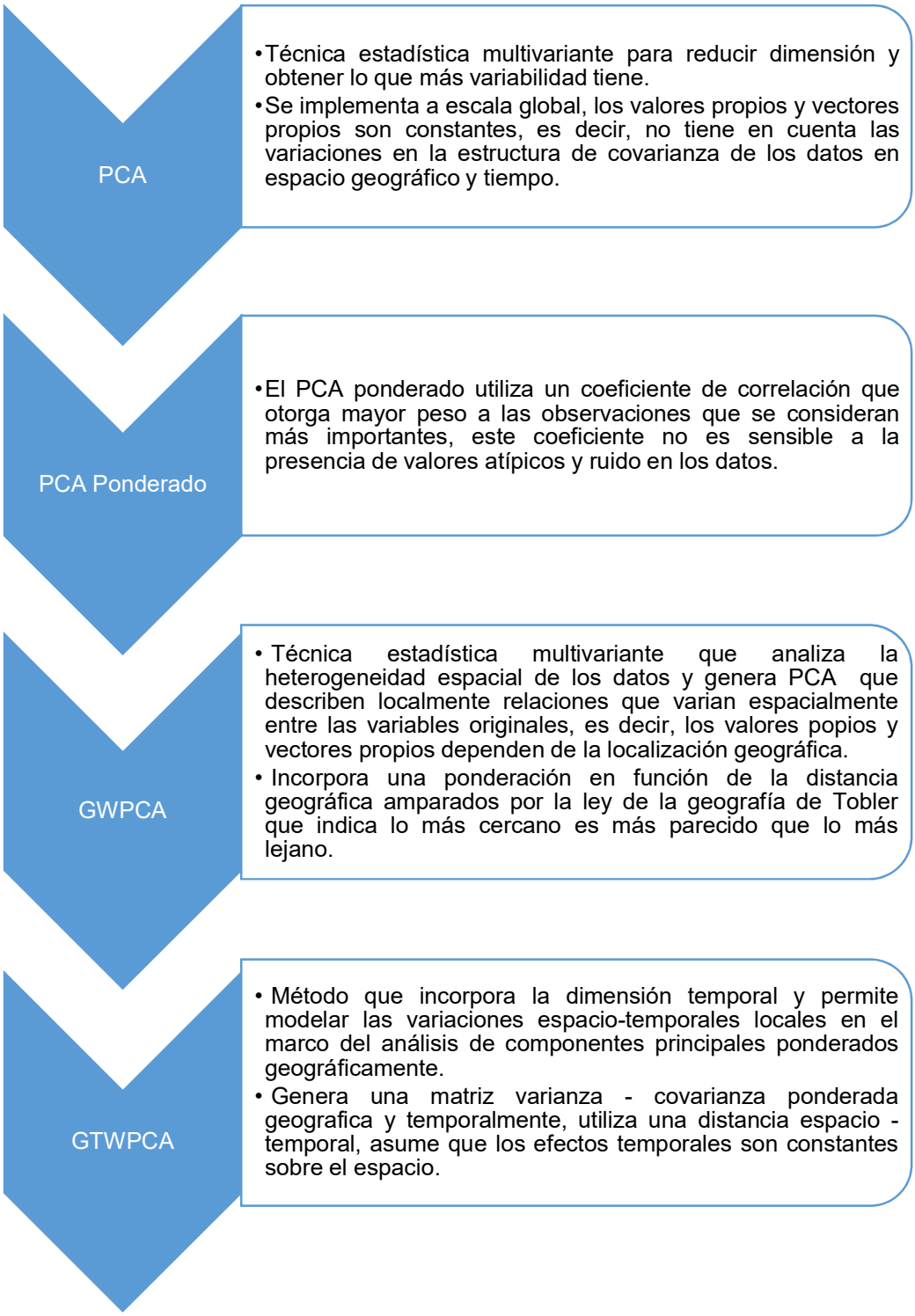


Figura 3. Alternativas del análisis de componentes principales - PCA

Además, otra extensión es GWnnegPCA propuesto por (Tsutsumida et al., 2022) donde presenta una opción para resolver el problema de patrones espaciales discontinuos de las cargas locales.

La revisión bibliográfica, (Charlton et al., 2010) (Demšar et al., 2013) (Lloyd, 2010) muestra a los modelos GWPCA como técnicas efectivas para el análisis de la heterogeneidad espacial, generando componentes principales locales que describen las relaciones entre las variables originales en cada ubicación y podrían también usarse, para construir índices compuestos locales, ha sido aplicado a diversas áreas, economía, social, ambiental, salud, entre otras, (Li et al., 2016) (Libório et al., 2022) (Han et al., 2022) (Perera et al., 2022)

2.3 TÉCNICAS BIPLOTS

Los métodos Biplot han sido fuente de continuas contribuciones a la ciencia desde su origen en 1971 hasta la fecha. Los distintos tipos de Biplots, sus usos y aplicaciones en otros contextos, han producido una gran diversidad de investigaciones en distintos ámbitos, enfocándose en la representación de matrices de datos para generar información multivariante.

(Gabriel, 1971) proporciona los métodos Biplot como una herramienta útil para el análisis y visualización de grandes matrices de datos, permitiendo representar gráficamente las filas y columnas de una matriz de datos sobre un espacio óptimo y de menor dimensión.

Denominaremos a X una matriz de orden (IxJ) y de rango r, donde las I filas corresponden a individuos y las J columnas a las variables, los métodos Biplot

consisten en aproximar la matriz de datos X por una de menor rango q, siendo q < r, a través de la descomposición en valores singulares (DVS) de X, para lo cual, considera dos factorizaciones denominadas GH-Biplot que consigue una alta calidad de representación de las columnas (variables) y el JK-Biplot que permite una alta calidad de representación en las filas (individuos).

(Galindo Villardón, 1986) propone el HJ-Biplot una representación simultánea de filas y columnas, elegidos de tal forma que puedan superponerse en el mismo sistema de referencia con una máxima calidad de representación, capturando la estructura subyacente y las relaciones entre individuos (filas) y variables (columnas) de la matriz de datos X. El HJ-Biplot ha sido aplicado ampliamente en distintos ámbitos, tales como:

En el campo de la medicina (Pedraz & Galindo, 1986) (Correa Londoño et al., 2007) (Cisneros et al., 2020; Miranda et al., 2022) (Riera-Segura et al., 2022).

En el campo de la biología (Pérez-Mellado & Galindo, 1986) (Fernández Gómez et al., 1996) (Herrera et al., 2014) (Jaramillo-Feijoo, Galindo-Villardon, Real-Cotto, et al., 2020);

En el campo de la economía (C. Santos et al., 1991) (Vicente et al., 1993) (Cabrera et al., 2006) (Mendes et al., 2009) (Díaz-Faes et al., 2013) (Gallego-Álvarez et al., 2015) (Amor-Esteban, García-Sánchez, et al., 2018) (Amor-Esteban, Galindo-Villardón, et al., 2018) (Medina-Hernández et al., 2023).

En el campo de la educación (Díaz-Faes et al., 2013) (Andrade-Sánchez et al., 2015) (García & Villardón, 2018) (González-García et al., 2019) (Ruiz-Toledo et al., 2021) (Vairinhos et al., 2022).

En el campo ambiental (Gallego-Álvarez et al., 2014) (Martínez-Hidalgo et al., 2014) (Carrasco et al., 2019); en el campo agrícola (Valenzuela-Cobos et al., 2022) (Valenzuela-Cobos et al., 2023).

En el campo social (Tavera et al., 1994) (M. P. Galindo Villardón et al., 2007 (Vázquez-Pérez et al., 2011) (Rodríguez et al., 2014) (Jaramillo-Feijoo, Galindo-Villardon, & Real-Cotto, 2020); en transporte (Frutos Bernal et al., 2020)

(Villardón, 1992) propone una alternativa a las técnicas factoriales mediante una generalización de los métodos Biplot

(Martín-Rodríguez et al., 2002) propone una integración de subespacios desde una perspectiva Biplot. (Amaro et al., 2004) es una generalización del Manova-Biplot de modelos lineales generales multivariantes para diseños de dos vías.

(González & Villardón, 2013) propone un método de biplot robusto.(Álvarez & Villardon, 2015) propone un análisis espacio-temporal de matrices de tráfico utilizando HJ-Biplot.

(Caballero-Juliá et al., 2017) propone una herramienta para el análisis de contenido combinando los métodos JK-Meta_Biplot y STATIS Dual.

(Hernández Suárez et al., 2016) propone un nuevo enfoque denominado HJ-Biplot composicional. (Nieto-Librero et al., 2017) propone un nuevo algoritmo denominado Clustering Disjoint HJ-Biplot.

(Cubilla-Montilla et al., 2021) propone una modificación del Hj-Biplot para datos masivos denominada Biplot HJ disperso. (González-García et al., 2021) propone un nuevo algoritmo matemático para biplots restringidos orotogonales y dispersos, llamado C_{enet}Biplots

(Pilacuan-Bonete et al., 2022) propone una herramienta HJ-Biplot por asignación latente de Dirichlet (LDA) para análisis de contenido.

Cuando los datos son binarios, los métodos Biplots antes descritos no son adecuados, por lo que (Vicente-Villardón et al., 2006a) propone un Biplot logístico donde cada individuo se representa como un punto y cada carácter como una dirección a través del origen.

La proyección de un punto sobre la dirección del carácter predice la probabilidad de presencia de ese carácter. Luego este método se amplía mediante un enfoque integrado como lo propuso (Demey et al., 2008a), dicho método comprende un análisis de coordenadas principales y una regresión logística para construir un Biplot logístico externo. Se ha realizado diversas aplicaciones del biplot logístico, como se indican a continuación:

La evaluación de las empresas a la sostenibilidad y su relación con el desempeño económico (Galindo et al., 2011); el análisis de empresas internacionales que divulgan indicadores sobre emisiones de gases de efecto invernadero (Gallego-Álvarez & Vicente-Villardón, 2012).

(Vicente-Villardón & Sánchez, 2014) propone un biplots logístico para datos ordinales con aplicación a la satisfacción laboral de los doctores en España.

Análisis de indicadores de sostenibilidad del Global Reporting Initiative. Una mirada desde del Biplot logístico (Montilla et al., 2015); Empresas innovadoras detrás de las regiones: análisis del rendimiento de la innovación en Portugal mediante Biplots logísticos externos (Noronha Vaz et al., 2015); Estudio de la sostenibilidad de las empresas mexicanas utilizando Biplot Logístico Externo (Murillo, 2015).

Cómo las corporaciones manejan los informes de sostenibilidad: evaluación utilizando el enfoque Biplot Logístico multicriterio; (Vicente Galindo et al., 2015); Caracterización de germoplasma de maíz local a través de marcadores SSR asistido por Biplot Logístico Externo (BLE) (Cañizares et al., 2016).

(Hernández-Sánchez & Vicente-Villardón, 2017) propone un biplot logístico para datos nominales; (Vicente-Villardón & Hernández-Sánchez, 2020) propone un Biplot logístico externo para tipo mixto de datos.

(Martínez-Regalado et al., 2021) propone las técnicas HJ-Biplot y Biplot logístico externo para aprendizaje automático; Métodos multivariantes para evaluar tumores neuroendocrinos (Montes Escobar, 2022)

2.4 PRUEBA ESTADÍSTICA NO PARAMÉTRICA MANN-KENDALL

El comportamiento de una serie temporal a largo plazo se denomina tendencia, existen métodos estadísticos que permiten su análisis, uno de los más utilizados es la prueba estadística no paramétrica Mann-Kendall (MK) basada en rangos (Mann, 1945) (Kendall, 1975) se ha utilizado ampliamente para explorar la importancia de las tendencias monótonas crecientes o decrecientes en serie temporales hidrometerológicas, como la calidad del agua, el caudal, la temperatura y la precipitación., aunque la prueba se usa ampliamente, se desconoce el poder que tiene la prueba MK para la detección de tendencias en diversas áreas.

La principal razón para usar pruebas estadísticas no paramétricas es que, los datos no deben seguir una distribución normal, existen varios estudios del uso de la prueba MK para detectar tendencias en series temporales hidrológicas e hidrometereológica, tales como: (Hipel et al., 1988), (Yue & Wang, 2004), entre otros.

La hipótesis nula a contrastar es H_o: no hay tendencia en los datos y la hipótesis alternativa $H_{1:}$ existe una tendencia en el conjunto de datos, se calcula el valor estadístico de la prueba de Mann-Kendall que viene representado por **Z.**

Los valores positivos de Z indican tendencias crecientes en la serie temporal y los valores negativos, tendencias decrecientes. Cuando |Z| > $Z_{1-\alpha/2}$, se rechaza la hipótesis nula y se dice que existe una tendencia significativa en la serie temporal al nivel de significancia de α. El término $Z_{1-\alpha/2}$ es el valor crítico de Z de la tabla normal estándar (en α = 0,05, $Z_{1-\alpha/2} = 1{,}96$)

Así también, el valor de p, se compara con el nivel de significancia que por lo general es α = 0,05 y si éste es menor al nivel de significancia definido se rechaza el Ho.

(Yue et al., 2002) realizó un análisis comparativo de las pruebas de Mann-Kendall y Spearman para detectar tendencias monótonas en series hidrológicas, evidenciándose que el poder que tiene la prueba MK va en relación de la función creciente de la pendiente de la tendencia, el tamaño de la muestra y el nivel de significación preasignado; mientras que es una función decreciente de la variación de la serie temporal.

La prueba Man-Kendall, ha sido aplicada en distintos ámbitos pero mayormente en datos ambientales, climatológicos, imágenes de tipo ráster y de salud.(Ozocak et al., 2023) (Sodagari & Varga, 2023), a continuación se presentan algunas aplicaciones y modificaciones:

(Hamed & Ramachandra Rao, 1998) propone una modificación a la prueba de tendencia Mann-Kendall cuando existe autocorrelación en los datos.

(Neeti & Eastman, 2011) propone un enfoque de Mann-Kendall para evaluar la tendencia en series temporales de imágenes de tipo ráster; (Drápela & Drápelová, 2011) propone una aplicación de la prueba Mann-Kendall y las estimaciones de la pendiente de Sen.

(Mondal et al., 2012) realizó un análisis de tendencia de precipitaciones mediante la prueba de Mann-Kendall: un estudio de caso del distrito de Cuttack, Orissa.

(Karmeshu, 2012) realizó un estudio para detectar tendencias en la temperatura y la precipitación anual mediante la prueba de Mann-Kendall en estados del noreste de los Estados Unidos.

(Urresti Estala et al., 2012) realizó una evaluación de tendencias de contaminantes en la masa de agua al sur de España mediante el test estadístico de Mann-Kendall, concluyendo que es una herramienta útil para obtener una visión en conjunto sobre la evolución en el tiempo de la calidad de las aguas subterráneas.

(Fathian et al., 2016) propone un estudio de tendencia de las variables hidrológicas y climáticas afectadas por cuatro variaciones del enfoque Mann-Kendall en la cuenca del lago Urmia, Irán, evidenciando que al incorporar los distintos enfoques al estudio, proporcionan mejor información sobre la cantidad de estaciones que muestran tendencias significativas de datos hidroclimáticos.

(García-Garizábal, 2017) propone un análisis de variabilidad y tendencia en la costa de Ecuador, mediante la prueba de Mann-Kendall.

(Sa'adi et al., 2019) realizó un análisis de tendencia de precipitaciones y lluvias extremas en Sarawak, Malasia, utilizando la prueba de Mann-Kendall modificada; (Grassi et al., 2019) realizó un análisis espacio temporal de la homogeneidad de estaciones de precipitación en una zona árida y semiárida de Venezuela.

(Villerias Salinas et al., 2020) realizó un análisis espacial de vulnerabilidad y riesgo en salud por COVID-19 mediante la prueba estadística Mann-Kendall, se identificaron tendencias en los datos.

(V. C. dos Santos et al., 2023) realizó una evaluación de tendencia del inicio, fin, duración y precipitación total de la temporada de lluvia de Palmas – Brasil, encontrándose cambios en las tendencias anuales de la estación lluviosa.

(Isensee et al., 2023) propone un análisis de series temporales de caudales extremos: tendencias, longitud de registro y persistencia, mediante la prueba estadística de Mann-Kendall.

2.5 ANÁLISIS ESPACIO TEMPORAL EN EL CAMPO DE LA SALUD

El análisis de la triada epidemiológica es tiempo, lugar y persona. Los análisis espacial y temporal son de interés en las investigaciones en salud, siendo el estudio de las variaciones geográficas un componente epidemiológico importante. (Waller & Gotway, 2004)

En el contexto de salud, los análisis espacio temporales, permiten estudiar de manera sistémica un fenómeno en particular, donde la mayoría de actores, agentes, factores ambientales que intervienen se localizan en tiempo y espacio y muchas relaciones se basan en la proximidad, según la primera ley de Tobler “Todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas lejanas”. (Tobler, 1970) Por lo que tener en cuenta la localización se vuelve esencial para entender el fenómeno en salud. (Souris, 2019)

Uno de los campos es la epidemiología espacial que describe la variación espacial en relación al riesgo de enfermedad, identifica la correlación geográfica de los factores de riesgo en relación con los resultados de salud medidos en un entorno geográfico, patrones de distribución espacial y su evolución en el tiempo, resultando análisis mucho más enriquecedores,

(Souris, 2019) (Pou et al., 2019) sin embargo, es un desafío incorporar la dimensión temporal en los análisis espaciales, por la complejidad de los modelos espaciotemporales. (Fotheringham et al., 2015b)

Se presentan algunos estudios relacionados con los análisis espacial y temporal y las técnicas que han sido utilizadas:

(Núñez-González et al., s. f.) realizó un estudio de tendencia de mortalidad por enfermedades cerebrovasculares en Ecuador 2001-2015, utilizando el modelo de regresión joinpoint.

(Núñez-González, Delgado-Ron, et al., 2018) realizó un estudio sobre las tendencias y patrones espaciales de la mortalidad por cáncer bucal en Ecuador, 2001 – 2016, utilizando un modelo de regresión joinpoint y el Índice Global de Moran.

(Núñez-González, Aulestia-Ortiz, et al., 2018) propone un estudio de tendencia de mortalidad por enfermedades isquémicas del corazón en Ecuador, 2001 – 2016, para analizar los cambios en la tendencia utilizó el análisis de regresión de punto de unión.

(Núñez González et al., 2018) realizó un estudio sobre los cambios en la tendencia temporal de mortalidad por cáncer de mama en Ecuador, 2001 – 2016, utilizando el análisis de regresión joinpoint, y concluye que se evidencia un incremento estadísticamente significativo de las muertes por esta enfermedad, duplicando los años potenciales de vida perdidos.

(Núñez-González et al., 2019) propone un análisis espacial de la mortalidad por dengue, cisticercosis y enfermedad de Chagas en Ecuador en un periodo

2011 al 2016, utilizando el índice de Morán Global para evaluar la autocorrelación espacial y la formación de conglomerados por el índice local de asociación espacial.

(Núñez-González et al., 2020) propone un análisis de tendencia y análisis espacio-temporal de la mortalidad por diabetes mellitus en Ecuador, 2001-2016, utilizando análisis de regresión de punto de inflexión para el análisis de tendencia e identificó conglomerados espacio-temporales.

(Lalangui et al., 2022) realizó un estudio espacio-temporal dinámico de la mortalidad infantil en Ecuador, 2010 – 2019, donde propone un método de priorización que combina un método para detectar clusters (LISA – Indicador local de asociación espacial) y la prueba estadística Mann-Kendall para identificar tendencia temporal monótona.

Basado en esta revisión bibliográfica se pone de manifiesto la necesidad de herramientas innovadoras que integren los componentes espacial y temporal, por lo que se propone en esta investigación, el método Biplot logístico ponderado espacio temporal (TSWLB) como una metodología de priorización de las unidades espaciales. En la siguiente gráfica se resumen las aportaciones de los métodos Biplots incluido el propuesto en esta investigación.

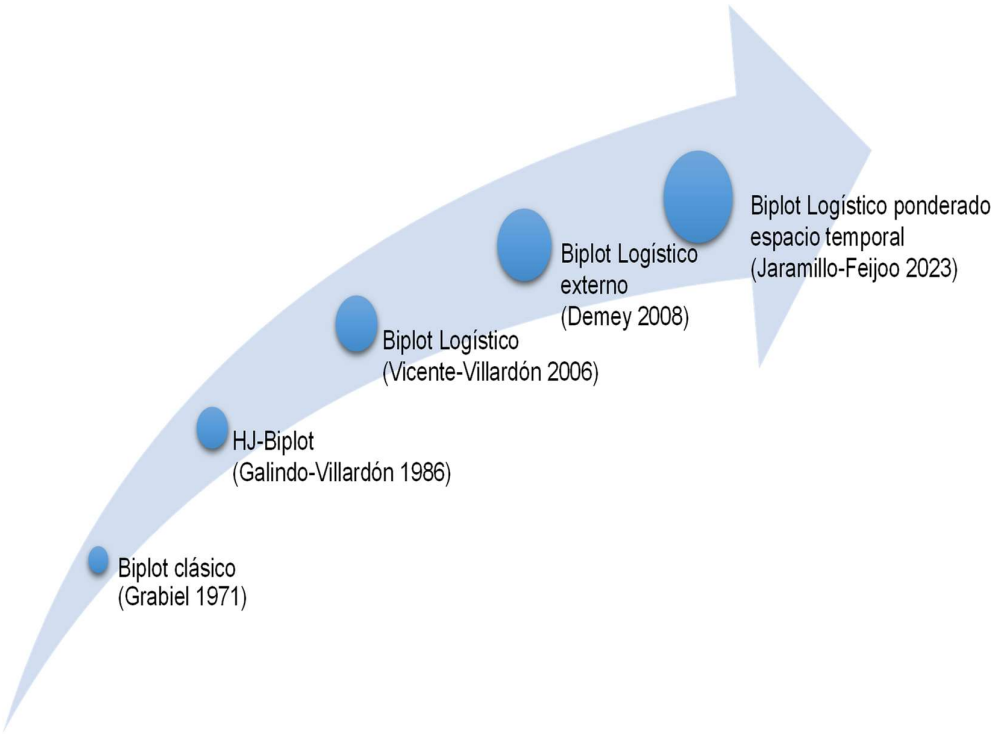


Figura 4. Aportaciones de los métodos Biplots

CAPÍTULO III

3. MÉTODOLOGÍA

MÉTODO PROPUESTO. BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB)

En el presente capítulo, se presenta una propuesta de una técnica multivariante espacio temporal que hemos denominado Biplot logístico ponderado espacio temporal (TSWLB), que combina los métodos multivariantes GWPCA para el componente espacial y la prueba estadística no paramétrica Mann-Kendall para el componente temporal y se representan gráficamente mediante un Biplot logístico externo.

Este método permite capturar y representar los componentes espacial y temporal de forma simultánea dando un ordenamiento a los sitios geográficos en función de la presencia de las características analizadas, generando clústeres prioritarios.

3.1 MÉTODO BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB)

Se tiene una matriz de datos X_{np} conformada por unidades espaciales (filas) y unidades de tiempo (columnas), es decir, las filas corresponden a sitios geográficos y las columnas corresponden a variables medidas para cada sitio geográfico y en distintos años y las dos últimas columnas corresponden a las coordenadas geográficas X y Y. .

Figura 5. Esquema de matriz de datos espaciales

En la Fig. 6 se muestra el esquema con las etapas para desarrollar la técnica TSWLB

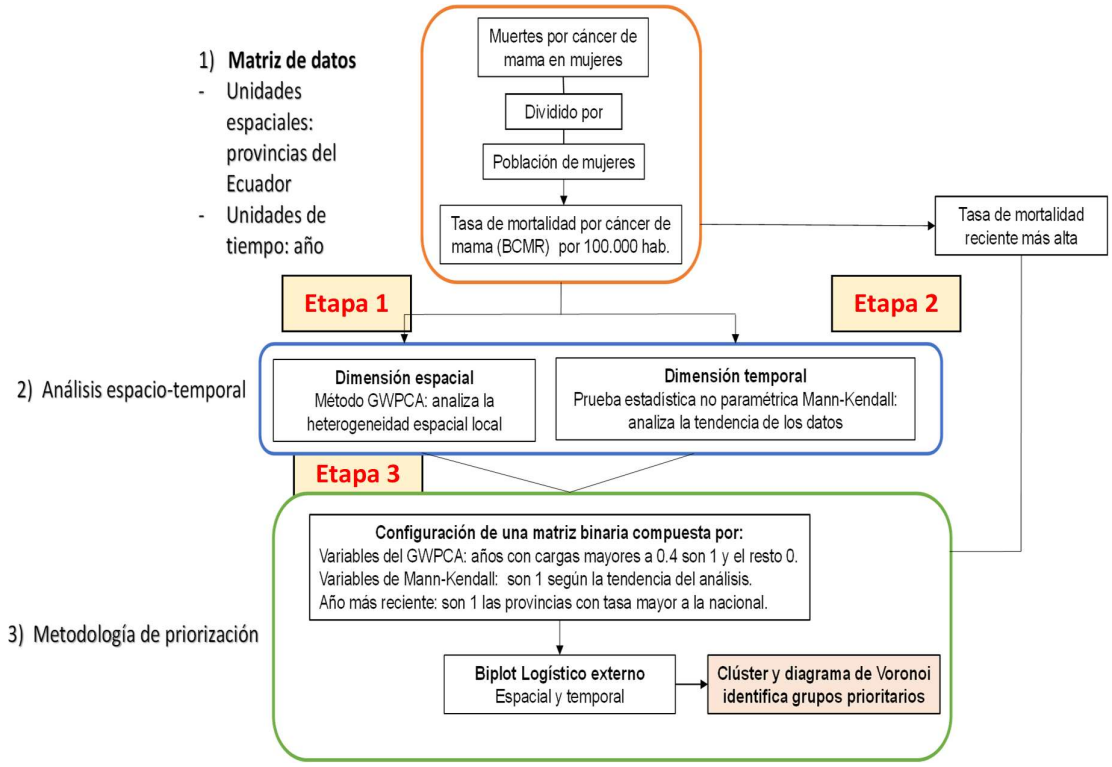


Figura 6. Esquema del Biplot logístico ponderado espacio temporal (TSWLB)

3.2 ETAPA 1: ANÁLISIS DE LA COMPONENTE ESPACIAL

La técnica GWPCA examina la parte no estacionaria de los datos localmente en el espacio geográfico, así tenemos, para GWPCA un vector de variables observadas x_i en la ubicación espacial i se asume que sigue una distribución normal multivariante con vector media μ y matriz de varianza-covarianza ∑, tal que $x_i \sim N(\mu, \sum)$.

Además, la ubicación espacial i tiene coordenadas (u,v), entonces el PCA con efectos geográficos locales implica considerar a x_i como condicional en u y v, y haciendo a μ y ∑ funciones de u y v; por lo tanto, $x_i|(u,v) \sim N\left(\mu(u,v), \sum(u,v)\right)$, así μ y ∑ son funciones de u y v, esto implica que cada elemento de μ (u, v) y ∑(u, v) es también función de u y v.

Por lo tanto los momentos μ(u,v) y ∑(u,v) son el vector de media geográficamente ponderados (GW) y la matriz de varianza-covarianza geográficamente ponderado, respectivamente. Para obtener los componentes principales geográficamente ponderados la descomposición de la matriz de varianza-covarianza GW provee los valores propios GW y los vectores propios GW.

El producto de la i fila de la matriz de datos con los vectores propios GW para la ubicación i provee la fila i de las puntuaciones de los componentes GW. La matriz de varianza-covarianza GW es;

$$\sum(u,v) = X^T \, W(u,v) X$$

Donde W(u,v) es una matriz diagonal de pesos geográficos que puede ser generada usando una función de kernel. En el caso de estudio, se utilizó una función de kernel bi-square:

$$w_{ij} = \left(1 - \left({}^{d_{ij}}/_{r}\right)^2\right)^2 \quad si\ \ d_{ij} \leq r; \quad w_{ij} = 0 \quad en\ otro\ caso$$

Donde el ancho de banda es la distancia geográfica r y d_{ij} es la distancia entre la ubicación espacial de la i y j filas en la matriz de datos X. Los componentes principales GW para la ubicación (u_i, v_i) puede escribirse como:

$$LVL^T|(u_i, v_i) = \sum(u_i, v_i)$$

Donde $= \sum(u_i, v_i)$ es la matriz de varianza-covarianza GW para la ubicación (u_i, v_i).

En el siguiente esquema, se resume los pasos que deben realizarse para aplicar la técnica GWPCA.

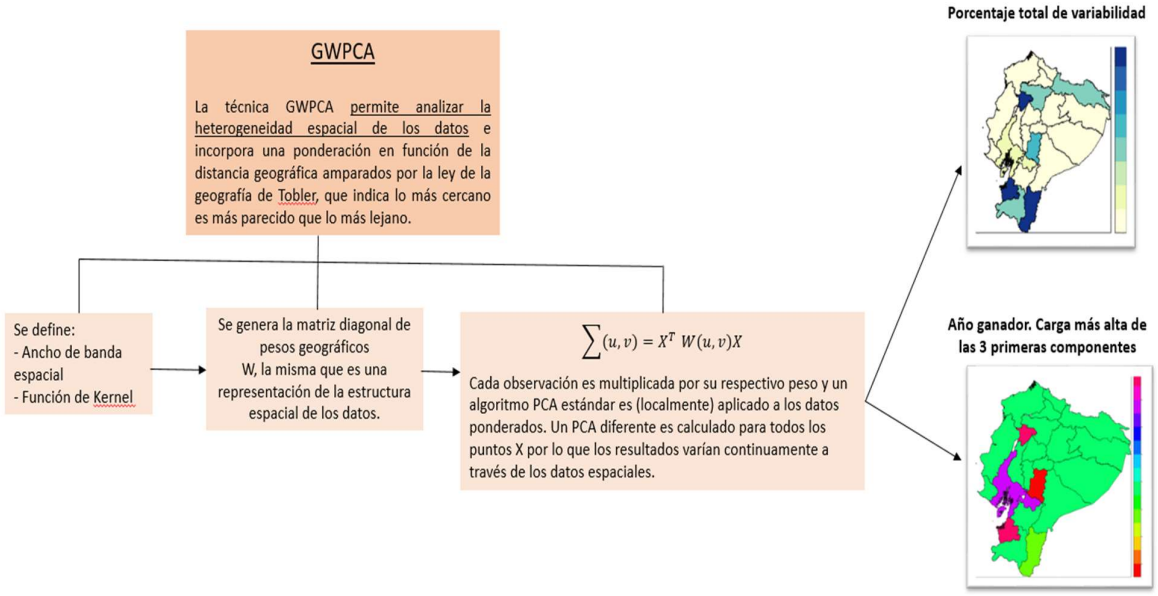


Figura 7. Esquema de la técnica Análisis de las componentes principales geográficamente ponderado (GWPCA)

3.2.1 Configuración de la matriz binaria

La técnica Análisis de componentes principales geográficamente ponderado (GWPCA), calcula un PCA estándar para cada sitio geográfico, generando n PCA locales, para resumir esta información y obtener las que están más correlacionados, se escogen las cargas máximas de los tres primeros componentes principales y se lo define con la variable o año ganador, obteniéndose una matriz con cargas máximas C_{np} dicha matriz luego se transforma a una matriz binaria con el siguiente criterio las cargas mayores a 0.4 son 1 y resto cero.

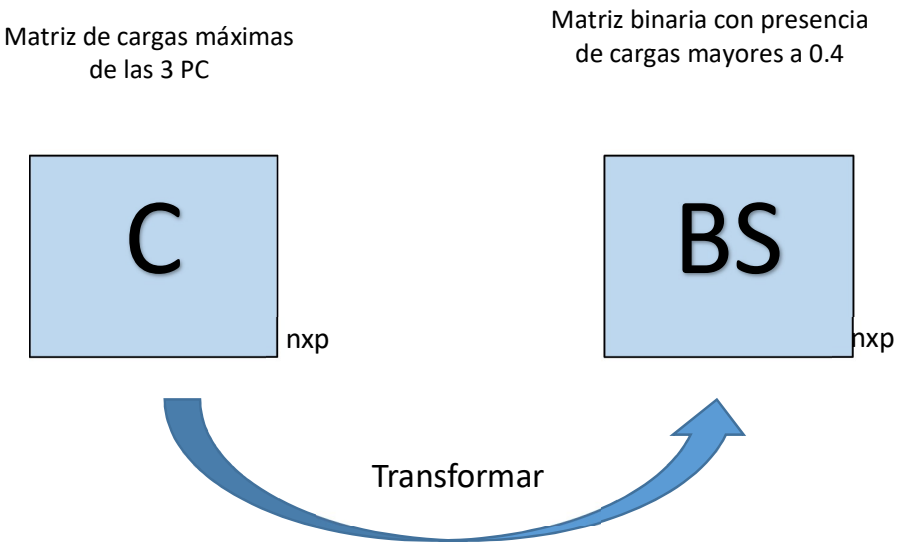


Figura 8. Estructura de la matriz binaria espacial

3.2.2. Herramienta informática en lenguaje R

El paquete GWModel, se puede utilizar para aplicar la técnica multivariante GWPCA, a continuación, se detalla el algoritmo y código fuente:

3.2.2.1 Algoritmo del GWPCA

1. Lectura de la matriz de datos espaciales.
2. Estandarización de la matriz de datos
3. Calculo del PCA estándar
4. Cálculo del ancho de banda óptimo

5. Definición de la función Kernel a utilizar
6. Cálculo del GWPCA
7. Calculo del porcentaje de variabilidad total de las 3 primeras componentes
8. Generación de la matriz de cargas.
9. Generación de la máxima carga de los 3 CP. (variable o año ganador)
10. Generación del mapa del porcentaje de variabilidad local.
11. Generación del mapa de la variable o año ganador.

3.2.2.2 Código fuente del GWPCA

```
libraries("sp", "rgdal",
"raster","lattice","latticeExtra","sf","RColorBrewer","robustbase","Rcpp","spDa
ta","Matrix","spatialreg","maptools","GWmodel","spDataLarge","rgeos",'geosp
here')

library(openxlsx)

library(sf)

library(ggplot2)


# Lectura de la matriz de datos y coordenadas geográficas.

datos.cancer <- read.xlsx("Matriz_07_21.xlsx", sheet = 1)

datos.cancer <- datos.cancer %>%

  mutate(PROV=c("AZUAY",

                "BOLIVAR",
                "CAÑAR",
                "CARCHI",
                "COTOPAXI",
                "CHIMBORAZO",
```

```
"EL ORO",
"ESMERALDAS",
"GUAYAS",
"IMBABURA",
"LOJA",
"LOS RIOS",
"MANABI",
"MORONA SANTIAGO",
"NAPO",
"PASTAZA",
"PICHINCHA",
"TUNGURAHUA",
"ZAMORA CHINCHIPE",
"GALAPAGOS",
"SUCUMBIOS",
"ORELLANA",
"SANTO DOMINGO DE LOS TSACHILAS",
"SANTA ELENA"))

datos.cancer <- datos.cancer %>% set_rownames(datos.cancer$PROV)

# CARGAMOS DATOS SHP PARA ECUADOR

Prov_Ecu<-
readOGR("C:/Users/Leyda/OneDrive/Escritorio/1.Doctorado_Salamanca/4.
Tesis/R-GWHJ-
Biplot/Analisis_2007_2021/GWPCA_CLUSTER_MK/SHP_2015/nxprovincias
.shp")

Prov_Ecu <- Prov_Ecu[!(Prov_Ecu@data$DPA_PROVIN==90) ,]

Prov_Ecu <- Prov_Ecu[order(as.numeric(Prov_Ecu@data$DPA_PROVIN)),]

Prov_Ecu$DPA_PROVIN <- as.numeric(Prov_Ecu$DPA_PROVIN)
```

```
# Estandarizamos la data (escalamos y centralizamos)
Data.scaled <- scale(as.matrix(datos.cancer[,4:18]))

# Coordenadas Provincias.
Coords <- as.matrix(cbind(datos.cancer$X, datos.cancer$Y))

# Creamos SPDF
Data.scaled.spdf <- SpatialPointsDataFrame(Coords,
                                + as.data.frame(Data.scaled))

# MODELO PCA CON DATOS ESCALADOS
pca.basic <- princomp(Data.scaled, cor = FALSE)
pca.basic$loadings
pca.basic$scores
pca_loadings <- pca.basic$loadings
write.csv2(pca_loadings,file="pcaloadigsmama10mayo.csv")
pca_scores <- pca.basic$scores
write.csv2(pca_scores,file="pcascoresmama10mayo.csv")
pca.basic$sdev
(pca.basic$sdev^2 / sum(pca.basic$sdev^2)) * 100
dt1 <-  as.data.frame(unclass(pca.basic$loadings))
Biplot(pca.basic)
correlaciones=cor(Data.scaled)
write.csv2(correlaciones,file="correlacioesmama10mayo.csv")
```

#Kernel bandwidths for GW PCA

#Función para el bandwidth bw.gwpca con k=3. k es el número de componentes.

#Bi-square

```
bw.gwpca.basic <- bw.gwpca(Data.scaled.spdf, vars = colnames(
 + Data.scaled.spdf@data), k = 3, robust = FALSE, kernel = "bisquare", adaptive = TRUE)
bw.gwpca.basic
```

#Box-Car

```
bw.gwpca.box <- bw.gwpca(Data.scaled.spdf, vars = colnames(
 + Data.scaled.spdf@data), k = 3, robust = FALSE, kernel = "boxcar", adaptive = TRUE)
bw.gwpca.box
```

 # Se utiliza el bandwidth para calibrar el Modelo GWPCA

```
gwpca.basic <- gwpca(Data.scaled.spdf, vars = colnames(
 + Data.scaled.spdf@data), bw = bw.gwpca.basic , k = 3, robust = FALSE, adaptive = TRUE)
gwpca.basic
str(gwpca.basic)
loading_pca_geo <- data.frame(gwpca.basic$loadings)
write.csv2(loading_pca_geo,file = "loadingsGWPCABW22mama10mayo.csv")
```

#Empecemos con la varianza (PTV)

```
prop.var <- function(gwpca.obj, n.components) {return((rowSums(gwpca.obj$var[, 1:n.components]) / rowSums(gwpca.obj$var)) * 100)}

var.gwpca.basic <- prop.var(gwpca.basic, 3)

Prov_Ecu$var.gwpca.basic <- var.gwpca.basic

Prov_Ecu@data

write.csv2(Prov_Ecu@data,file = "Pro.var.gwpcamama10mayo.csv")
```

 # GUARDAMOS NUEVO ARCHIVO SHAPEFILE

```
writeOGR(Prov_Ecu,
"C:/Users/Leyda/OneDrive/Escritorio/1.Doctorado_Salamanca/4. Tesis/R-GWHJ-Biplot/Analisis_2007_2021/SHP_GWPCA_CANCER",
"ecuador_GWPCAma10mayo",
      driver = "ESRI Shapefile") #also you were missing the driver argument
```

 # ESCALAS

```
map.na = list("SpatialPolygonsRescale", layout.north.arrow(),offset = c(329000, 261500), scale = 4000, col = 1)

map.scale.1 = list("SpatialPolygonsRescale", layout.scale.bar(), offset = c(326500, 217000), scale = 5000, col = 1, fill = c("transparent", "blue"))

map.scale.2 = list("sp.text", c(326500, 217900), "0", cex = 0.9, col = 1)

map.scale.3 = list("sp.text", c(331500, 217900), "5km", cex = 0.9, col = 1)

map.layout <- list(map.na, map.scale.1, map.scale.2, map.scale.3)
```

#colores

```
library(RColorBrewer)

mypalette.4 <- brewer.pal(8, "YlGnBu")
```

#Mapa-Resultado

```
spplot(Prov_Ecu, "var.gwpca.basic", key.space = "right", col.regions = mypalette.4, cuts = 7, sp.layout = map.layout, main = "PTV for local components 1 to 3 (basic GW PCA) BW=22 (K=3)")


#----
```

#Variable ganadora

#Colores

```
mypalette.5 <- rainbow(15)

loadings.pc.basic <- gwpca.basic$loadings[, , 3]

win.item.basic = max.col(abs(loadings.pc.basic))

Prov_Ecu$win.item.basic <- win.item.basic

p<-loadings.pc.basic

write.csv2(p,file = "variablewinmama10mayo3PC.csv")

spplot(Prov_Ecu, "win.item.basic", key.space = "right", col.regions = mypalette.5, at = seq(1:16), main = "Winning año: highest abs BW=22 (K=3)", sp.layout = map.layout)

)),main = "Winning variable: highest abs BW=22 (K=3)", sp.layout = map.layout)
```

3.3 ETAPA 2: ANÁLISIS DE LA COMPONENTE TEMPORAL

La prueba estadística no paramétrica Mann-Kendall examina la tendencia de una serie de datos temporales, una de las ventajas es que se basa en la importancia de las diferencias o variaciones y no directamente a los valores aleatorios, por lo que no es afectado por valores atípicos. (Mann, 1945) (Kendall, 1975)

La prueba estadística no paramétrica Mann-Kendall es también llamada prueba tau de Kendall y ha sido aplicada en muchos estudios para identificar si existen tendencias monótonas creciente o decreciente.

Esta prueba no requiere que los datos sigan alguna distribución en particular y una de sus propiedades es no hacer suposiciones de los datos sino comprobarlos, por lo que la hipótesis nula a contrastar es H_o: no hay tendencia en los datos y la hipótesis alternativa H_1: existe una tendencia en el conjunto de datos.

Se define un vector X que tiene una serie temporal de observaciones, tales que, $X = \{x_1, x_2, \ldots, x_n\}$. La prueba realiza combinaciones de cada par de datos ordenados de forma secuencial en el tiempo y comprueba si $X_j > X_i$, o $X_j < X_i$ contabilizando el número de pares que incrementan o decrecen en el tiempo.

El cálculo de la prueba Mann-Kendal **S** expresa la frecuencia de incrementos menos la frecuencia relativa de decrementos, siendo calculado para cada

unidad espacial. La fórmula de cálculo de la prueba de Mann-Kendall S es como sigue:

$$\mathbf{S} = \frac{2(\mathrm{t}-2)!}{\mathrm{t}!} \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} sing(x_j - x_i)$$

Siendo x_j y x_i valores de datos secuenciales; j más grande que i y n la longitud del conjunto de datos. (Lalangui et al., 2022)

La función signo está dada por:

$$Sign(x_j - x_i) = \begin{cases} +1 \; si \; (x_j - x_i) > 0 \\ 0 \; si \; (x_j - x_i) = 0 \\ -1 \; si \; (x_j - x_i) < 0 \end{cases}$$

x_i es el tiempo en i ϵ {1,2,….t - 1} y x_j es el tiempo en j = (i + 1) ϵ {1,2,….,t}.

La varianza de **S** está determinada por:

$$VAR(S) = \frac{1}{18}(n(n-1)(2n+5))$$

El resultado de **S** indica el tipo de tendencia, siendo **S** diferente de cero, el H_o puede ser rechazada y H_1 aceptada.

El valor estadístico de la prueba de Mann-Kendall está representado por **Z** y se expresa en la siguiente ecuación:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var(S)}} & ; \quad S > 0 \\ 0 & ; \ S = 0 \\ \dfrac{S+1}{\sqrt{Var(S)}} & ; \quad S < 0 \end{cases}$$

La existencia de una tendencia estadísticamente significativa es evaluada por el valor **Z**, si es positivo indica que hay tendencia creciente y si es un valor negativo indica que hay tendencia decreciente, para un nivel de significancia, por lo general se utiliza un α = 0,05, donde el criterio es si el valor p menor al nivel de significancia, se rechaza H_o.

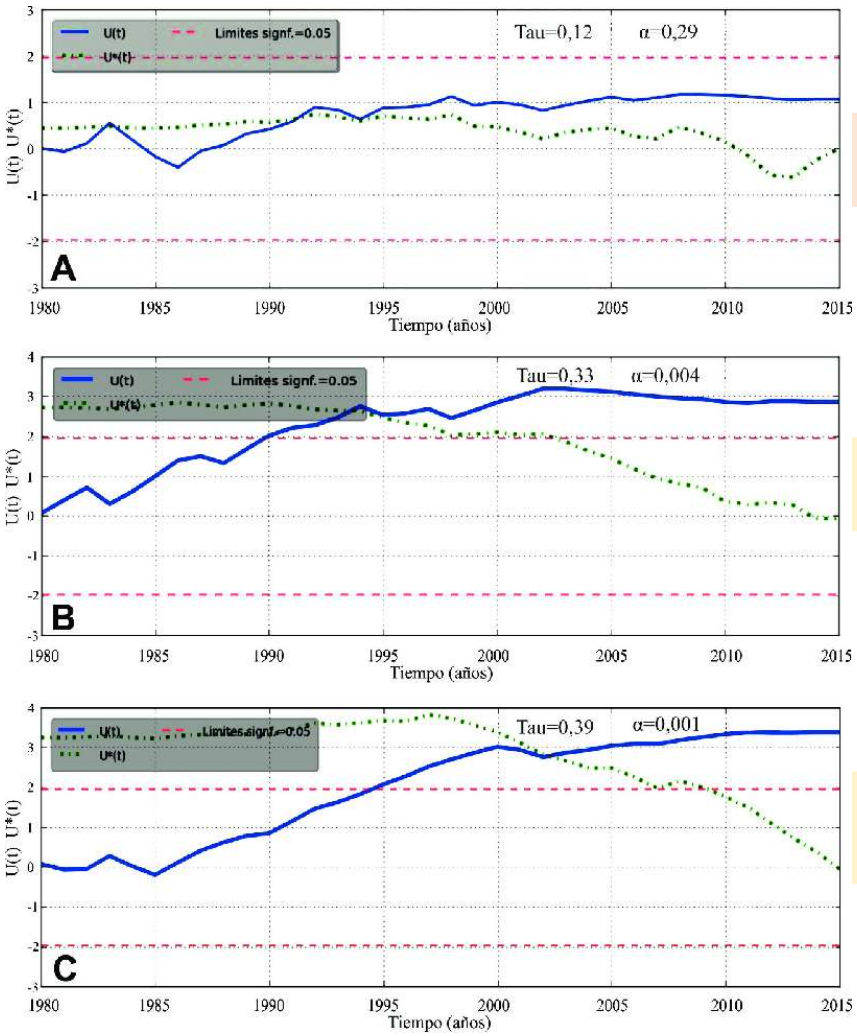
3.3.1 Interpretación de la prueba estadística Mann-Kendall

La interpretación de la prueba estadística no paramétrica Mann-Kendall se muestra en la tabla 1.

Tabla 1. Interpretación de la significancia de la prueba Mann-Kendall

Significancia	Simbología	Z
Sin tendencia	ST	0
Tendencia significativa creciente	TSC	> + 1,96
Tendencia significativa decreciente	TSD	< - 1,96
Tendencia no significativa creciente	TNSC	< + 1,96
Tendencia no significativa decreciente	TNSD	> - 1,96

A continuación, se muestran ejemplos de tendencia de una serie de datos temporales.



Tendencia no significativa creciente

Tendencia significativa creciente

Tendencia significativa creciente

Figura 9. Ejemplos de tendencia de una serie temporal

3.3.2 Configuración de la matriz binaria

De acuerdo a los resultados obtenidos de la prueba estadística no paramétrica Mann-Kendall, los valores de tau y valor p, se interpretan el tipo de tendencia y si es significativa estadísticamente, siendo estos resultados los que se definen como variables para luego consolidarlas en una matriz binaria, donde 1 indica presencia de esa variable y 0 en caso contrario, y esto es para cada sitio geográfico, como se muestra en el siguiente esquema:

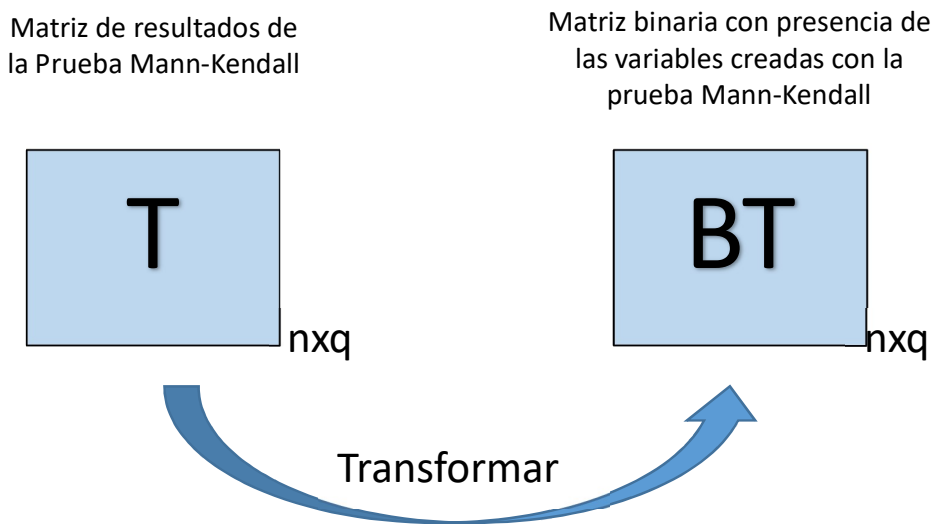


Figura 10. Estructura de la matriz binaria temporal

3.3.3 Herramienta informática en lenguaje R

Para aplicar la prueba estadística no paramétrica Mann-Kendall se lo realizó con la librería Kendall del lenguaje R.

3.3.3.1 Algoritmo de la prueba Mann-Kendall

1. Lectura de la matriz de datos temporal
2. Calculó del valor tau.
3. Calculó del valor estadístico Z.
4. Cálculo del valor p.

3.3.3.2 Código fuente de la prueba Mann-Kendall

```
##Prueba Mann Kendall para tasas de mortalidad del cáncer de mama
library(tidyverse)
library(readxl)
library(Kendall)

c50 <- read_csv("matriz_07_21.csv",
                col_types="innnnnnnnnnnnnn",
                locale = locale(encoding = "latin1"))

prv_mk <- data.frame(id_prv=unique(c50$ID_PROVIN))
prv_mk$tau<-0.0
prv_mk$p<-0.0
for (prv in prv_mk$id_prv)
{
  prv_mk$tau[prv]<-MannKendall(c50[prv,2:16])$tau[1]
  prv_mk$p[prv]<-MannKendall(c50[prv,2:16])$sl[1]
}
write_csv(prv_mk,"prv_mk.csv")
```

3.4 ETAPA 3: INTEGRACIÓN DE LA COMPONENTE ESPACIAL Y TEMPORAL

Las metodologías antes señaladas de análisis de componentes principales geográficamente ponderado GWPCA y la prueba estadística no paramétrica de Mann-Kendall, han sido muy utilizados en la literatura en áreas de la salud, ambiente, agricultura, entre otros, en un contexto espacial y temporal, siendo aplicadas de forma separada, por ejemplo: (Tejedor Flores, 2018) analiza eficazmente el nexo entre el agua, la energía y los alimentos; (Lalangui et al., 2022) analiza la mortalidad infantil en el Ecuador; (Zymaroieva et al., 2019) analiza el rendimiento de la soya basado en las técnicas GWPCA.

Los resultados de las técnicas GWPCA y Mann-Kendall se interpretan de manera individual resultando algo complejo, a su vez, las técnicas actuales con enfoque geográficamente ponderado que incluyen el componente temporal no tienen una validez estadística, por esta razón se seleccionó la prueba estadística Mann-Kendall que analiza la tendencia temporal con una significancia estadística; el componente espacial fue examinado mediante la técnica GWPCA que analiza localmente la heterogeneidad espacial.

Con la finalidad de integrar gráficamente tanto los resultados del GWPCA y la prueba de Mann-Kendall y caracterizar los sitios geográficos y su relación con las dimensiones espacial y temporal, se aplica un algoritmo, como el propuesto por (Vicente-Villardón et al., 2006b) y que luego es ampliado por (Demey et al., 2008a) donde combina un análisis de coordenadas principales (PCoA) y una regresión logística (LR) para construir un Biplot logístico externo.

En el paso de PCoA se utilizó el coeficiente de Russel y Rao para datos dicotómicos, esto evita indeterminación en el cálculo, dado que hay pares de

sitios geográficos en los cuales ninguna de las características está presente, generando d (dobles ausencias) en el denominador.

$$S_{RR} = \frac{a}{a + b + c + d}$$

S_{RR} coeficiente de Russell y Rao está acotado entre cero y uno; uno indica máxima similaridad y cero disimilaridad total.

3.4.1 Configuración de la matriz binaria

Previo a la aplicación de la técnica Biplot logístico, se configuro la matriz de datos binarios compuesta por variables obtenidas de las componentes principales del GWPCA y de la prueba Mann-Kendall.

Para el GWPCA se partió de la matriz que consolida la máxima carga entre las tres primeras componentes de las variables del estudio y para cada unidad espacial, dicha matriz es convertida a binaria considerando el siguiente criterio, las cargas o correlaciones mayores a 0.4 son uno y el resto cero, es de mencionar que se excluyeron las columnas donde todos los valores fueron cero

Para la prueba de Mann-Kendall, se identificaron variables según el tipo de tendencia obtenida por cada unidad espacial, como se muestra en la tabla 1. tendencia creciente significativa (TSC), tendencia creciente no significativa

(TNSC), tendencia significativa decreciente (TSD), tendencia no significativa decreciente (TNSD).

Además, para el caso puntual del estudio se consideró una variable adicional que correspondía al año más reciente y donde se identificaba con 1 los sitios geográficos que presentaban la tasa de mortalidad local superior a la tasa nacional y 0 en el resto.

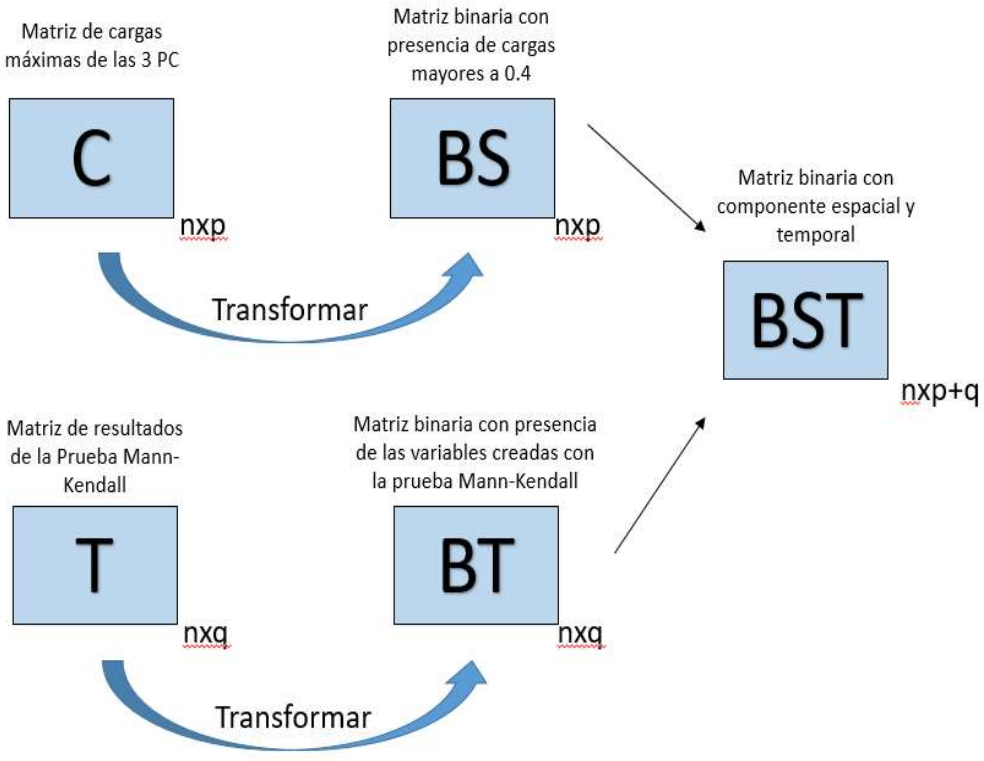


Figura 11. Estructura de la matriz binaria espacio temporal

3.4.2 Formulación del método TSWLB

Sea X la matriz de datos de orden nxm, donde m = p+q que proviene de n unidades espaciales a los que se les cuantifican m atributos o caracteres cualitativos que se asocian a variables binarias, en este caso son atributos espaciales con la técnica GWPCA y atributos temporales con la técnica Mann-Kendall, que toman el valor 0 si la característica (componente espacial o temporal) está ausente y el valor 1 si está presente.

Sea $\pi_{ij} = E(x_{ij})$ la probabilidad de que el j-ésimo componente esté presente en una unidad espacial cualquiera, con coordenadas $y_{is}(i = 1, \ldots\ldots, n; s = 1, \ldots\ldots, k)$ y que está representado en el plano k-dimensional generado por el Análisis de Coordenadas Principales (ACoP), π_{ij} puede escribirse en función de las coordenadas principales como sigue:

$$\pi_{ij} = \frac{e^{b_{jo} + \sum_{s=1}^{k} b_{js} y_{is}}}{1 + e^{b_{jo} + \sum_{s=1}^{k} b_{js} y_{is}}}$$

Donde $b_{js}(j = 1, \ldots\ldots, p)$ son los coeficientes de la regresión logística que corresponden a la j-ésima variable (componente espacial o temporal) en la k-ésima dimensión. La ecuación anterior es equivalente al modelo lineal generalizado que utiliza la función logit, como función de enlace para evitar problemas de escala.

Este procedimiento genera un gráfico bi o tri dimensional donde las y_s representadas como puntos (unidades espaciales) y los b_s estimados para

cada componente espacial y temporal son representados como vectores los cuales determinan direcciones de los ejes Biplot.

La proyección de cada unidad espacial sobre el segmento que representa a cada componente (atributo), permite obtener la probabilidad estimada de presencia de un componente (espacial o temporal) en particular para cada unidad espacial.

Es importante considerar proyectar aquellos componentes que se relacionan directamente con la configuración, es decir, los parámetros que presenten la mejor calidad de representación después de ajustar la regresión logística.

3.4.3 Geometría e Interpretación del método TSWLB

En nuestro contexto, se interpreta como que la proyección de una unidad espacial en la dirección de un vector (espacio temporal) cualquiera predice la probabilidad de la presencia de esa componente en la unidad espacial.

Para facilitar la interpretación gráfica, en los extremos de cada vector se fijan puntos de predicción con probabilidad conocida, es así como el 0,50 se fija como punto corte para la predicción de la presencia y 0,75 para la dirección de mayor probabilidad creciente.

La longitud del vector debe ser interpretada como una medida inversa de la capacidad discriminatoria de los componentes espacio temporal, es decir, vectores más cortos discriminan mejor a las unidades espaciales. La relación

entre los diferentes componentes proyectados sobre el plano Biplot, se interpretan según el ángulo que formen.

Cuando dos componentes espacio temporal tengan el mismo sentido de predicción se dice que están positivamente correlacionados, cuando tengan direcciones opuestas se correlacionan negativamente, y cuando formen un ángulo cerca de 90° se dice que son independientes.

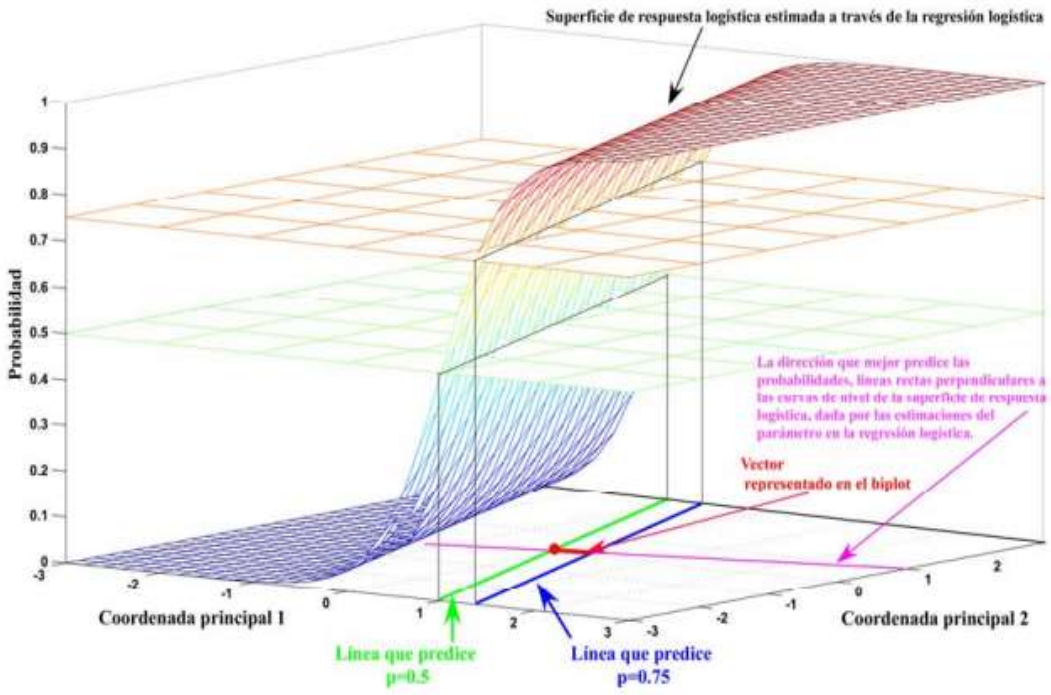


Figura 12. Geometría de la curva de respuesta logística ajustada (adaptada de *Demey et al., 2008a)*

La técnica TSWLB representa a los sitios geográficos como puntos y las variables como vectores en un diagrama de dispersión en un espacio euclidiano.

Los vectores indican dirección y están más correlacionados con la presencia de la característica, esto permite caracterizar y dar un ordenamiento a los sitios geográficos en función de la dimensión espacial y temporal.

Para interpretar las características asociadas al agrupamiento de los sitios geográficos proyectamos los puntos sobre la dirección de los vectores, cuanto más lejos se proyecte el punto en la dirección de la flecha, mayor es la probabilidad de la característica.

El origen del vector es el punto que predice una probabilidad 0.5 y la flecha indica la dirección de probabilidad creciente. También proporciona información adicional sobre la bondad de ajuste de cada variable.

El Biplot logístico ponderado espacio temporal cumple con las siguientes reglas para su interpretación:

- La distancia entre los puntos (sitios geográficos) en el gráfico, están inversamente relacionadas con las similitudes de sus perfiles, es decir, los sitios cercanos tienen características similares.

- El ángulo entre vectores y el eje factorial, indica el grado de relación entre la variable y la dimensión latente.

- El ángulo entre vectores, indican el grado de asociación entre ellos, los ángulos agudos (pequeños) indican que los vectores están estrechamente relacionados.

- Las proyecciones de los puntos (sitios geográficos) sobre el vector (variable), estiman la probabilidad esperada de la característica para ese sitio geográfico.

- La longitud del vector, indica el poder discriminante de la variable en el ordenamiento de los sitios geográficos, mientras más pequeños mayor poder discriminante.

3.4.4 Metodología de priorización

Se utilizó el método de mínima varianza de Ward por ser el mejor que clasifica y permite crear clústeres de priorización, el cual es un procedimiento donde el criterio para la elección del par de clústeres a mezclar en cada paso se basa en el valor óptimo de una función objetivo, la suma de cuadrados de la varianza y para producir una solución integral, se aplica el diagrama de Voronoi, el cual divide el espacio en función de las relaciones geométricas espacio temporales determinado por las distancias a un conjunto discreto de puntos.

3.4.5 Herramienta informática

Para realizar los cálculos y obtener las representaciones gráficas, se utilizó un programa de computadora basado en el código Matlab llamado MultBiplot. (Vicente-Villardón, 2013) (Vicente-Villardón, 2021)

A continuación, se resume los pasos para realizar el método propuesto Biplot logístico ponderado espacio temporal (TSWLB)

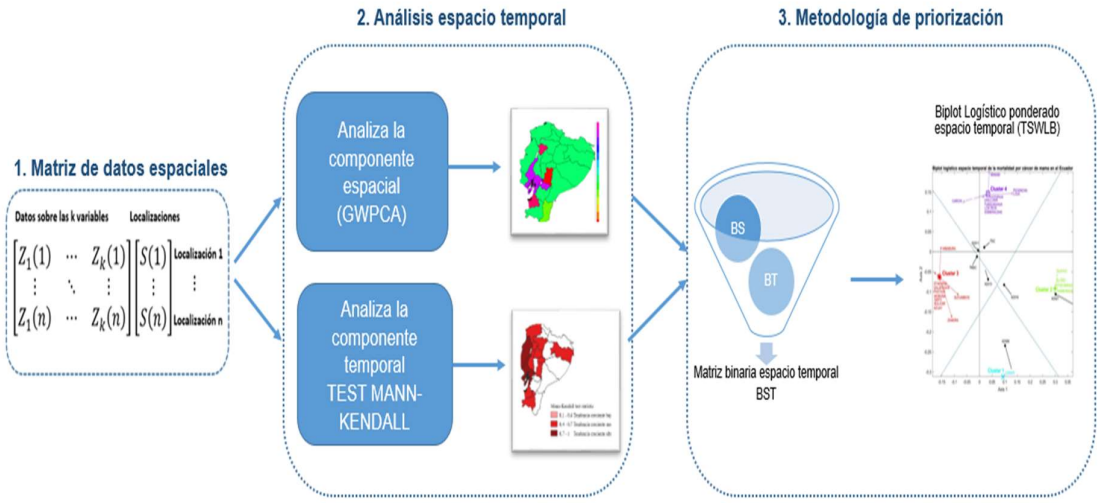


Figura 13. Esquema del método propuesto "Biplot logístico ponderado espacio temporal" (TSWLB)

CAPÍTULO IV

4. ESTUDIO DE LA MORTALIDAD POR CÁNCER DE MAMA EN EL ECUADOR MEDIANTE LA APLICACIÓN DEL BIPLOT LOGISTICO PONDERADO ESPACIO TEMPORAL

El cáncer de mama es la principal causa de mortalidad en las mujeres a nivel mundial, la mayoría de las muertes se registran en países de ingresos bajos y medianos. (Beltrán & Martínez, 2021) Según datos del Globocan 2020 en Ecuador la tasa de mortalidad por cáncer de mama (BCMR) es de 10.9 por 100.000 habitantes, representando la cuarta causa de mortalidad en el país. (Globocan - IARC, 2020, s. f.) (Tanca-Campozano et al., 2019) (Jaramillo-Feijoo, Galindo-Villardon, Real-Cotto, et al., 2020)

La Organización mundial de la salud (OMS), entre uno de los objetivos de la Iniciativa Mundial contra el Cáncer de Mama, es reducir un 2.5% anual la mortalidad mundial por esta enfermedad, en este sentido, se realiza el siguiente estudio de la mortalidad por cáncer de mama en el Ecuador utilizando el método propuesto en esta investigación TSWLB, con la finalidad de proporcionar información base para la planificación en la prevención y control de esta enfermedad. (OMS, Cáncer de mama)

4.1 MATERIALES

4.1.1 Localización

Esta investigación se realizó en el Ecuador, que está localizado en América del Sur limitando al norte con Colombia, sur y este Perú y al oeste con el Océano Pacífico.

4.1.2 Período de investigación

El periodo del estudio fue entre el 2007 al 2021

4.1.3 Recursos empleados

4.1.3.1 Humanos

- Investigador
- Director
- Tutor

4.1.3.2 Materiales

Paquetes de análisis espacio-temporal en lenguaje RStudio, software MultiBiplot Laptop,

4.1.4 Universo y muestra

4.1.4.1 Universo

La base de datos de acceso abierto son las defunciones generales del. Instituto Nacional de Estadísticas y Censos - INEC. El período de estudio es de 15 años del 2007 al 2021.

4.1.4.2 Muestra

La base de datos de defunciones para el estudio incluye las mujeres fallecidas por cáncer de mama registradas por cada provincia del Ecuador.

4.2 ASPECTOS ÉTICOS Y LEGALES

Es de mencionar que, en esta investigación se garantiza la confidencialidad de los datos, a su vez, se expresa que los datos fueron anonimizados de acuerdo con las leyes de protección de datos, confidencialidad y seguridad vigentes en el Ecuador. Finalmente, indicar que la base de datos con la que se trabajó esta investigación es de acceso abierto

4.3 MÉTODO

4.3.1 Tipo de investigación

Análisis espacio-temporal multivariante

4.3.2 Criterios de inclusión/exclusión

4.3.2.1 Criterios de inclusión

Las muertes por cáncer de mama en mujeres del periodo 2007 al 2021.

4.3.2.2 Criterios de exclusión

Las muertes tardías y las muertes de cáncer de mama en hombres.

4.3.3 Variables

- Causa de muerte: Cáncer de mama
- Sexo: mujeres
- Provincia de residencia de la persona fallecida
- Año de mortalidad
- Número de defunciones
- Población de mujeres del Ecuador
- Tasa de mortalidad por cáncer de mama
- Coordenadas X, Y según sistema geodésico UTM del Ecuador

4.4 ETAPAS PARA APLICAR EL MÉTODO TSWLB

El Biplot logístico ponderado espacio temporal (TSWLB) es una técnica multivariante que captura y representa simultáneamente los componentes espacial y temporal.

Es una metodología de priorización porque ordena a los sitios geográficos en función de los años con cargas mayores a 0.4 y las tendencias temporales crecientes o decrecientes; el análisis de la componente espacial se realizó con el modelo GWPCA que mide la heterogeneidad espacial; y para el componente temporal se utilizó la prueba estadística no paramétrica Mann-Kendall que determina si es una tendencia creciente o decreciente y si ésta es estadísticamente significativa.

Se excluyeron los años que no cumplieron el criterio con cargas mayores a 0.4 en ningún sitio geográfico, además, se consideró los sitios geográficos con tasas de mortalidad por cáncer de mama BCMR mayores a la tasa nacional durante el año más reciente.

Para aplicar el método TSWLB se inicia con la preparación de la matriz de datos espaciales.

4.4.1 Matriz de datos espaciales

Para preparar la matriz de datos espaciales se seleccionó la escala, en este caso será el nivel provincial, y se contabilizaron las muertes de mujeres fallecidas por cáncer de mama por provincia de residencia. Se descartaron el registro de defunciones tardías y de no residentes en Ecuador.

Luego, se calculó la tasa de mortalidad por cáncer de mama (BCMR), mediante la siguiente fórmula:

$$BCMR = 100000\ x\ \frac{Defunciones\ por\ cáncer\ de\ mama\ en\ mujeres}{Población\ de\ mujeres}$$

La matriz de datos geoespaciales está compuesta por las BCMR medidas en las 24 provincias del Ecuador. Las coordenadas geográficas fueron obtenidas con el sistema UTM WGS84 que para Ecuador corresponde el 17S.

4.4.2 Etapa 1: Análisis de la componente espacial

La técnica GWPCA que examina la parte no estacionaria de los datos localmente en el espacio geográfico, fue utilizado en este estudio para analizar las tasas de mortalidad por cáncer de mama (BCMR) del periodo 2007 al 2021 en las 24 provincias del Ecuador.

4.4.3 Etapa 2: Análisis de la componente temporal

Se utilizó la prueba estadística no paramétrica Mann-Kendall para examinar la tendencia de la serie temporal de las tasas de mortalidad por cáncer de mama del periodo 2007 al 2021 en las 24 provincias del Ecuador. Para cada provincia fue identificada el tipo de tendencia si es creciente o decreciente de las tasas de mortalidad por cáncer de mama. Para la aplicación de la prueba los datos no requieren que sigan alguna distribución en particular.

4.4.4 Etapa 3: Integración de la componente espacial y temporal

Una vez que se ha configurado la matriz de datos binarios en función de los resultados obtenidos de cada componente espacial y temporal, se ejecuta el Biplot Logístico externo (Demey et al., 2008a) utilizando el programa MultiBiplot (Vicente-Villardón, 2021), obteniéndose una representación gráfica de las unidades espaciales como puntos y las componentes espacial y temporal como vectores, generando un análisis integral para evidenciar las interacciones espacio temporales a nivel provincial, ofreciendo una comprensión holística de la estructura de datos. Se seleccionó el coeficiente de Russel y Rao para datos dicotómicos, a fin de cuantificar la similaridad de las características presentes y evitar las dobles ausencias.

4.5 RESULTADOS Y DISCUSIÓN

Desde el año 2007, (Fig.13) se observa el incremento sostenido de las tasas de mortalidad por cáncer de mama en el Ecuador, pasando de 4,9 a 8,5 por 100.000 habitantes en el año 2021, sin embargo, este comportamiento es diferente por provincia, como se observa en la Fig. 14.

La heterogeneidad espacial local fue analizada con la técnica GWPCA. El porcentaje de variabilidad total de los primeros 3 componentes que analizan las tasas de mortalidad por cáncer de mama en las provincias del Ecuador, se muestra en la Fig.15, donde se observa que las provincias de mayor variabilidad se encuentran al norte y sur del país, siendo al norte: Esmeraldas, Carchi e Imbabura; y al sur: Loja, El Oro y Zamora Chinchipe; dicha variabilidad podría deberse a que son zonas fronterizas y dinámicas en su población.

En la Fig. 16 se muestra el año con mayor carga de los primeros 3 componentes, donde se observa una variación geográfica en la influencia de la mortalidad por cáncer de mama del periodo 2007 al 2021 en las provincias del Ecuador, se tiene que al norte del país los años que han influenciado en la mortalidad de cáncer de mama fueron los años 2007 y 2008; en la parte sur fue el año 2018, pero en el centro fue el año 2017 y en el centro - este fue el año 2009 a diferencia del centro – oeste que fue el año 2019.

Es importante evidenciar los cambios que se han dado en este periodo de estudio donde la presentación de muertes por cáncer de mama pudiera deberse a distintos factores, como, el acceso a la atención oncológica, movilidad, desarrollo de nuevos servicios y capacidad resolutiva especializada, e incluso el registro de las estadísticas vitales, entre otros.

En la Fig. 17 muestra la tendencia temporal analizada con la prueba estadística no paramétrica Mann-Kendall en las 24 provincias del Ecuador, las tendencias muestran que las tasas no son constantes espacialmente. A nivel regional se observa en las provincias de la región costa y sierra un mayor predominio de tendencias crecientes en las tasas de mortalidad por cáncer de mama.

El Biplot logístico integra los resultados de las técnicas GWPCA y prueba Mann Kendall, esto se muestra en la Fig. 18 donde se observa cuatro clústeres, que están agrupados en función de la evolución en el tiempo de la mortalidad por cáncer de mama y el año que ha influido mayormente en su mortalidad.

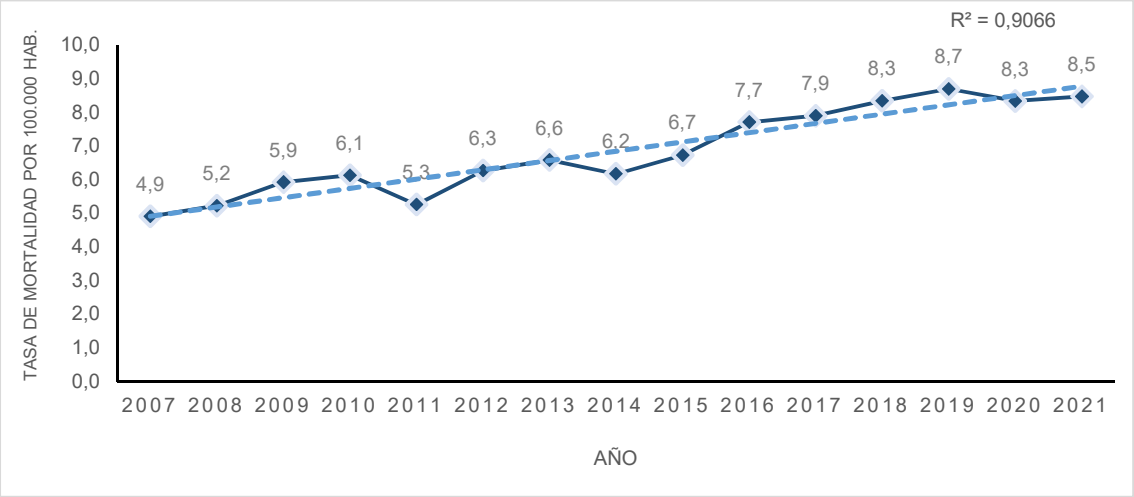


Figura 14. *Comportamiento de las tasas de mortalidad por cáncer de mama en Ecuador. 2007-2021*

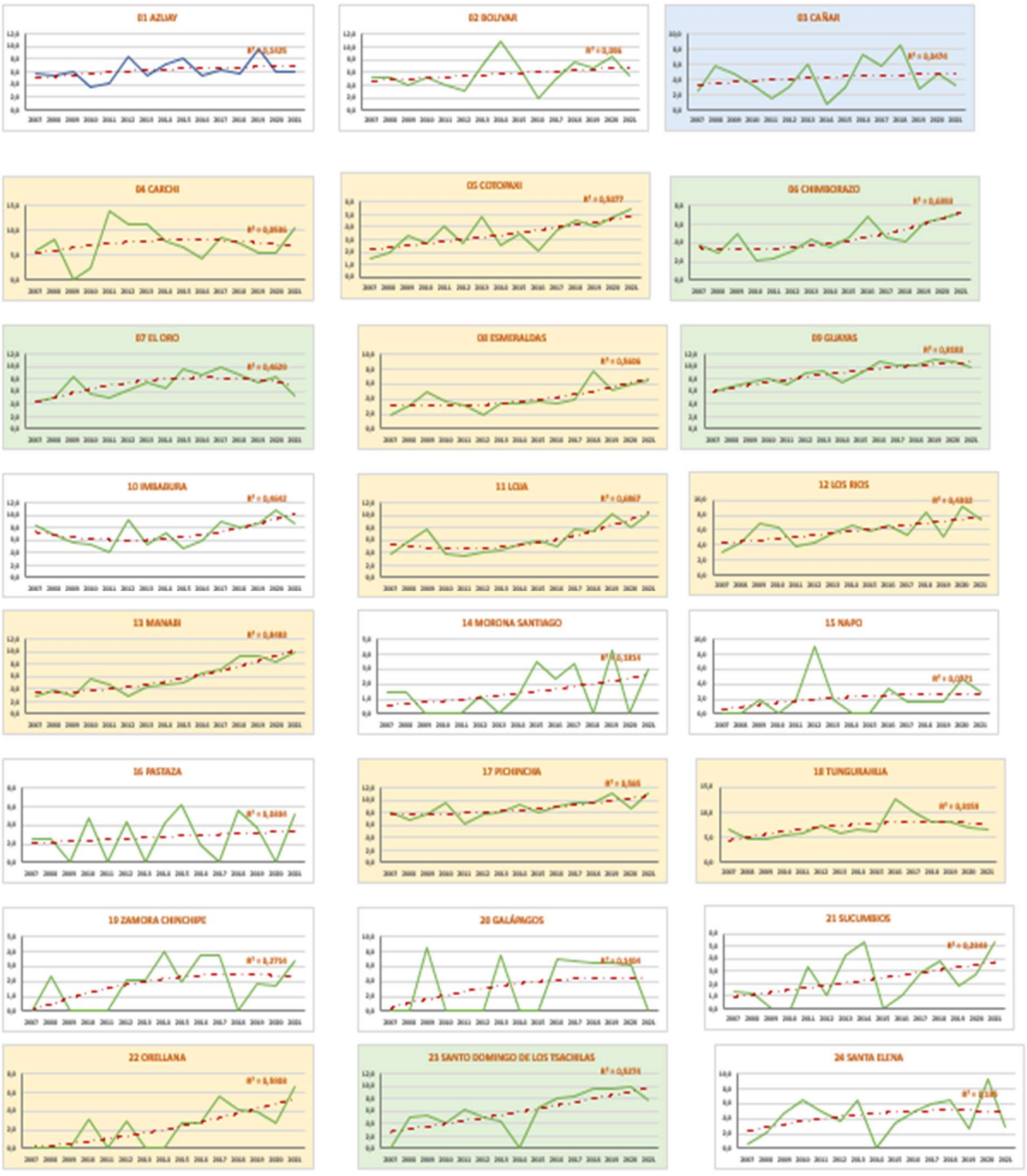


Figura 15. Comportamiento de las tasas de mortalidad por cáncer de mama por provincias del Ecuador. 2007 - 2021

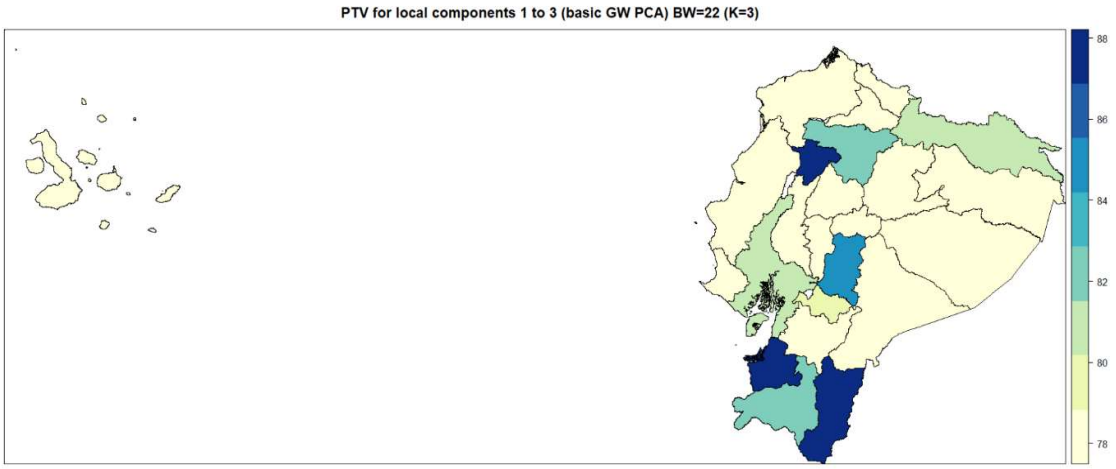


Figura 16. Mapa del porcentaje de variabilidad local de las 3 primeras componentes

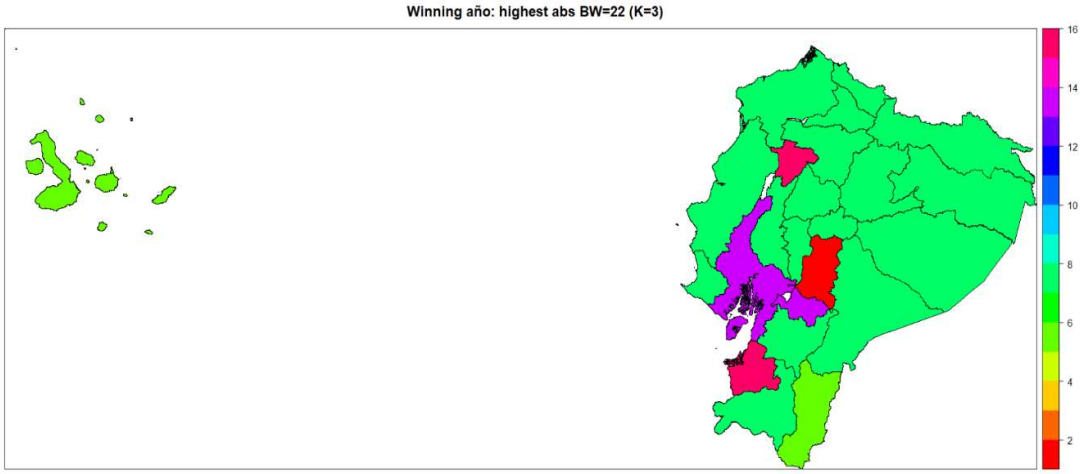


Figura 17. Mapa representando la variable o año ganador con la mayor carga

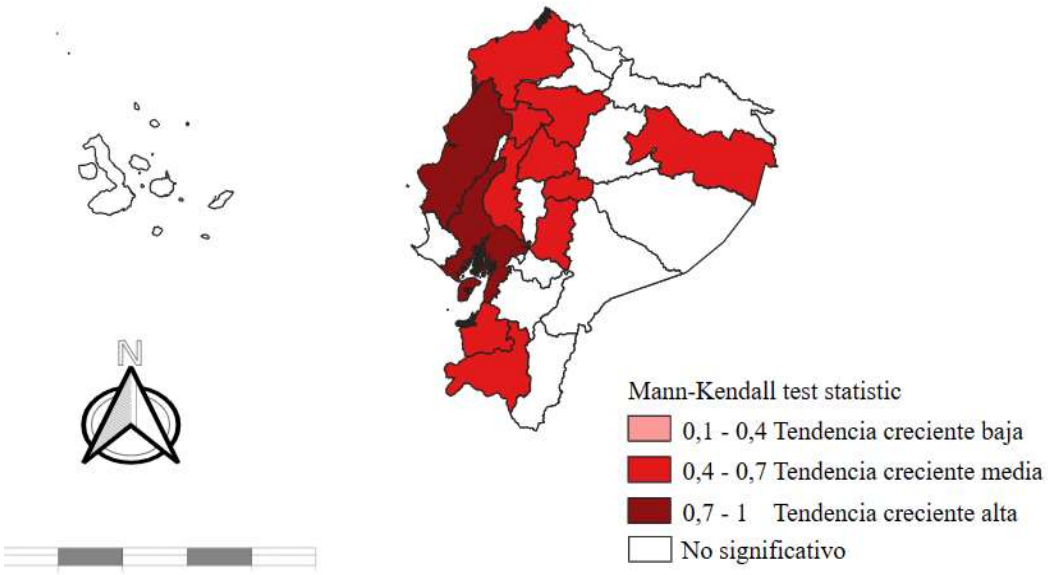


Figura 18. Mapa de la tendencia temporal Mann-Kendall (Tau). 2007-2021

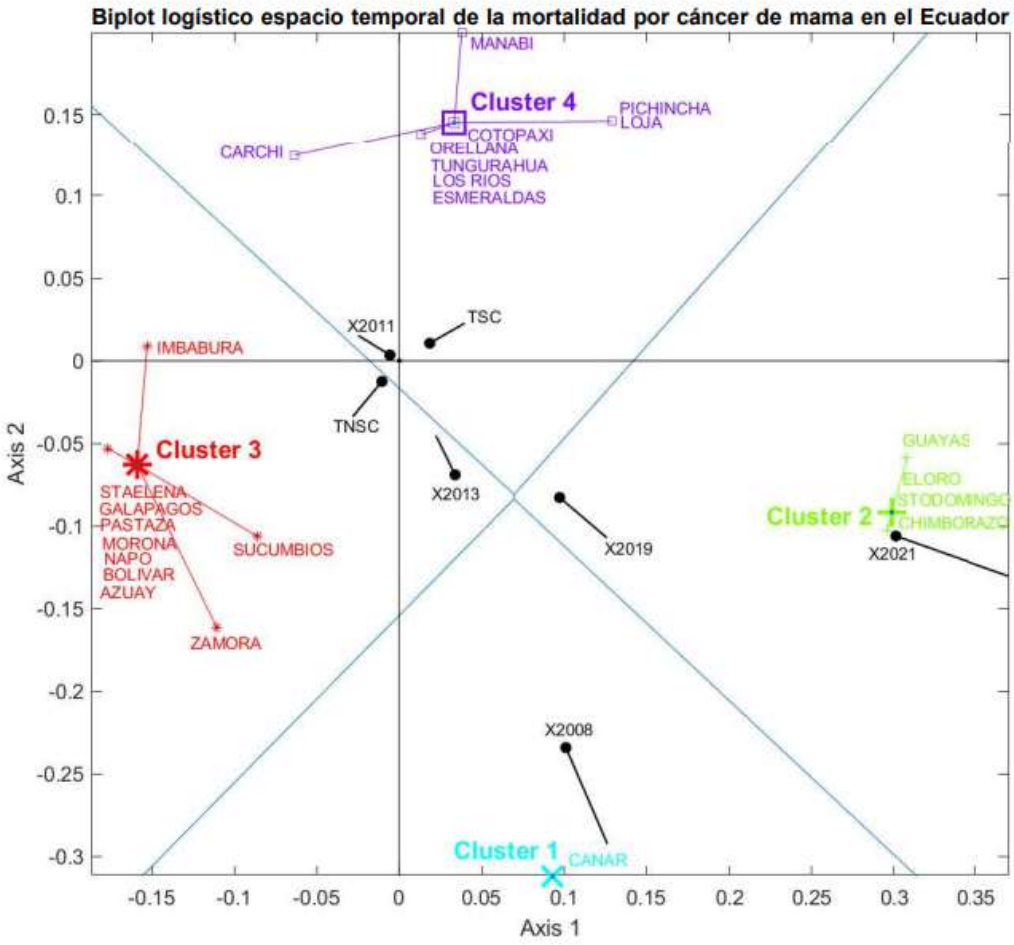


Figura 19. Biplot logístico ponderado espacio temporal y clústeres de priorización

La tasa de mortalidad por cáncer de mama en el Ecuador tiene un incremento sostenido en los últimos 15 años, esto se asemeja a estudios regionales que indican que el cáncer de mama es un problema de salud en América Latina y el Caribe donde anualmente mueren unas 300.000 mujeres por esta enfermedad. (Tanca-Campozano et al., 2019) (Robles & Galanis, 2002)

Existe una variabilidad espacial a nivel provincial de las tasas de mortalidad por cáncer de mama, siendo las provincias ubicadas al sur y al norte del país con mayor variabilidad en la mortalidad de esta enfermedad, así se tiene al norte: Esmeraldas, Carchi e Imbabura; y al sur: Loja, El Oro y Zamora Chinchipe; dicha variabilidad podría ser que sea debida a que son zonas fronterizas y dinámicas en su población, estos resultados son similares a estudios donde se evidencia diferencias en la mortalidad a nivel urbano y rural. (Jaramillo-Feijoo, Galindo-Villardon, Real-Cotto, et al., 2020)

Durante el periodo de estudio se pudo comprobar una variación geográfica en los años que presentaron mayores cargas dentro de los primeros 3 componentes principales a nivel local, es decir, años que influyeron en la mortalidad por cáncer de mama a nivel provincial, así se tiene, las provincias al norte del país fueron influenciados por los años 2007 y 2008; en las provincias al sur fue el año 2018, mientras que en el centro del país fue el año 2017 y en el centro - este fue el año 2009 a diferencia del centro – oeste que fue el año 2019. Dichos resultados son parecidos a estudios previos, donde se observan diferencias de la mortalidad de cáncer a nivel geográfico. (Ramos-Herrera et al., 2020)

Se identificaron cuatro clústeres mediante el Biplot logístico el cual integró los resultados de las técnicas GWPCA y Mann-Kendall, el clúster 2 y 4 tienen un comportamiento creciente estadísticamente significativo de las muertes por cáncer de mama durante los 15 años de estudio, en el clúster 2 los años más recientes está influenciado su mortalidad, mientras que los clústeres 1 y 3 tienen un comportamiento creciente estadísticamente no significativo, los clústeres 3 y 4 están más influenciados su mortalidad por los años intermedios 2011 y 2013, el clúster 1 está influenciado su mortalidad por el año 2008.

Considerando que, el clúster 2 tiene una tendencia con comportamiento creciente significativa estadísticamente e influenciada su mortalidad por los años recientes, se puede identificar como grupo prioritario, dicho grupo lo conforman 4 provincias que representan el 17% del total país, siendo dos provincias de la región costa: Guayas y El Oro; y dos provincias de la región sierra: Santo Domingo y Chimborazo; esta representación mediante el Biplot logístico permitió capturar e integrar información espacial y temporal, siendo una técnica innovadora e integradora para ser aplicada en distintas áreas. (Demey et al., 2008b) (Galindo et al., 2011) (Gallego-Álvarez & Vicente-Villardón, 2012) (de Noronha Vaz et al., 2015)

CAPÍTULO V

5.- CONCLUSIONES

Los resultados obtenidos de esta investigación nos permiten concluir:

1. La amplia revisión bibliográfica ha puesto de manifiesto el interés por las técnicas multivariantes que capturan la dimensión espacio temporal de los datos; sin embargo, no existe ni una sola técnica estadística que integre ambas componentes en un solo análisis; es decir, hasta el momento, no existía una comprensión holística de la estructura de los datos.

2. La revisión del software pone de manifiesto que existen librerías en lenguaje R para analizar cada componente por separado, corroborando la necesidad de integrar la variabilidad espacial y temporal en un solo análisis para estudiar, no solo los efectos aislados, sino también las interacciones entre ellos. El paquete GWModel en lo espacial y el test Mann-Kendall en lo temporal, son los más utilizados.

3. En esta tesis doctoral se propone un nuevo método estadístico que hemos denominado **Biplot Logístico ponderado espacio temporal** – **TSWLB**, el cual analiza la componente espacial, a través de un análisis de componentes principales geográficamente ponderados, y contrasta la tendencia temporal con el test de Mann-Kendall, e integra ambas componentes en un solo análisis. Los resultados se presentan en un plano factorial mediante un análisis Biplot Logístico Externo.

4. El método TSWLB propuesto, permite integrar las componentes espacio temporales en una única representación simultánea donde se evidencia, con alta calidad de representación, la información de las relaciones entre las unidades espaciales y las características presentes en ellas, con metodología de priorización, de fácil interpretación.

5. El cáncer de mama representa la cuarta causa de muerte en las mujeres ecuatorianas, con un incremento de las muertes en los tres últimos quinquenios; mediante el método propuesto TSWLB ha sido posible capturar y representar las variaciones geográficas locales y temporales, priorizando las provincias que tienen una tendencia significativamente creciente en la mortalidad por cáncer de mama.

6. Los clústeres 2 y 4 se caracterizan por tener alta probabilidad de tendencia creciente. Las provincias que conforman el clúster 4 son: Manabí, Pichincha, Loja, Cotopaxi, Orellana, Tungurahua, Los Ríos, Esmeraldas y Carchi; y en el clúster 2 están: Guayas, El Oro, Santo Domingo y Chimborazo.

7. Los clústeres 1 y 3 se caracterizan por tener alta probabilidad de tendencia decreciente. Las provincias que conforman el clúster 3 son: Santa Elena, Galápagos, Pastaza, Morona, Napo, Bolívar, Azuay, Sucumbíos e Imbabura; y en el clúster 1 está: Cañar.

8. Se ha demostrado que, la técnica multivariante TSWLB capturó y representó simultáneamente, de manera gráfica, la mortalidad por cáncer de mama en los últimos 15 años, comprobándose diferencias espacio-temporales a nivel provincial e identificándose cuatro clústeres con características propias que permitieron priorizar provincias. Guayas, El Oro, Chimborazo y Santo Domingo de los Tsáchilas, son las provincias que requieren de intervenciones urgentes en estrategias de detección precoz y gestión integral del cáncer de mama.

APORTACIONES CIENTÍFICAS

ARTÍCULO 1:

ARTÍCULO 1: “ANÁLISIS CLÚSTER PARA BIG DATA: UNA APLICACIÓN CON VARIABLES DEMOGRÁFICAS EN PROVINCIAS DEL ECUADOR”

J. health med. sci.,
6(1):45-50, 2020.

Análisis Clúster para Big Data: una aplicación con variables demográficas en provincias del Ecuador

Cluster analysis for big data: an application with demographic variables in provinces of Ecuador

Jaramillo-Feijoo Leyda Elizabeth[1], Galindo-Villardon María Purificación[2], Real-Cotto Jhony Joe[3]

JARAMILLO, L.; GAINDO, M. & REAL, J. Análisis clúster para big data: una aplicación con variables demográficas en provincias del ecuador. *J. health med. sci., 6(1)*:45-50, 2020.

RESUMEN: Los métodos de clasificación permiten explorar y analizar grandes conjuntos de datos visualmente, lo cual es de gran utilidad para tomar decisiones rápidas. El objetivo fue comparar dos métodos de análisis de clúster para big data en variables demográficas de las provincias del Ecuador. Se hizo uso de un estudio observacional de tipo comparativo mediante la representación simultanea del HJ-Biplot y el método Two Step (clúster bietápico), a través del software MultBiplot y SPSS. Los datos corresponden a variables demográficas de interés sociosanitarias tasa de mortalidad general, tasa de mortalidad infantil, tasa de natalidad, densidad poblacional, porcentaje urbano y esperanza de vida, medidas en las provincias del Ecuador. Se utilizaron datos provenientes del Instituto de Estadísticas y Censos INEC. Se analizó la asociación entre variables y se identificaron clústeres de las provincias del Ecuador según estas variables demográficas. Según la representación simultánea del HJ-Biplot se identificaron 3 clústeres, el clúster 1 son provincias con mayor densidad poblacional y tasas de mortalidad general, pero valores bajos de tasas de natalidad, el clúster 2 agrupa provincias con mayor esperanza de vida y tasas de mortalidad infantil pero bajos valores de tasa de natalidad y el clúster 3 están las provincias con valores altos de tasas de natalidad y valores bajos de densidad poblacional, esperanza de vida, tasas de mortalidad general y mortalidad infantil, distintos resultados se obtuvieron con el método Two Step. Se pudo concluir que estos métodos son de utilidad para explorar las similitudes entre las provincias según variables demográficas.

PALABRAS CLAVE: clúster, demográficas, HJ-Biplot, método two step.

INTRODUCCIÓN

Según cifras del censo del año 2010 por el Instituto Nacional de Estadísticas y Censos (INEC), Ecuador registra una población de 14 millones de habitantes, de los cuales el 66% es población urbana. El crecimiento de la población se ha visto afectada por la reducción de la tasa bruta de natalidad de 32,4 a 11,4 nacimientos por 1000 habitantes entre 1981 y 2010 (Lucio et al., 2011), la disminución de la tasa de mortalidad de 6,7 muertes por 1000 habitantes en 1981 a 4,3 en 2008 y la tasa de mortalidad infantil en 2009 fue de 20 por 1000 nacidos vivos.

Las técnicas de minería de datos son herramientas que tienen como propósito descubrir conocimiento, siendo el agrupamiento o clúster uno de estos métodos, (Shirkhorshidi et al., 2014).

El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado, (Vicente, 2007) es una técnica no supervisada que se utiliza para clasificar grandes conjuntos de datos en grupos correlativos, tiene aplicación en muchos campos, se incluyen otras técnicas para el análisis de big data, tales como: el aprendizaje automático, el reconocimiento de patrones, entre otros. La agrupación consiste en clasificar objetos similares en grupos distintos, es decir la partición de un conjunto de datos en subconjuntos, los mismos que tienen características similares.

La generación de indicadores socio demográficos, epidemiológicos y de producción,

[1] Ingeniera en Estadística e Informática. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.
[2] Vicerrectora de ordenación académica y profesorado en la Universidad de Salamanca, España.
[3] Docente de la Universidad de Guayaquil. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.

permiten conocer la situación de salud de la población, facilitando la comparación y análisis de los avances en la salud individual y colectiva (OPS, 2020). Resulta interesante e importante identificar clústeres de estos indicadores epidemiológicos medidos en las áreas geográficas, determinar hallazgos y analizar patrones, siendo una herramienta que contribuye a determinar los puntos críticos y falencias a ser superadas; así como, un aporte significativo e importante en la epidemiología y salud pública. (Gonzalez, 2015). Desde el punto de vista de la información sanitaria, es beneficioso determinar si hay similitudes entre los países de cada región y de distintas regiones (Verhasselt & Mansourian, 1991).

En términos del tipo de datos que se utiliza, se pueden considerar para la aplicación de clúster, el jerárquico que está limitado a conjuntos de datos pequeños, K-Means está restringido a valores continuos y Two Step que permite crear modelos de clúster basado tanto en variables continuas como categórica y el número de clúster se determina automáticamente, se muestra la aplicación del método Two Step, y del HJ-Biplot, dando a conocer y destacando sus ventajas (Şchiopu, 2010) (Bacher et al., 2004).

Este tipo de estudio no se ha realizado en el país y más aún con la aplicación de diversos métodos para evaluar variables sociodemográficas de impacto sanitario, por lo que se tuvo como objetivo comparar dos métodos de análisis de clúster para big data en variables demográficas de las provincias del Ecuador.

MATERIAL Y MÉTODO

Estudio de tipo observacional y comparativo (Santos, 2015) donde se realizó un análisis de clúster con variables demográficas de interés sociosanitarias, medidas en provincias del Ecuador de la población femenina. Los datos fueron tomados de los registros administrativos del año 2017 del INEC. Las variables evaluadas fueron: tasa de natalidad, tasa de mortalidad, tasa de mortalidad infantil, esperanza de vida, densidad poblacional, población urbana. Se utilizaron métodos multivariantes y de clasificación, tales como: HJ-Biplot y Two step.

HJ-Biplot

El HJ-Biplot es una representación gráfica multivariante de una matriz Xnxp mediante los marcadores j_1.... jn para sus filas y h_1...h_p para sus columnas,(Gabriel, 1971) (Galindo, 1986) elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación. Se utilizó el software MultBiplot (Vicente, 2010) para realizar el gráfico donde se obtiene la representación simultanea de las provincias, los indicadores epidemiológicos y la identificación de clúster.

$X = UDV^T$ $\quad J = U\,D$

$H = V\,D$

Two Step (cluster bietápico)

El origen del método "TwoStep" fue el algoritmo BIRCH, (Zhang et al., 1996) que fue desarrollado por Chiu et al. (2001) para realizar análisis de clúster, maneja variables de tipo continuas y categóricas. Consiste en dos fases: primero se realiza un proceso de pre-clusterización para todo el conjunto de registros agrupando éstos en muchos pequeños subclústeres y posteriormente se agrupan estos subclústeres mediante un algoritmo de agrupamiento jerárquico hasta obtener el número deseado de clúster.

Siguiendo esta metodología, como el número de elementos a procesar es mucho menor que el número total original de registros y dado que requiere un análisis para todos ellos, este algoritmo es muy eficiente desde el punto de vista de coste operacional. El método Two-Step permite dos tipos de medidas en función del tipo de variables que se tiene en la matriz, las mismas que son:

- Distancia Euclídea
- Distancia Máxima Verosimilitud

Además, maneja dos criterios de agrupamiento, el AKAIKE y el SCHWARTZ:

AKAIKE: $AIC = -2 \bullet lnL(\theta) + 2K$

SCHWARTZ: BIC $= -2 \bullet lnL(\theta) + (\ln(n) \bullet K)$

El modelo con el valor BIC más bajo se considera el mejor en explicar los datos del análisis con el mínimo número de parámetros. Se utilizó el software SPSS versión 20 (Bacher et al., 2004) para realizar este análisis.

RESULTADOS

La Tabla I, muestra la calidad de representación de las provincias del Ecuador sobre el HJ-Biplot. Una vez preparada la matriz de datos, con las provincias como filas y las variables demográficas como columnas; siendo valores relativos, para el análisis HJ-Biplot se estandarizó por columnas y se realizó una descomposición de valores singulares, considerando una dimensión de dos componentes.

La Figura 1, resume la representación simultánea de la matriz de datos a través del HJ-Biplot y la técnica del clúster jerárquico con coordenadas biplot y distancia euclídea. El análisis HJ-Biplot permite hacer un ordenamiento de las provincias al proyectar los marcadores filas sobre las variables demográficas, a continuación, se detallan los clústeres formados:

- Clúster1: Guayas y Pichincha.
- Clúster2: Bolívar, Cotopaxi, Carchi, Azuay, Tungurahua, Chimborazo, Loja, Santa Elena, El Oro, Los Ríos, Manabí, Cañar, Imbabura.
- Clúster 3: Esmeraldas, Santo Domingo de los Tsáchilas, Orellana, Pastaza, Zamora Chinchipe, Sucumbíos, Morona Santiago, Napo.

Adicionalmente, se aplicó otra técnica de clasificación, el análisis de clúster bietápico con el software SPSS, dicha técnica combina variables continuas y categóricas, se incorporó la variable categórica tipo de región, que agrupa las provincias en 4 regiones que son: costa, sierra, oriente e insular, esta última fue excluida por corresponder a una provincia con poca población. Estos resultados son mostrados en la Figura 2.

DISCUSIÓN

La Tabla I mostró un 67,7% como varianza explicada del análisis, lo cual pone de manifiesto que todas las variables se encuentran bien representadas en el eje 1 y 2, aunque la tasa de natalidad está mejor representada en el eje 1; por otro lado, las provincias de Pastaza, Zamora Chinchipe, Sucumbíos, Orellana están bien representadas en el eje 1 y las demás provincias en el eje 2.

Tabla I. Calidad de representación de las provincias del Ecuador sobre el HJ-Biplot.

Calidad de Representación de las Provincias			
	Provincias	Axis 1	Axis2
1	Azuay	443	603
2	Bolivar	204	771
3	Cañar	39	110
4	Carchi	445	691
5	Cotopaxi	3	593
6	Chimborazo	299	432
7	Imbabura	643	654
8	Loja	242	270
9	Pichincha	568	605
10	Tungurahua	527	696
11	Santo Domingo de los Tsáchilas	155	881
12	El Oro	400	748
13	Esmeraldas	562	809
14	Guayas	366	872
15	Los Ríos	9	333
16	Manabí	11	252
17	Santa Elena	130	221
18	Morona Santiago	764	826
19	Napo	495	655
20	Pastaza	981	982
21	Zamora Chinchipe	610	615
22	Sucumbíos	738	796
23	Orellana	847	859

En la Figura 1, se pueden destacar tres clústeres, donde se caracterizó al clúster 1 con las provincias de alta concentración y desarrollo; el clúster 2, agrupó las provincias medianamente desarrolladas, mientras que el clúster 3 son las provincias de poco desarrollo, evidenciando una agrupación de forma más homogénea en los clústeres con las variables demográficas del estudio.

La clasificación con la técnica del HJ-Biplot, describe al clúster 1 por tener valores altos de densidad poblacional, población urbana, mortalidad general, mortalidad infantil y esperanza de vida. En contraste, se observan bajas tasas de natalidad, este grupo es considerado por tener provincias con alta concentración y desarrollo; el clúster 2, presenta valores altos de esperanza de vida, tasas de mortalidad general y mortalidad

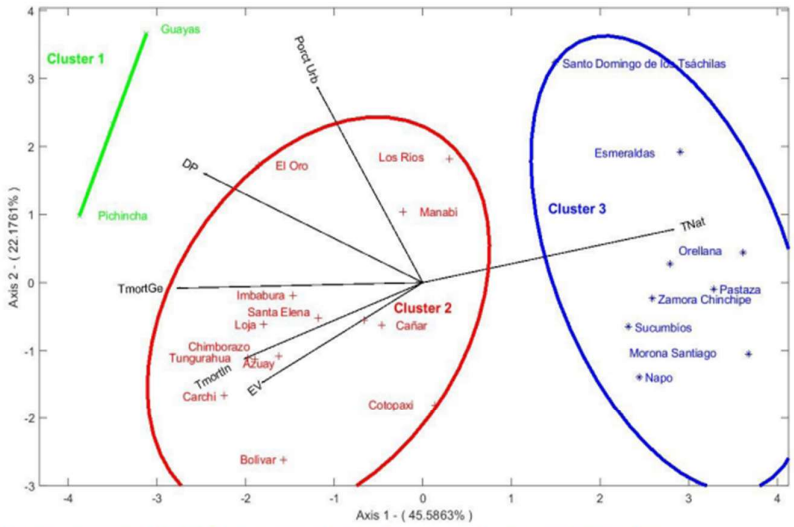


Fig. 1. Clúster y representación HJ-Biplot de provincias del Ecuador según variables demográficas.

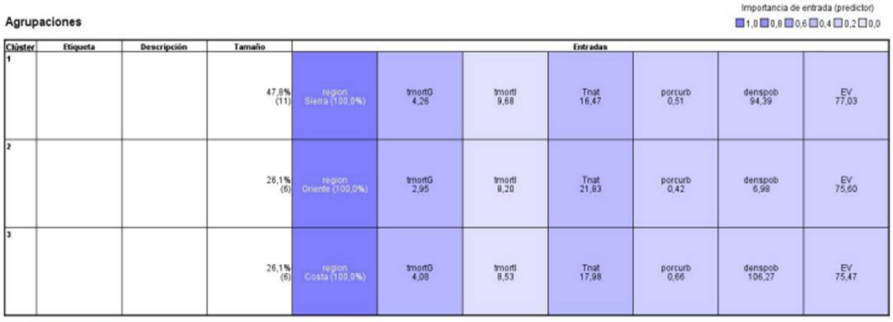


Fig. 2. Clúster Bietápico de regiones del Ecuador y variables demográficas.

infantil. En contraste tiene valores bajos de tasas de natalidad, densidad poblacional y población urbana; sin embargo, las provincias de El Oro, Los Ríos y Manabí muestran valores altos de población urbana. Este clúster se lo identifica como provincias medianamente desarrolladas; finalmente el clúster 3, se caracterizan por tener valores altos de tasa de natalidad. En contraste tienen valores bajos de esperanza de vida, densidad poblacional, población urbana, tasa de mortalidad infantil. Estas provincias se las identifica como de poco desarrollo.

Resultados similares se puede observar en el estudio realizado por Lucila Blanco y Carlos Mujica, donde se propuso una metodología aplicando la técnica de escalamiento multidimensional (MDS), para definir una configuración de países latinoamericanos con respecto a características sociodemográficas y económicas. Los resultados sugieren que las técnicas aplicadas pueden ofrecer una visión global del comportamiento sociodemográfico y económico de los países latinoamericanos, apoyados de análisis multivariante (Blanco & Mujica, 1996).

Según los resultados mostrados en la Figura 2, se identificaron tres clústeres donde la clasificación fue según la variable tipo de región, así se tiene que el clúster 1 representa el 47,8% y corresponde a la región sierra, el clúster 2 representa el 26,1% y corresponde a la región del Oriente y el clúster 3 representa el 26,1% y corresponde a la región costa.

En el clúster 1 se encuentran las 11 provincias que corresponde a la región Sierra, las mismas que son: Bolívar, Cotopaxi, Carchi, Azuay, Tungurahua, Chimborazo, Loja, Cañar, Imbabura, Santo Domingo de los Tsáchilas y Pichincha; en el clúster 2, están 6 provincias de la región Oriente que son: Orellana, Pastaza, Zamora Chinchipe, Sucumbíos, Morona Santiago, Napo; y en el clúster 3 están 6 provincias que corresponden a la región Costa: Guayas, Santa Elena, Los Ríos, Manabí, El Oro y Esmeraldas.

Con la técnica de Two Step, se obtuvo que la mayoría de los valores promedio del clúster 1 y 3 son similares, se diferencia porque el clúster 1 presenta mayor valor promedio de la tasa de mortalidad general, tasa de mortalidad infantil y de esperanza de vida; en cambio el clúster 3, presenta mayor promedio en la población urbana y densidad poblacional. El clúster 2, tiene mayor valor promedio de la tasa de natalidad y promedios bajos de densidad poblacional; dicho método, evidenció tres clústeres que son similares a la variable región, siendo una estructura distinta al método antes descrito.

Los resultados obtenidos según las técnicas de clúster reflejan la existencia de tres grandes estructuras sociodemográficas en el Ecuador; siendo una estructura, con la mayor concentración y desarrollo, la segunda estructura de mediano desarrollo y finalmente una tercera estructura de poco desarrollo.

Esta investigación, provee información sobre técnicas de clasificación aplicados a variables demográficas e indicadores básicos de salud medidos en las provincias del Ecuador; la identificación de clúster en provincias del Ecuador con características similares dentro del clúster y diferentes entre los mismos, ayudan a comprender la importancia de las técnicas de clasificación que permiten combinar variables socio-demográficas y las interacciones con las áreas geográficas (Mueller et al., 2019).

Limitaciones

Se han representado diferentes métodos para demostrar su utilidad y la aplicación en variables sociodemográficas, sería importante evaluar otras variables que complemente a este estudio a fin de brindar una información adecuada que refleje grupos prioritarios en la parte sanitaria, y por ende sea de apoyo a la toma de decisiones.

CONCLUSIONES

El HJ-Biplot, como técnica de representación gráfica multivariante, ha demostrado ser de mejor utilidad, ya que permite conocer las relaciones entre las variables, realizando un adecuado ordenamiento de los individuos y clasificación de los clústeres, siendo un aporte al análisis multivariante, reflejando información de posibles estructuras de conglomerados de las variables sociodemográficas medidas en las provincias del Ecuador.

Por otro lado, el método Two Step identificó clústeres en función de las variables estudiadas distintos al método HJ-Biplot, obteniéndose de este último una agrupación de forma más homogénea en los clústeres con las variables demográficas del estudio; con lo cual basados en estos análisis, se obtuvieron tres estructuras de desarrollo a nivel nacional.

JARAMILLO, L.; GAINDO, M. & REAL, J. Cluster analysis for big data: an application with demographic variables in provinces of Ecuador. *J. health med. sci., 6(1):*45-50, 2020.

ABSTRACT: The classification methods allow to explore and analyze big data sets visually, which is very useful for making quick decisions. This work aimed to compare of two methods of cluster analysis for big data in demographic variables of the provinces of Ecuador. An observational study of comparative type was carried out through the simultaneous representation of the HJ/Biplot and the Two Step method (two-stage cluster), through the MultBiplot and SPSS software. The data correspond to demographic variables of socio-health interest, general mortality rate, infant mortality rate, birth rate, population density, urban percentage and life expectancy, measured in the provinces of Ecuador. Data from Statistics and Census Institute were used. The association between variables was analyzed and clusters of the provinces of Ecuador were identified according to these demographic variables. According to the simultaneous representation of the HJ-Biplot, 3 clusters were identified, cluster 1 are provinces

with higher population density and general mortality rates, but low birth rates values, cluster 2 are provinces with higher life expectancy and mortality rates infantile but low birth rate values and cluster 3 are the provinces with high birth rates values and low population density, life expectancy, general mortality and infant mortality rates, different results were obtained with the Two Step method. It was concluded that these methods are useful for exploring the similarities between provinces according to demographic variables.

KEY WORDS: cluster, demographic, HJ-Biplot, two step method.

REFERENCIAS BIBLIOGRÁFICAS

Blanco, L. & Mujica, C. Representación de variables sobre una configuración de objetos obtenida a través de un escalamiento multidimensional. Rev. Venez. Anál. Coyunt., 4(2):223-36, 1998.

Bacher, J.; Wenzig, K. & Vogler, M. SPSS TwoStep Cluster-a first evaluation. SSOAR Social Sciences Open Access Repository, 2004.

Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453–67, 1971.

Gonzalez, M. V. Modelo extendidos para el análisis espacial en epidemiología del cáncer. Universidad Nacional de Córdoba. Trabajo de Tesis para optar al Título de Magister en Estadística Aplicada, 2015.

Lucio, R.; Villacrés, N. & Henríquez, R. Sistema de salud de Ecuador. Salud Pública Méx., 53(2):s177–s87, 2011. Disponible en: https://www.redalyc.org/pdf/106/10619779013.pdf

Mueller, E.; Sandoval, J. S. O.; Mudigonda, S. & Elliott, M. 2019. A Cluster-Based Machine Learning Ensemble Approach for Geospatial Data: Estimation of Health Insurance Status in Missouri. ISPRS Int. J. Geo-Inf., 8(1):13, 2019. Disponible en: https://doi.org/10.3390/ijgi8010013

Santos, C. Two-step Cluster" en SPSS y técnicas relacionadas. Universidad de Salamanca. Máster en Análisis Avanzado de Datos Multivariantes. Trabajo de Fin de Máster. 2015.

Şchiopu, D. Applying TwoStep cluster analysis for identifying bank customers' profile. Buletinul, 62(3): 66–75, 2010.

Shirkhorshidi, A. S.; Aghabozorgi, S.; Wah, T. Y. & Herawan, T. Big data clustering: a review. International Conference on Computational Science and Its Applications. ICCSA 2014. Springer, pp. 707–20, 2014.

Organización Panamericana de la Salud (OPS). Situación de la salud, 2020. Disponible en: https://www.paho.org/ecu/index.php?option=com_content&view=article&id=25:situacion-salud&Itemid=135

Verhasselt, Y. & Mansourian, B. Método para la clasificación de los países de acuerdo con sus indicadores de salud, 1991. Disponible en: https://iris.paho.org/handle/10665.2/16636

Vicente, J. MULTBIPLOT: A package for Multivariate Analysis using Biplots. Departamento de Estadistica. Universidad de Salamanca, 2010.

Vicente, J. Introducción al análisis de clúster. Departamento de Estadística. Universidad de Salamanca. 22pp., 2007.

Galindo, M. Una alternativa de representacion simultánea: HJ-Biplot. Qüestiió: quaderns d'estadística i investigació operative, 10(1):13–23, 1986.

Zhang, T.; Ramakrishnan, R. & Livny, M. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, 25(2):103–14, 1996.

Direccion para Correspondencia:
Leyda Jaramillo Feijoo
Departamento Gestión de la Información y Productividad SOLCA- Guayaquil
Av. Pedro Menéndez Gilbert y Atahualpa, parroquia Tarqui.
Guayaquil
ECUADOR
Teléfono: (593) 3718300

Email:
leydaj14@hotmail.com; ljaramillo@solca.med.ec

Recibido: 20-12-2019
Aceptado: 30-01-2020

ARTÍCULO 2:

ARTÍCULO 2: “CLÚSTER ESPACIAL DE MORTALIDAD POR CÁNCER DE MAMA EN ECUADOR”

J. health med. sci.,
6(1):29-36, 2020.

Clúster Espacial de Mortalidad por Cáncer de Mama en Ecuador

Space Cluster of mortality for breast Cancer in Ecuador

Jaramillo-Feijoo Leyda Elizabeth[1]; Galindo-Villardon María Purificación[2]; Real-Cotto Jhony Joe[3]; González-Rugel José Luis[4] & Idrovo-Madezco Segundo Enrique[5]

JARAMILLO, L.; GALINDO, M.; REAL, J.; GONZÁLEZ, J. & IDROVO, S. Clúster espacial de mortalidad por cáncer de mama en Ecuador. *J. health med. sci., 6(1)*:29-36, 2020.

RESUMEN: En la actualidad, los análisis de distribución espacial mediante el uso de técnicas de clusters para enfermedades crónicas como el cáncer de mama, son revelantes para la identificación de patrones espaciales de la mortalidad por cáncer según áreas geográficas. Identificar clústeres espaciales de la mortalidad por cáncer de mama en mujeres a nivel de las provincias del Ecuador, entre 2004 al 2018. Estudio observacional, de tipo descriptivo, ecológico multigrupal que compara a nivel espacio – temporal las tasas de mortalidad por cáncer de mama en mujeres según las provincias del Ecuador, utilizando el índice de Móran para el análisis de autocorrelación y el algoritmo de k-medias para el análisis de agrupamiento en períodos quinquenales mediante el programa informático ArcGIS versión 10.5. Resultados. En el Ecuador, el 86,5% de las muertes por cáncer de mama en mujeres se registraron en el área urbana, dichas muertes tienen un patrón no aleatorio según el índice de Morán, distinto al área rural que tiene un patrón aleatorio; se identificó diferencia en el agrupamiento de la mortalidad por cáncer de mama en las provincias urbanas y rurales, donde se obtuvo para el área urbana, clústeres con altas, media-altas, media-baja y bajas tasas de mortalidad, mientras que en lo rural se obtuvieron solo clústeres con altas, medias y bajas tasas de mortalidad. La distribución espacial y el análisis de agrupamiento identificó clústeres de la mortalidad por cáncer de mama en el Ecuador, evidenciando entre lo urbano y rural diferencias en los clústeres obtenidos, siendo esta información de utilidad para la implementación de estrategias de control del cáncer en el país.

PALABRAS CLAVE: clúster espacial, análisis de agrupamiento, cáncer de mama, mortalidad.

INTRODUCCIÓN

En el año 2015, el cáncer ocasionó 8.8 millones de defunciones, cerca del 70% de las muertes por cáncer se registran en países de ingresos medios y bajos; (OMS, 2018) (Brome et al, 2018) siendo el cáncer de mama una de las principales causas de muerte en mujeres en el mundo, y en América Latina también constituye la primera causa de muerte por neoplasias malignas femeninas que incluso en diversos países ha desplazado al cáncer de cuello del útero (Ramos *et al*., 2015); según datos del Instituto Nacional de Estadísticas y Censos (INEC), en el Ecuador, durante el año 2016 se registraron 641 muertes por cáncer de mama. (INEC, 2017) Además, es considerado como el tumor maligno más frecuente en mujeres, con alrededor de 1.2 millones de casos que se diagnostican a nivel mundial. (Martín *et al.*, 2015).

El análisis espacial es el conjunto de técnicas que analiza la dinámica de los eventos en salud según su zona geográfica y atributos de la misma, con la finalidad de identificar patrones espaciales; así también, el análisis de cluster espacial identifica aumento de casos en ubicaciones específicas o un patrón inusual. (Valbuena & Rodríguez, 2018). Los diversos países en especial los de América Latina son los que tienen mayores tasas de incidencia y mortalidad por cáncer, entre ellos la Argentina, donde vieron necesario representar de manera espacial la distribución del cáncer de mama, para explicar las

[1] Ingeniera en Estadística e Informática. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.
[2] Vicerrectora de ordenación académica y profesorado en la Universidad de Salamanca, España.
[3] Docente de la Universidad de Guayaquil. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.
[4] Docente de la Universidad Espíritu Santo.
[5] Médico internista del Hospital Clínica "San Francisco".

variaciones de la incidencia o mortalidad entre las diferentes áreas geográficas (Tumas *et al.*, 2017).

Actualmente, existen diversos métodos de estadística espacial para identificar patrones de distribución y clústeres de enfermedades como las crónicas según las áreas geográficas tanto en su incidencia y mortalidad, en el país se tiene poca evidencia del uso de estos métodos de agrupamiento, siendo importante aplicar dichas técnicas para enfermedades crónicas como el cáncer de mama, con la finalidad de descubrir estructuras espaciales o comportamiento de la distribución espacial en un periodo de tiempo, observando la variabilidad de las tasas de mortalidad por este tipo de cáncer; por lo antes mencionado, este artículo tuvo como objetivo identificar clústeres espaciales de la mortalidad por cáncer de mama en mujeres según las provincias del Ecuador.

MATERIAL Y MÉTODO

Población

Se realizó un estudio de enfoque cuantitativo, de diseño observacional, de tipo descriptivo, ecológico multigrupal que compara a nivel espacio – temporal las tasas de mortalidad por cáncer de mama en mujeres según las provincias del Ecuador (Valbuena & Rodríguez). La población fueron mujeres fallecidas por cáncer de mama en el periodo del 2004 al 2018, registradas por el INEC; las variables consideradas fueron: año de fallecimiento, tipo de cáncer, provincia de residencia, área de residencia, tasa de mortalidad, coordenada x y coordenada y.

Procedimiento y estadística

Para el análisis de agrupamiento espacial se trabajaron con las tasas de mortalidad cuyo cálculo es el número de muertes por cáncer de mama entre la población en riesgo por 100 000 habitantes en cada provincia; así como para el área urbana y rural, (Betanzos *et al.*, 2017) con las proyecciones de población por año del INEC. Además, se utilizó la división geopolítica de las provincias del Ecuador; excluyéndose las áreas no delimitadas; el análisis fue realizado en el software de sistemas de información geográfico ArcGIS versión 10.5 (ESRI, 2020) (ArcMap, 2018).

El periodo 2004 al 2018, se lo dividió en quinquenios, donde se calcularon tasas de mortalidad en cada periodo según provincia de residencia, tanto en el área urbana y rural. Luego se realizó un análisis de agrupamiento en las provincias según la tasa de mortalidad, con las variables discriminatorias que fueron los tres periodos del estudio (2004-2008, 2009-2013 y 2014-2018), lo que permitió identificar provincias con características homogéneas de la mortalidad por cáncer de mama.

La identificación de clústeres es una clasificación que intenta encontrar grupos basados en atributos de características y restricciones espaciales o temporales, siendo la mejor solución aquella donde las características dentro de cada grupo sean lo más homogéneas posibles y heterogénas entre los grupos. El método de agrupamiento utiliza un algoritmo de k-medias, que consiste en particionar un conjunto de n observaciones en k grupos, en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Para cada división la mejor solución es la que maximiza tanto la similitud dentro del grupo como la diferencia entre grupos. El análisis de agrupamiento calcula un valor R2 que indica la variabilidad de los datos, cuanto mayor sea el valor R2 mejor será esa variable para discriminar entre sus características; para este estudio se consideró un R2 mayor al 55% (porcentaje de variabilidad) mientras más cercano al 100% la variable será válida para discriminar entre los distintos grupos (Schabenberger & Gotway, 2017).

Además, se utilizó el índice de Morán para medir la autocorrelación espacial, que se caracteriza por la correlación de una señal entre otras áreas en el espacio, los valores oscilan entre -1 que indica dispersión perfecta a 1 correlación perfecta y un valor de cero indica un patrón espacial aleatorio, usando un nivel de confianza menor a 0,05 para indicar autocorrelación espacial (Tumas *et al.*) (Rocha *et al.*, 2017) (Aponte *et al.*, 2015).

Ética

Cabe indicar, que se trabajó con grupos de población y áreas geográficas, por el cual, no se ha tomado referencia alguna de las personas involucradas en esta investigación; y se solicitó la autorización respectiva a los directivos de SOLCA Guayaquil.

RESULTADOS

La Figura 1 analiza 7149 muertes en mujeres por cáncer de mama en el Ecuador, durante el periodo

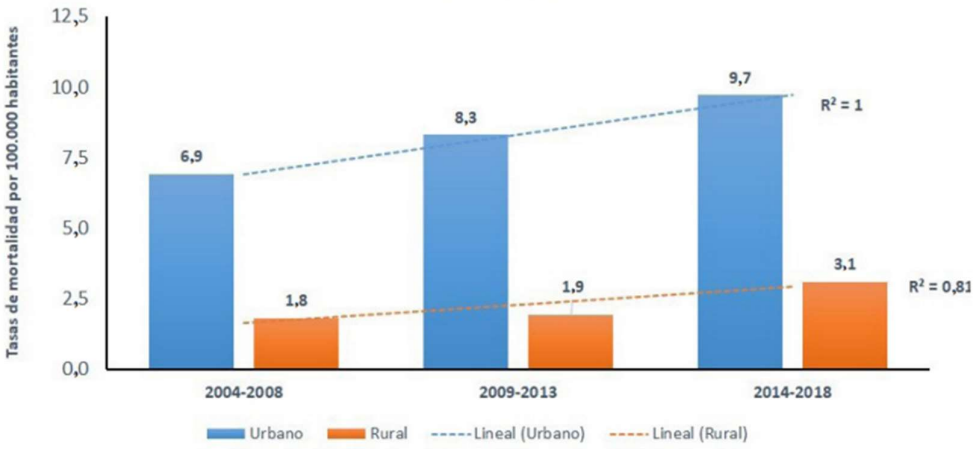


Fig. 1. Tasas de mortalidad en mujeres por cáncer de mama según área urbana y rural del Ecuador. 2004-2018.

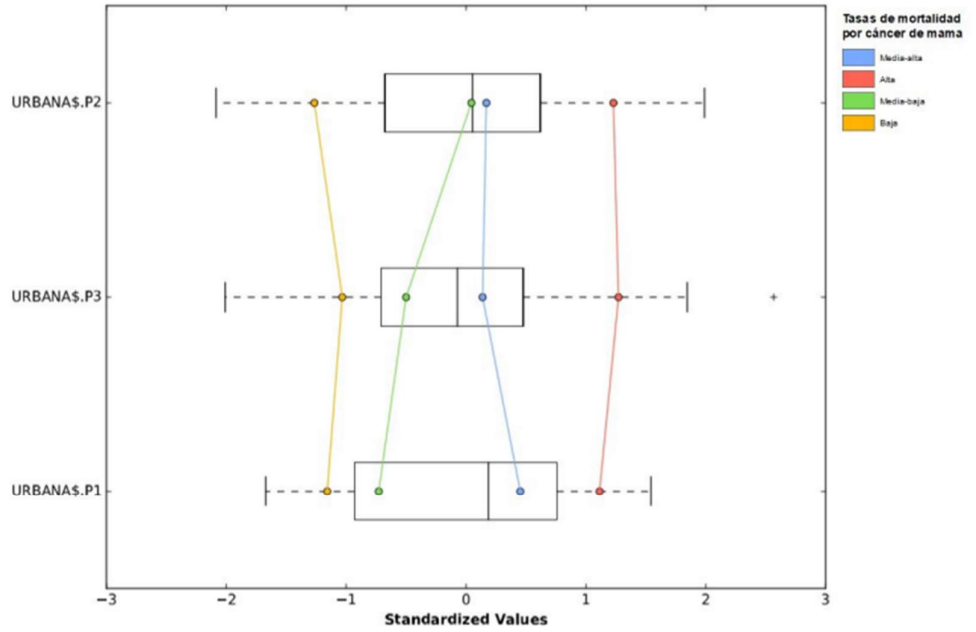


Fig. 2. Diagrama de caja en paralelo de los promedios de las tasas de mortalidad por cáncer de mama en el área urbana.
*P1 corresponde al periodo 2004-2008; P2 al periodo 2009-2013; y P3 al periodo 2014-2018.

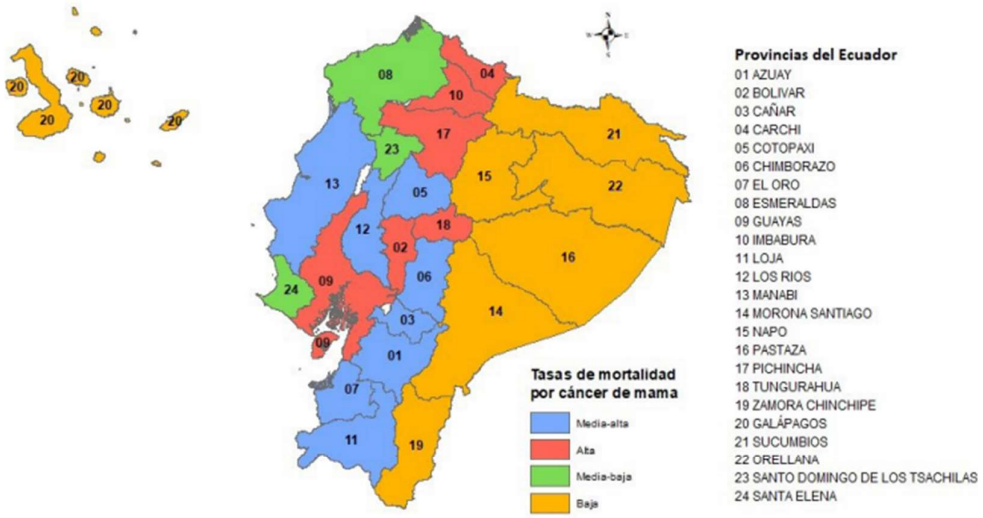


Fig. 3. Agrupamiento espacial de la mortalidad por cáncer de mama en provincias del área urbana, Ecuador 2004-2018.

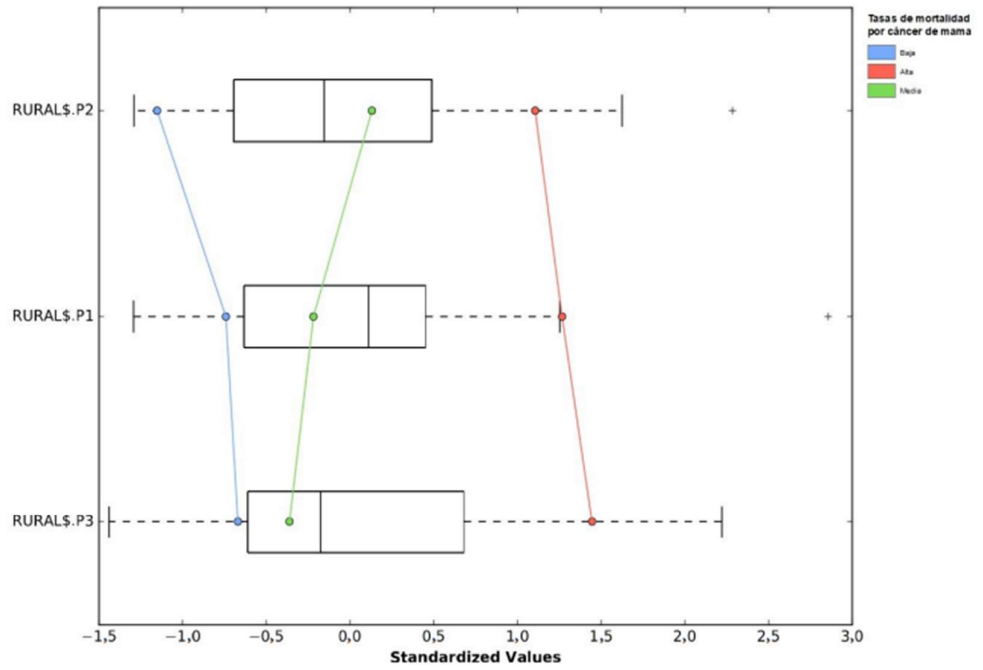


Fig. 4. Diagrama de caja en paralelo de los promedios de las tasas de mortalidad por cáncer de mama en el área rural.
*P1 corresponde al periodo 2004-2008; P2 al periodo 2009-2013; y P3 al periodo 2014-2018.

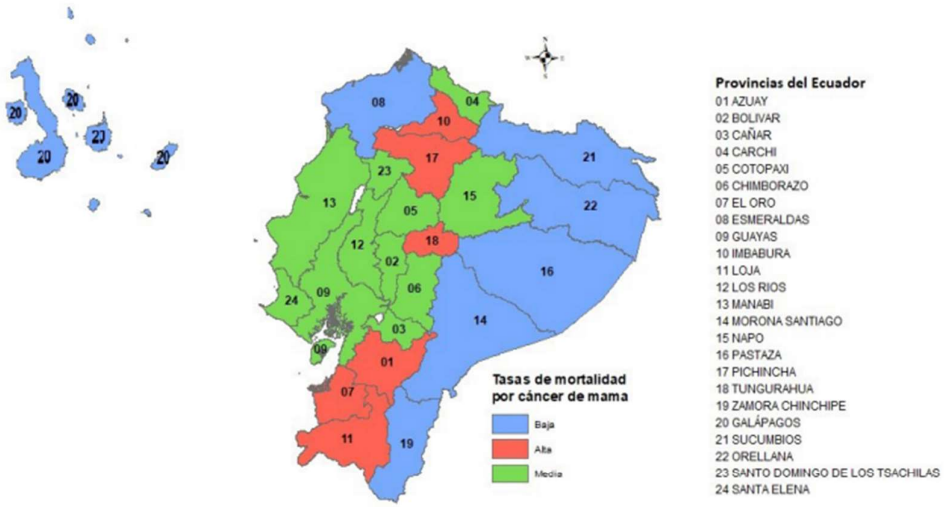


Fig. 5. Agrupamiento espacial de la mortalidad por cáncer de mama en provincias del área rural, Ecuador 2004-2018.

2004 al 2018; obteniéndose el 86,5% para el área urbana y 13,4% en la parte rural.

Basado en los tres periodos quinquenales, el método de clasificación mediante análisis de agrupamiento de provincias para el área urbana identificó cuatro grupos proporcionando la mejor diferenciación entre los mismos como se observa en la Figura 2, resultando un R2 mayor al 75% en los tres periodos, evidenciando una alta variabilidad en los datos, lo cual indica, que estas variables son mejores discriminadores. En la Figura 3, se muestra la mortalidad por cáncer de mama para los cuatro grupos de provincias del área urbana.

En el área rural, mediante el análisis de agrupamiento de las provincias, se obtuvieron tres grupos que diferenciaron de forma adecuada las tasas de mortalidad en los tres periodos quinquenales, como se observa en la Figura 4, resultando un R2 del 58% en el periodo 2004 – 2008, del 70% en 2009 – 2013 y del 71% para el periodo 2014 – 2018, indicando una alta variabilidad en los datos, siendo variables que mejor discriminan al momento de aplicar el análisis de agrupamiento. En la Figura 5, se muestra la mortalidad por cáncer de mama para los tres grupos de provincias del área rural.

DISCUSIÓN

La Figura 1 mostró el 41% de incremento en la tasa de mortalidad por cáncer de mama en los últimos 15 años para el área urbana, incrementándose la tasa de 6,9 en el periodo 2004-2008 a 9,7 en 2014 – 2018 por cada 100 000 habitantes; asimismo, en el área rural el incremento fue 72%, siendo la tasa 1,8 en el periodo 2004-2008 a 3,1 en el 2014 – 2018 por 100 000 habitantes.

Se aplicó el índice de Morán en el área urbana y rural, para la determinación del comportamiento espacial que tienen los datos; observándose en el área urbana para los tres periodos un resultado de un índice positivo con un valor p menor al 10%; lo cual indica que los datos se comportan de forma espacial de manera concordante, es decir, se agrupan valores altos con altos y valores bajos con bajos. Para el área rural, el índice en el periodo de estudio fue negativo con un valor p mayor al 10%, donde indica que los datos tienen un comportamiento aleatorio; estos resultados son parecidos al estudio de identificación de determinantes sociodemográficas asociadas a la distribución espacial de la incidencia de cáncer de mama realizado en Argentina, donde la mayoría de las variables presentaron un índice de Moran significativo y positivo que indica no aleatoriedad de

la distribución espacial; así como, valores pequeños que indican procesos espaciales subyacentes de forma aleatoria. (Tumas *et al*.)

En la Figura 3 se muestran los cuatro grupos de provincias del área urbana de la siguiente manera: en el grupo 1 consta 8 provincias que son: Manabí, Los Ríos, Cotopaxi, Chimborazo, Cañar, Azuay, El Oro y Loja, donde se registran tasas de mortalidad por cáncer de mama por encima de la media global en cada quinquenio y que incrementan en el tiempo; así se obtuvo, una tasa media de mortalidad en el 2004 – 2008, de 6,3, 2009 – 2013 de 7,2 y 2014 – 2018 de 8,6 por cada 100 000 habitantes, considerando a este grupo con tasas de mortalidad media – alta por cáncer de mama.

El grupo 2, está conformado por 6 provincias: Carchi, Imbabura, Pichincha, Tungurahua, Bolívar y Guayas, donde se registran tasas de mortalidad por cáncer de mama por encima del cuartil superior en los tres periodos de estudio, en el que se observa una tasa media de mortalidad en el 2004 – 2008 de 8,2, 2009 – 2013 de 10,3 y 2014 – 2018 de 12,2 por cada 100 000 habitantes, siendo este grupo que obtuvo las tasas de mortalidad más altas.

El grupo 3, agrupa 3 provincias: Esmeraldas, Santo Domingo de los Tsáchilas y Santa Elena, en el que se registran tasas de mortalidad por cáncer de mama por debajo de la media global en los periodos 2004 – 2008 y 2014 – 2018, con 2,7 y 6,5 por cada 100 000 habitantes respectivamente, siendo este grupo con las tasas de mortalidad media – baja.

El grupo 4, está constituida por 7 provincias: Sucumbíos, Napo, Orellana, Pastaza, Morona Santiago, Zamora Chinchipe y Galápagos, donde se registran tasas de mortalidad por cáncer de mama por debajo del cuartil inferior en cada quinquenio, obteniéndose para el periodo 2004-2008 una tasa de 1,5, 2009-2013 de 3,0 y 2014 – 2018 de 4,8 por cada 100 000 habitantes, por lo que, en este grupo se observó tasas de mortalidad bajas.

En la Figura 5, se observan los tres grupos de provincias del área rural: el grupo 1 comprende 7 provincias: Esmeraldas, Sucumbíos, Orellana, Pastaza, Morona Santiago, Zamora Chinchipe y Galápagos, donde se registran las tasas de mortalidad de cáncer de mama por debajo del cuartil inferior; obteniéndose, una tasa media de mortalidad en el periodo 2004 – 2008 de 0,5, 2009 – 2013 de 0,15 y 2014 – 2018 de 1,25 por cada 100 000 habitantes; por lo que este grupo obtuvo tasas de mortalidad bajas.

El grupo 2, comprende 6 provincias: Imbabura, Pichincha, Tungurahua, Azuay, El Oro y Loja, registrándose tasas de mortalidad por encima del cuartil superior en cada quinquenio, observándose, una tasa media de mortalidad en el 2004 – 2008 de 2,6, 2009 – 2013 de 2,7 y 2014 – 2018 de 4,6 por cada 100 000 habitantes, este grupo registró tasas de mortalidad altas.

El grupo 3, comprende 11 provincias: Santa Elena, Guayas, Manabí, Cañar, Los Ríos, Chimborazo, Bolívar, Cotopaxi, Napo, Santo Domingo y Carchi, en las que se registran tasas de mortalidad por cáncer de mama cercanas a la mediana (cuartil 2); así se obtuvo, una tasa media de mortalidad en el 2004 – 2008 de 1,1 2009 – 2013 de 1,6 y 2014 – 2018 de 1,7 por cada 100 000 habitantes, siendo este grupo con tasas de mortalidad medias.

Este estudio detectó diferencias en los clústeres según las tasas de mortalidad por cáncer de mama en el país, para el área urbana identificó 4 clústeres y en lo rural fueron 3 clústeres, los cuales se midieron en función de la variación de dichas tasas, obteniéndose en el área de urbana el cluster con altas tasas de mortalidad durante los tres periodos quinquenales con un incremento sostenido en dichos periodos; a diferencia del área rural que refleja variabilidad de los datos para identificar los grupos, teniendo tasas más bajas que el área urbana; sin embargo, el método agrupó a provincias con mayores tasas de mortalidad en el área rural, siendo similar al estudio de análisis espacio-temporal de eventos asociados al cáncer realizado en Cuba, donde se observó la distribución geográfica del cáncer de mama asociado a variables demográficas, evidenciando la distribución de la incidencia con predominio en la zonas urbanas. (Hernández *et al.*, 2012) Asimismo, es muy parecido al estudio sobre el Análisis estadístico espacial para la identificación de conglomerados de cáncer de mama realizado en la ciudad de La Paz que identificó posibles conglomerados de cáncer apoyados con sistemas de información geográfico y herramientas de análisis espacial. (Agúndez *et al*., 2018)

Se identificaron diferencias en la estructura espacial de agrupamiento entre el área urbana y rural, siendo las tasas de mortalidad por cáncer de

mama en los clústeres del área urbana más altos que en el área rural; estos resultados son similares al estudio realizado del análisis espacial por cluster de la mortalidad por cáncer de mama en la provincia de Shandong – China, donde se encontró que las tasas de mortalidad fueron más altas en áreas urbanas que en áreas rurales y además, obtuvieron entre el área urbana y rural diferencias en la distribución espacial y clústeres de cáncer de mama (Chu *et al.*, 2017)

Limitaciones

La aplicación de métodos estadísticos espaciales permite clasificar a un área geográfica según variables epidemiológicas durante un período de tiempo, sería importante evaluar otras variables demográficas o ambientales que permitan identificar posibles factores asociados o causales a los grupos generados en este estudio. Además, otra limitante fue no encontrar literatura especializada en el país de estudios similares que permitan comparar los resultados obtenidos.

CONCLUSIONES

El análisis de agrupamiento, identificó clústeres de mortalidad por cáncer de mama en las provincias del Ecuador, evidenciando entre lo urbano y rural diferencias en los clústeres obtenidos; a su vez, el área urbana presentó tasas con valores más altos y un patrón no aletorio. En el área rural se observaron tasas bajas y un patrón aleatorio según el índice de Morán, donde el agrupamiento de provincias visualizó diferencias en las tasas.

JARAMILLO, L.; GALINDO, M.; REAL, J.; GONZÁLEZ, J. & IDROVO, S. Space cluster of mortality for breast cancer in Ecuador. *J. health med. sci., 6(1)*:29-36, 2020.

ABSTRACT: Currently spatial distribution analyzes through the use of cluster techniques for chronic diseases such as breast cancer are revealing for the identification of spatial patterns of cancer mortality according to geographic areas. Objective. Identify spatial clusters of breast cancer mortality in women at the level of the provinces of Ecuador, between 2004 to 2018. We used an observational, descriptive, ecological multigroup study that compares at a Spatio-temporal level the rates of breast cancer mortality in women according to the provinces of Ecuador, using the Moran index for the autocorrelation analysis and the k-, means algorithm for cluster analysis in five-year periods using the ArcGIS version 10.5 software. Results. In Ecuador, 86.5% of breast cancer deaths in women were recorded in the urban area, these deaths have a non-random pattern according to the Morán Index different from the rural area that has a random pattern; difference was identified in the grouping of breast cancer mortality in urban and rural provinces, where it was obtained for urban areas, clusters with high, medium. high, medium-low and low mortality rates. While in rural areas only clusters with high, medium and low mortality rates were obtained. Conclusions. The spatial distribution and cluster analysis identified clusters of breast cancer mortality in Ecuador; evidencing between urban and rural differences in the clusters obtained, this information is useful for the development of cancer control strategies in the country.

KEY WORDS: Spatial cluster, cluster analysis, breast cancer, mortality.

REFERENCIAS BIBLIOGRÁFICAS

Agúndez, M.; Sánchez, C.; Martínez, G.; Romero, R. & Luna, J. Análisis estadístico espacial para la identificación de conglomerados de cáncer de mama en la ciudad de La Paz, BCS. Pistas Educativas, No. 114. México, Instituto Tecnológico de Celaya, 2015

Aponte, C.; Romero, E. & Santa, L. Análisis de datos espaciales del Índice de Necesidades Básicas Insatisfechas en la Región Andina. Perspectiva Geográfica, 20(2):391–418, 2015.

ArcMap. Análisis de agrupamiento. Ayuda. ArcGIS Desktop, 2020. Disponible en: https://desktop.arcgis.com/es/arcmap/latest/tools/spatial-statistics-toolbox/grouping-analysis.htm

Betanzos, F.; Escoto, M. & Chávez, J. Estadística aplicada en Psicología y Ciencias de la salud. Manual Moderno, 2017.

Brome, M.; Montoya, D. & Salcedo, L. Incidencia y mortalidad por cáncer en Medellín, Colombia. 2010-2014. Colom. Med., 49(1):81–88, 2018.

Chu, J.; Zhou, C.; Guo, X.; Sun, J.; Xue, F.; Zhang, J.; Lu, Z.; Fu, Z. & Xu, A. Female Breast Cancer Mortality Clusters in Shandong Province, China: A Spatial Analysis. Sci. Rep., 7(1):1–8, 2017.

ESRI. Clustering multivariante. ArcGIS Pro. ArcGIS Desktop. 2020. Disponible en: https://pro.arcgis.com/es/pro-app/tool-reference/spatial-statistics/multivariate-clustering.htm

Hernández, B.; Antón, O. & Alegret, M. Análisis espacio-temporal de eventos asociados al cáncer: una herramienta para apoyar estudios epidemiológicos. MediSur 10(2):171–181, 2012.

Instituto Nacional de Estadística y Censos (INEC). El cáncer de mama en Ecuador. 2017. Disponible en: https://www.ecuadorencifras.gob.ec/el-cancer-de-mama-en-ecuador/

Martín, M.; Herrero, A. & Echavarría, I. El cáncer de mama. Arbor, 191(773):a234, 2015.

Organización Mundial de la Salud (OMS). Cáncer. Datos y Cifras. 2018. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/cancer

Rocha, A.; Castañeda, R. & Brum, I. Análise espacial exploratória com o emprego do índice de moran. GEOgraphia 19(4):161–179, 2017.

Ramos, Y.; Marimón, E.; Crespo, C.; Junco, B. & Valiente, W. 2015. Cáncer de mama, su caracterización epidemiológica. Rev. Ciencias Médicas, 19(4):619–29, 2015.

Schabenberger, O. & Gotway, C. Statistical methods for spatial data analysis. Chapman and Hall/CRC. 1st Edition, New York, 2017.

Tumas, N.; Pou, S. & Díaz, M. Inequidades en salud: análisis sociodemográfico y espacial del cáncer de mama en mujeres de Córdoba, Argentina. Gac. Sanit., 31(5):396–403, 2017.

Valbuena, A. & Rodríguez, L. Análisis espacial en epidemiología: revisión de métodos. Rev. Univ. Ind. Santander. Salud, 50(4):358-65, 2018.

Direccion para Correspondencia:
Leyda Jaramillo Feijoo
Departamento Gestión de la Información y Productividad SOLCA- Guayaquil
Av. Pedro Menéndez Gilbert y Atahualpa, parroquia Tarqui.
Guayaquil
ECUADOR
Teléfono: (593) 3718300

Email:
leydaj14@hotmail.com; ljaramillo@solca.med.ec

Recibido: 16-12-2019
Aceptado: 25-01-2020

ARTÍCULO 3:

ARTÍCULO 3: "BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB): Una aplicación a datos de mortalidad por cáncer de mama en el Ecuador"

(En proceso de publicación)

Artículo original

SPATIO TEMPORAL WEIGTHED LOGISTIC BIPLOT (TSWLB): An application to breast cancer mortality data in Ecuador.

BIPLOT LOGÍSTICO PONDERADO ESPACIO TEMPORAL (TSWLB): Una aplicación a datos de mortalidad por cáncer de mama en el Ecuador.

Jaramillo-Feijoo Leyda[1,2], Galindo-Villardon Purificación[1,3,4], González-Rugel José[6]; Real-Cotto Jhony[2,5]

1 Department of Statistics, University of Salamanca, 37008 Salamanca, Spain.
2 Departamento Gestión de la Información y Productividad, Hospital SOLCA, Guayaquil 090514, Ecuador
3 Escuela Superior Politécnica del Litoral, ESPOL, Centro de Estudios e Investigaciones Estadísticas, Campus Gustavo Galindo, Km. 30.5 vía Perimetral, Guayaquil Apartado postal 09-01-5863, Ecuador.
4 Universidad Estatal de Milagro, UNEMI, Centro de Gestión de Estudios Estadísticos. Ciudadela Universitaria Km. 1.5 vía al Km. 26, Guayas 091050, Ecuador
5 Facultad de Ciencias Médicas, Universidad de Guayaquil, Guayaquil 090514 Ecuador
6 Facultad de Administración, Universidad Espíritu Santo, Guayaquil 170301, Ecuador

Grado de contribución:
Conceptualización LJ, PG, JG, JR; metodología, LJ, PG, JR; validación, LJ, PG; depuración de datos, LJ, PG, JG; aplicación de métodos espacio-temporales, LJ, PG, JG, JR; realización de análisis, redacción y edición, LJ, PG, JG, JR.

Correspondencia:
Leyda Jaramillo Feijoo: Departamento Gestión de la Información y Productividad SOLCA- Guayaquil. Av. Pedro Menéndez Gilbert y Atahualpa, parroquia Tarqui, Guayaquil – Ecuador; teléfono (593) 3718300 extensión 2464; 992813865. Correo electrónico: leydaj14@hotmail.com; leyda.e.jaramillo@solca.med.ec

Resumen

El objetivo del estudio es combinar las técnicas GWPCA y la prueba estadística no paramétrica Mann-Kendall que son ampliamente usadas para analizar la componente espacial y temporal siendo aplicadas de forma individual y no existiendo una representación simultanea de dichas técnicas. En este artículo se propone un Biplot Logístico ponderado espacio temporal (TSWLB) que es una técnica multivariante que combina los componentes espacial y temporal para representarlas en un gráfico que permite una fácil interpretación de las relaciones entre los sitios geográficos y las variables, es de interés en varias áreas siendo aplicado en la mortalidad por cáncer de mama en el Ecuador. Se utilizó el paquete GWmodel y la librería Kendall de R y el programa MultiBiplot. Se observó un incremento sostenido de las tasas de mortalidad por cáncer de mama en el Ecuador. Con una mayor variabilidad de las muertes por esta enfermedad al norte y sur del país. La técnica TSWLB representó simultáneamente las características espacio temporales dando un ordenamiento a los sitios geográficos e identificando 4 clústeres, siendo el clúster 2 con las provincias de Guayas, El Oro, Santo Domingo y Chimborazo el más prioritario por tener una tendencia creciente significativa estadísticamente de la mortalidad por cáncer de mama y con presencia de altas tasas de mortalidad en años recientes, orientando las intervenciones en salud por esta enfermedad.

Palabras clave: Geográficamente ponderado, tendencia temporal, análisis espacial, mortalidad, cáncer de mama

ABSTRACT

The objective of the study is to combine the GWPCA techniques and the Mann-Kendall non-parametric statistical test, which are widely used to analyze the spatial and temporal component, being applied individually and there being no simultaneous representation of said techniques. In this article, is proposed a spatio temporal weigthed logistic biplot (TSWLB), which is a multivariate technique that combines spatial and temporal components to represent them in a graph that allows easy interpretation of the relationships between geographic sites and variables, it is of interest in various areas. Applied in mortality from breast cancer in Ecuador. The GWmodel package and the Kendall R library and the MultiBiplot program were used. An increase in breast cancer mortality rates was observed in Ecuador. With a greater variability of deaths from this disease in the north and south of the country. The TSWLB technique simultaneously represented the spatio-temporal characteristics, ordering the geographical sites and identifying 4 clusters, the cluster 2 with the provinces of Guayas, El Oro, Santo Domingo and Chimborazo have the highest priority for having a statistically significant increasing trend of the mortality from breast cancer and with the presence of high mortality rates in recent years, guiding health interventions for this disease

Keywords: Geographically weighted, temporal trend, spatial analysis, mortality, breast cancer.

Introducción

El análisis de la triada epidemiológica es tiempo, lugar y persona. Los análisis espacial y temporal son herramientas importantes en las investigaciones en salud, uno de los campos es la epidemiología espacial que describe la variación espacial en relación al riesgo de enfermedad, identifica la correlación geográfica de los factores de riesgo en relación con los resultados de salud medidos en un entorno geográfico, patrones de distribución espacial y su evolución en el tiempo, resultando análisis mucho más enriquecedores, (Puranik et al., 2020) (Souris, 2019) (Pou et al., 2019) sin embargo, es un desafío incorporar la dimensión temporal en los análisis espaciales, por la complejidad de los modelos espaciotemporales. (Fotheringham et al., 2015)

Existen varias técnicas que se utilizan para los análisis espaciotemporales, que permiten identificar conglomerados de algún tipo de enfermedad y generar un mapa de riesgo y su evolución en el tiempo, (Santamaría Ulloa, 2003) en función de su incidencia y mortalidad, siendo importante las técnicas multivariantes exploratorias con enfoque geográficamente ponderado y temporal para determinar patrones espaciales, evaluar la heterogeneidad espacial en el tiempo y correlación en los datos geoespaciales. (Hernández et al., 2012)

Las técnicas análisis de componentes principales ponderados geográficamente GWPCA y la prueba estadística no paramétrica Mann-Kendall son ampliamente usadas para analizar la componente espacial y temporal, es de anotar que, se han desarrollado varias técnicas con el enfoque GW que incluyen: regresión ponderada geográficamente GWR propuesto por (Brunsdon et al., 1996) que permite conocer las distintas relaciones espaciales en diferentes puntos del espacio geográfico y sugiere que cualquier modelo que puede ser ponderado puede ser ponderado geográficamente; además, se tiene las estadísticas de resumen de GW (Brunsdon et al., 2002) el cual utiliza la estimación de la función de Kernel para ponderar geográficamente los puntos y obtener un resumen de las estadísticas y relaciones espaciales; así mismo el análisis discriminante ponderado geográficamente GWDA (Brunsdon et al., 2007) esta técnica adapta el enfoque del modelo GWR permitiendo el modelado y la predicción de variables de respuesta categóricas; de igual manera el modelo lineal generalizado ponderado geográficamente GWGLM, (Nakaya et al.,

2009) que es una ampliación de GWR para la predicción de variables no continuas, permitiendo el uso de distintos modelos de regresión, tales como regresión logística para datos binarios, regresión de Poisson para datos de conteo, que a través del predictor lineal estos modelos de regresión se integran como modelos lineales generalizados; finalmente, el GWPCA (Harris et al., 2011) evalúa la heterogeneidad y autocorrelación espacial para conocer la estructura espacial subyacente del conjunto de datos.

Cabe indicar que, para el uso de los modelos ponderados geográficamente antes descritos, fue desarrollado un paquete en el software estadístico R GWmodel. (Comber et al., 2020) Luego, se realizaron dos adaptaciones al modelo GWR, una con respecto a la temporalidad, que se denominó regresión ponderada geográfica y temporal GTWR, (Fotheringham et al., 2017) la cual analiza los efectos locales tanto en el espacio como en el tiempo, destacando la importancia de la temporalidad; y la otra con respecto a las escalas, es la regresión ponderada geográficamente multiescala MGWR que parte de la suposición que diferentes procesos operan a distintas escalas espaciales, para lo cual propone un vector de ancho de banda óptimo en el que cada elemento indica la escala espacial en la que tiene lugar un proceso particular. (Fotheringham et al., 2017)

Las técnicas multivariantes GWPCA y la prueba estadística no paramétrica Mann-Kendall se aplican de forma individual, genera resultados muy extensos y de difícil interpretación, por otro lado, no existen técnicas multivariantes que representen de manera simultánea dichas técnicas. Por lo que el objetivo de este estudio es proponer un Biplot Logístico ponderado espacio temporal (TSWLB) que es una técnica multivariante que combina los componentes espacial y temporal para representarlas en un gráfico que permite una fácil interpretación de las relaciones entre los sitios geográficos y las variables, es de interés en varias áreas siendo aplicado en la mortalidad por cáncer de mama en el Ecuador

Método propuesto: Biplot logístico ponderado espacio temporal (TSWLB)

El Biplot logístico ponderado espacio temporal es una técnica multivariantes que captura y representa simultáneamente las componentes espacial y temporal, siendo una metodología de priorización mediante el ordenamiento de los sitios geográficas en función de los años con cargas mayores a 0.4 y tendencias temporales crecientes o decrecientes; el análisis espacial es mediante el análisis de componentes principales geográficamente ponderados que mide la heterogeneidad espacial; el análisis temporal se realizó con la prueba estadística no paramétrica Mann-Kendall que determina si es una tendencia creciente o decreciente y si es estadísticamente significativa. Se excluyeron los años que no cumplieron el criterio con cargas mayores a 0.4 en ningún sitio geográfico, además, se consideró los sitios geográficos con tasas de mortalidad por cáncer de mama BCMR mayores a la tasa nacional durante el año más reciente (Fig. 1).

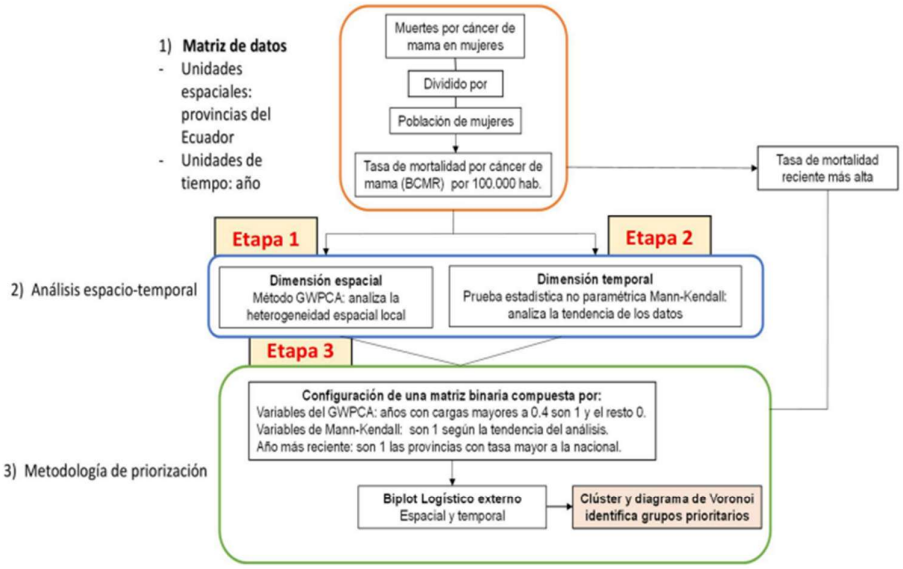


Fig. 1 Esquema del Biplot logístico ponderado espacio temporal (TSWLB)

Etapa 1: GWPCA

La técnica GWPCA examina la parte no estacionaria de los datos localmente en el espacio geográfico, en este estudio se utilizó GWPCA para las tasas de mortalidad por cáncer de mama BCMR del 2007 al 2021 medidos en las 24 provincias del Ecuador.

Así tenemos, para GWPCA un vector de variables observadas x_i en la ubicación espacial i se asume que sigue una distribución normal multivariante con vector media μ y matriz de varianza-covarianza ∑, tal que $x_i \sim N(\mu, \sum)$.

Además, la ubicación espacial i tiene coordenadas (u,v), entonces el PCA con efectos geográficos locales implica considerar a x_i como condicional en u y v, y haciendo a μ y ∑ funciones de u y v; por lo tanto, $x_i|(u,v) \sim N\left(\mu(u,v), \sum(u,v)\right)$. Así μ y ∑ son funciones de u y v, esto implica que cada elemento de μ(u, v) y ∑(u, v) es también función de u y v. Por lo tanto los momentos μ(u,v) y ∑(u,v) son el vector de media geográficamente ponderados (GW) y la matriz de varianza-covarianza geográficamente ponderado, respectivamente. Para obtener los componentes principales geográficamente ponderados la descomposición de la matriz de varianza-covarianza GW provee los valores propios GW y los vectores propios GW. El producto de la i fila de la matriz de datos con los vectores propios GW para la ubicación i provee la fila i de las puntuaciones de los componentes GW. La matriz de varianza-covarianza GW es

$$\sum(u,v) = X^T\, W(u,v) X$$

Donde W(u,v) es una matriz diagonal de pesos geográficos que puede ser generada usando una función de kernel. En el caso de estudio, se utilizó una función de kernel bi-square:

$$w_{ij} = \left(1 - \left({}^{d_{ij}}/_{r}\right)^2\right)^2 \quad si\ \ d_{ij} \le r; \quad w_{ij} = 0 \quad en\ otro\ caso$$

Donde el ancho de banda es la distancia geográfica r y d_{ij} es la distancia entre la ubicación espacial de la i y j filas en la matriz de datos X. Los componentes principales GW para la ubicación (u_i, v_i) puede escribirse como:

$$LVL^T|(u_i, v_i) = \sum(u_i, v_i)$$

Donde $= \sum(u_i, v_i)$ es la matriz de varianza-covarianza GW para la ubicación (u_i, v_i).

Etapa 2: Prueba estadística no paramétrica Mann-Kendall

La prueba estadística no paramétrica Mann-Kendall se utiliza para analizar los datos recopilados en un periodo de tiempo y determinar su tendencia si es creciente o decreciente. Para la aplicación de la prueba los datos no requieren que sigan alguna distribución en particular. El estadístico analiza las diferencias en los signos de las combinaciones de cada par de datos observados a lo largo del tiempo, es así, que comprueba si $BCMR_j > BCMR_i$ o $BCMR_j < BCMR_i$ y contabiliza los pares que aumentan o disminuyen en dicho periodo. Generando para cada unidad espacial la frecuencia relativa de incrementos menos la frecuencia relativa de las disminuciones. (Lalangui et al., 2022)

$$S = \frac{2(t-2)!}{t!} \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} sing(BCMR_j - BCMR_i)$$

Donde la función signo está dada por:

$$sign(BCMR_j - BCMR_i) = \begin{cases} 1\ si\ (BCMR_j - BCMR_i) > 0 \\ 0\ si\ (BCMR_j - BCMR_i) = 0 \\ -1\ si\ (BCMR_j - BCMR_i) < 0 \end{cases}$$

$BCMR_i$ es la BCMR en el año i ϵ {1,2,….t - 1} siendo t el número de años del periodo y $BCME_j$ es la BCMR en el año j = (i + 1) ϵ {1,2,….,t}.

Etapa 3: Biplot logístico ponderado espacio temporal (TSWLB)

Las metodologías antes señaladas de análisis de componentes principales geográficamente ponderado GWPCA y la prueba estadística no paramétrica de Mann-Kendall, han sido muy utilizados en la literatura en áreas de la salud, ambiente, agricultura, entre otros, en un contexto espacial y temporal, siendo aplicadas de forma separada, por ejemplo: (Tejedor Flores, 2018) analiza eficazmente el nexo entre el agua, la energía y los alimentos; (Lalangui et al., 2022) analiza la mortalidad infantil en el Ecuador; (Zymaroieva et al., 2019) analiza el rendimiento de la soya basado en las técnicas GWPCA.

Los resultados de las técnicas GWPCA y Mann-Kendall se interpretan de manera individual resultando algo complejo. Para integrar gráficamente tanto los resultados del GWPCA y la prueba de Mann-Kendall y caracterizar los sitios geográficos y su relación con las dimensiones espacial y temporal, se aplica un algoritmo, como el propuesto por (Vicente-Villardón et al., 2006) y que luego es ampliado por (Demey et al., 2008a) donde combina un análisis de coordenadas principales (PCoA) y una regresión logística (LR) para construir un biplot logístico externo. En el paso de PCoA se utilizó el coeficiente de Russel y Rao para datos dicotómicos, esto evita indeterminación en el cálculo, dado que hay pares de sitios geográficos en los cuales ninguna de las características está presente, generando d (dobles ausencias) en el denominador.

$$S_{RR} = \frac{a}{a+b+c+d}$$

S_{RR} coeficiente de Russell y Rao está acotado entre cero y uno; uno indica máxima similaridad y cero disimilaridad total.

Se utilizó el método de mínima varianza de Ward para crear clústeres, el cual es un procedimiento donde el criterio para la elección del par de clústeres a mezclar en cada paso se basa en el valor óptimo de una función objetivo, la suma de cuadrados de la varianza.

Previo a la aplicación de la técnica Biplot logístico, se configuro la matriz de datos binarios compuesta por variables obtenidas de las componentes principales del GWPCA y de la prueba Mann-Kendall. Para el GWPCA se partió de la matriz que consolida la máxima carga entre las tres primeras componentes de los años de estudio y para cada sitio geográfico, dicha matriz es convertida a binaria considerando el siguiente criterio, las cargas o correlaciones mayores a 0.4 son uno y el resto cero, es de mencionar que se excluyeron los años donde ningún sitio geográfico cumplió con el criterio. Para la prueba de Mann-Kendall, se identificaron variables según la tendencia que siguen los datos en cada sitio geográfico, es así, que se identificaron tres variables: tendencia creciente significativa (TSC), tendencia creciente no significativa (TNSC), tendencia decreciente no significativa (TNSD), finalmente, se agrega una variable que clasifica en el año 2021 como uno los sitios geográficos que presentaron una tasa de mortalidad por cáncer de mama superior a la tasa nacional y cero en el resto.

La técnica del biplot logístico representa a los sitios geográficos como puntos y las variables como vectores en un diagrama de dispersión en un espacio euclidiano. Los sitios geográficos con combinaciones similares en los años con mayor carga en su mortalidad y la tendencia en el tiempo se agrupan mientras que los sitios geográficos distintos tienden a separarse. Los vectores indican dirección y están más correlacionados con la presencia de la característica, esto permite caracterizar y dar un ordenamiento a los sitios geográficos en función de la dimensión espacial y temporal.

Para interpretar las características asociadas al agrupamiento de los sitios geográficos proyectamos los puntos sobre la dirección de los vectores, cuanto más lejos se proyecte el punto en la dirección de la flecha, mayor es la probabilidad de la característica. El origen del vector es el punto que predice una probabilidad 0.5 y la flecha indica la dirección de probabilidad creciente. También proporciona información adicional sobre la bondad de ajuste de cada variable. Para producir una solución integral, se aplica el diagrama de Voronoi.

El biplot logístico ponderado espacio temporal cumple con las siguientes reglas para su interpretación:

- La distancia entre los puntos (sitios geográficos) en el gráfico, están inversamente relacionadas con las similitudes de sus perfiles, es decir, los sitios cercanos tienen características similares.
- El ángulo entre vectores y el eje factorial, indica el grado de relación entre la variable y la dimensión latente.
- El ángulo entre vectores, indican el grado de asociación entre ellos, los ángulos agudos (pequeños) indican que los vectores están estrechamente relacionados.
- Las proyecciones de los puntos (sitios geográficos) sobre el vector (variable), estiman la probabilidad esperada de la característica para ese sitio geográfico.
- La longitud del vector, indica el poder discriminante de la variable en el ordenamiento de los sitios geográficos, mientras más pequeños mayor poder discriminante.

Para realizar los cálculos y obtener las representaciones gráficas, se utilizó un programa de computadora basado en el código Matlab llamado MultBiplot. (Vicente-Villardón, 2021) (Vicente-Villardón, 2013)

Etapa 4: Aplicación del método TSWLB

El cáncer de mama es la principal causa de mortalidad en las mujeres a nivel mundial, la mayoría de las muertes se registran en países de ingresos bajos y medianos. (Beltrán and Martínez, 2021) Según datos del Globocan 2020 en Ecuador la tasa de mortalidad por cáncer de mama (BCMR) es de 10.9 por 100.000 habitantes, representando la cuarta causa de mortalidad en el país. (Globocan - IARC, 2020, 2023) (Tanca-Campozano et al., 2019) (Jaramillo-Feijoo et al., 2020a)

Fuente de datos

La base de datos secundaria son las defunciones generales que se descargó del sitio web del INEC. (Instituto Nacional de Estadísticas y Censos: Defunciones generales) El período de estudio es de 15 años desde el 2007 al 2021. La base de datos de defunciones para el estudio incluye las mujeres fallecidas por cáncer de mama registradas por cada provincia del Ecuador.

Extracción de datos

Para aplicar el estudio espacial con enfoque geográficamente ponderado (Brunsdon et al., 2002) y temporal, se seleccionó el nivel provincial, para lo cual se contabilizaron las muertes de mujeres fallecidas por cáncer de mama por provincia de residencia. Se descartaron el registro de defunciones tardías y de no residentes en Ecuador.

Tasa de mortalidad por cáncer de mama

La fórmula aplicada es la siguiente:

$$BCMR = 100000\ x\ \frac{Defunciones\ por\ cáncer\ de\ mama\ en\ mujeres}{Población\ de\ mujeres}$$

La matriz de datos geoespaciales está compuesta por las BCMR medidas en las 24 provincias del Ecuador. Las coordenadas geográficas fueron obtenidas con el sistema UTM WGS84 que para Ecuador corresponde el 17S.

Resultados

Desde el año 2007, (Fig.2) se observa el incremento sostenido de las tasas de mortalidad por cáncer de mama en el Ecuador, pasando de 4,9 a 8,5 por 100.000 habitantes en el año 2021.

La heterogeneidad espacial local fue analizada con la técnica GWPCA. El porcentaje de variabilidad total de los primeros 3 componentes que analizan las tasas de mortalidad por cáncer de mama en las provincias del Ecuador, se muestra en la Fig.3, donde se observa que las provincias de mayor variabilidad se encuentran al norte y sur del país, siendo al norte: Esmeraldas, Carchi e Imbabura; y al sur: Loja, El Oro y Zamora Chinchipe; dicha variabilidad podría ser que sea debida a que son zonas fronterizas y dinámicas en su población.

En la Fig. 4 se muestra el año con mayor carga de los primeros 3 componentes, donde se observa una variación geográfica en la influencia de la mortalidad por cáncer de mama del periodo 2007 al 2021 en las provincias del Ecuador, se tiene que al norte del país los años que han influenciado en la mortalidad de cáncer de mama fueron los años 2007 y 2008; en la parte sur fue el año 2018, pero en el centro fue el año 2017 y en el centro - este fue el año 2009 a diferencia del centro – oeste que fue el año 2019.

Es importante evidenciar los cambios que se han dado en este periodo de estudio donde la presentación de muertes por cáncer de mama pudiera deberse a distintos factores, como, el acceso a la atención oncológica, movilidad, desarrollo de nuevos servicios y capacidad resolutiva especializada, e incluso el registro de las estadísticas vitales, entre otros.

En la Fig. 5 muestra la tendencia temporal analizada con la prueba estadística no paramétrica Mann-Kendall en las 24 provincias del Ecuador, las tendencias muestran que las tasas no son constantes espacialmente. A nivel regional se observa en las provincias de la región costa y sierra un mayor predominio de tendencias crecientes en las tasas de mortalidad por cáncer de mama.

El biplot logístico integra los resultados de las técnicas GWPCA y prueba Mann Kendall, esto se muestra en la Fig. 6 donde se observa cuatro clústeres, que están agrupados en función de la evolución en el tiempo de la mortalidad por cáncer de mama y el año que ha influido mayormente en su mortalidad.

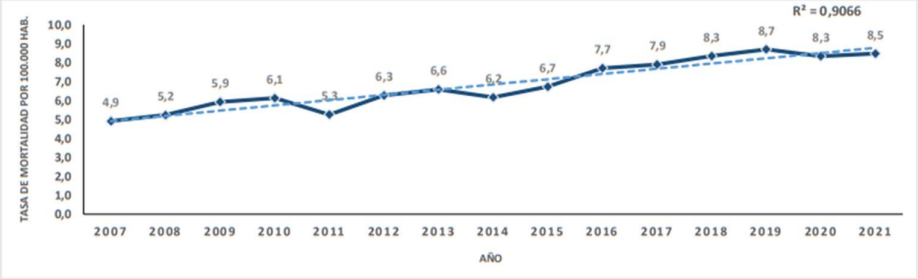


Fig. 2 Evolución en el tiempo de las tasas de mortalidad por cáncer de mama en Ecuador (2007-2021)

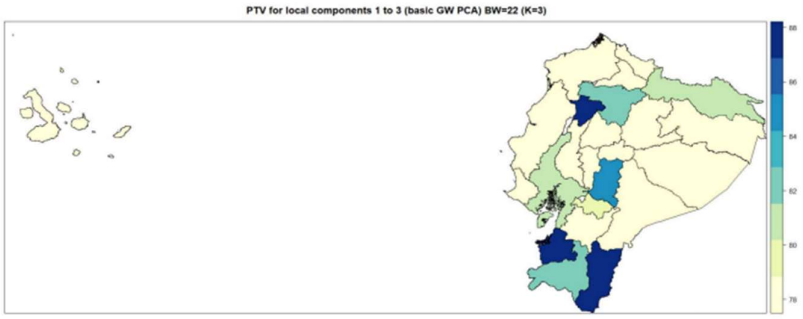


Fig. 3 Mapa representando el porcentaje de variabilidad de las componentes locales.

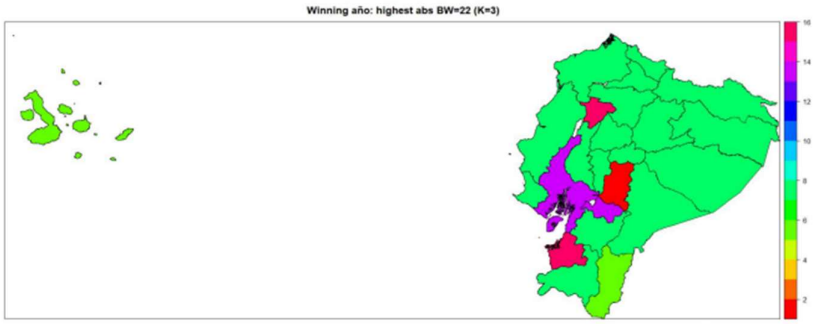


Fig. 4 Mapa representando el año ganador con la mayor carga

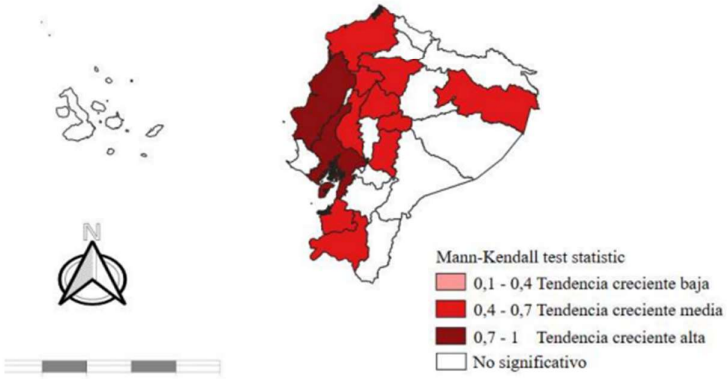


Fig. 5 Mapa de la tendencia temporal Mann-Kendall (Tau) desde 2007 al 2021

15

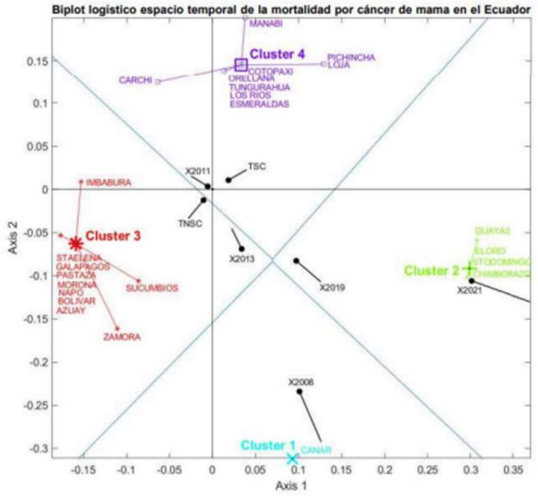


Fig. 6 Biplot logístico con los clústeres de priorización

Discusión

La tasa de mortalidad por cáncer de mama en el Ecuador tiene un incremento sostenido en los últimos 15 años, esto se asemeja a estudios regionales que indican que el cáncer de mama es un problema de salud en América Latina y el Caribe donde anualmente mueren unas 300.000 mujeres por esta enfermedad. (Tanca-Campozano et al., 2019) (Robles and Galanis, 2002)

Existe una variabilidad espacial a nivel provincial de las tasas de mortalidad por cáncer de mama, siendo las provincias ubicadas al sur y al norte del país con mayor variabilidad en la mortalidad de esta enfermedad, así se tiene al norte: Esmeraldas, Carchi e Imbabura; y al sur: Loja, El Oro y Zamora Chinchipe; dicha variabilidad podría ser que sea debida a que son zonas fronterizas y dinámicas en su población, estos resultados son similares a estudios donde se evidencia diferencias en la mortalidad a nivel urbano y rural. (Jaramillo-Feijoo et al., 2020b)

Durante el periodo de estudio se pudo comprobar una variación geográfica en los años que presentaron mayores cargas dentro de los primeros 3 componentes principales locales, es decir, años que influyeron en la mortalidad por cáncer de mama a nivel provincial, así se tiene, las provincias al norte del país fueron influenciados por los años 2007 y 2008; en las provincias al sur fue el año 2018, mientras que en el centro del país fue el año 2017 y en el centro - este fue el año 2009 a diferencia del centro – oeste que fue el año 2019. Dichos resultados son parecidos a estudios previos, donde se observan diferencias de la mortalidad de cáncer a nivel geográfico. (Ramos-Herrera et al., 2020)

Se identificaron cuatro clústeres mediante el biplot logístico el cual integró los resultados de las técnicas GWPCA y Mann-Kendall, el clúster 2 y 4 tienen un comportamiento creciente estadísticamente significativo de las muertes por cáncer de mama durante los 15 años de estudio, en el clúster 2 los años más recientes está influenciado su mortalidad, mientras que los clústeres 1 y 3 tienen un comportamiento creciente estadísticamente no significativo, los clústeres 3 y 4 están más influenciados su mortalidad por los años intermedios 2011 y 2013, el clúster 1 está influenciado su mortalidad por el año 2008.

Considerando que, el clúster 2 tiene una tendencia con comportamiento creciente significativa estadísticamente e influenciada su mortalidad por los años recientes, se puede identificar como grupo prioritario, dicho grupo lo conforman 4 provincias que representan el 17% del total país, siendo dos provincias de la región costa: Guayas y El Oro; y dos provincias de la región sierra: Santo Domingo y Chimborazo; esta representación pudo ser obtenida mediante el biplot logístico el cual integra información externa, dicha técnica es utilizada en varios estudios y aplicadas en distintas áreas. (Demey et al., 2008b) (Galindo et al., 2011) (Gallego-Álvarez and Vicente-Villardón, 2012) (de Noronha Vaz et al., 2015)

Conclusiones

La técnica multivariante TSWLB capturó y representó simultáneamente los componentes espacio temporal de la mortalidad por cáncer de mama en los últimos 15 años, comprobándose diferencias espacio-temporales a nivel provincial de las tasas de mortalidad e Identificándose cuatro clústeres con características propias que permitieron priorizar las provincias de Guayas, El Oro, Chimborazo y Santo Domingo de los Tsáchilas, que conforman el clúster 2 resultando el de mayor importancia para realizar intervenciones urgentes en salud por esta enfermedad.

Bibliografía

Brunsdon, C., Corcoran, J., Higgs, G., 2007. Visualising space and time in crime patterns: A comparison of methods. Computers, environment and urban systems 31, 52–75.

Brunsdon, C., Fotheringham, A.S., Charlton, M., 2002. Geographically weighted summary statistics—a framework for localised exploratory data analysis. Computers, Environment and Urban Systems 26, 501–524.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 28, 281–298.

Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., 2020. The GWR route map: a guide to the informed application of Geographically Weighted Regression. arXiv preprint arXiv:2004.06070.

de Noronha Vaz, T., Galindo, P.V., de Noronha Vaz, E., Nijkamp, P., 2015. Innovative firms behind the regions: Analysis of regional innovation performance in Portugal by external logistic biplots. European Urban and Regional Studies 22, 329–344.

Demey, Vicente-Villardón, J.L., Galindo-Villardón, M.P., Zambrano, A.Y., 2008a. Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. Bioinformatics 24, 2832–2838. https://doi.org/10.1093/bioinformatics/btn552

Demey, Vicente-Villardón, J.L., Galindo-Villardón, M.P., Zambrano, A.Y., 2008b. Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. Bioinformatics 24, 2832–2838. https://doi.org/10.1093/bioinformatics/btn552

Fotheringham, A.S., Crespo, R., Yao, J., 2015. Exploring, modelling and predicting spatiotemporal variations in house prices. The Annals of Regional Science 54, 417–436.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers 107, 1247–1265.

Galindo, P.V., De Noronha Vaz, M.T., de Noronha Vaz, E., 2011. Analysis of Regional Innovation Performance in Portugal-Results from an External Logistic Biplot Method.

Gallego-Álvarez, I., Vicente-Villardón, J.L., 2012. Analysis of environmental indicators in international companies by applying the logistic biplot. Ecological Indicators 23, 250–261.

Globocan - IARC, 2020, 2023. Cancer today [WWW Document]. URL http://gco.iarc.fr/today/home (accessed 5.12.23).

Harris, P., Brunsdon, C., Charlton, M., 2011. Geographically weighted principal components analysis. International Journal of Geographical Information Science 25, 1717–1736.

Hernández, N.E.B., Fleites, O.A., Rodríguez, M.A., 2012. Análisis espacio-temporal de eventos asociados al cáncer: una herramienta para apoyar estudios epidemiológicos. Medisur 10, 93–103.

Jaramillo-Feijoo, L.E., Galindo-Villardon, M.P., Real-Cotto, J.J., González-Rugel, J.L., Idrovo-Madezco, S.E., 2020a. Clúster espacial de mortalidad por cáncer de mama en Ecuador. J. health med. sci. (Print) 29–36.

Jaramillo-Feijoo, L.E., Galindo-Villardon, M.P., Real-Cotto, J.J., González-Rugel, J.L., Idrovo-Madezco, S.E., 2020b. Clúster espacial de mortalidad por cáncer de mama en Ecuador. J. health med. sci. (Print) 29–36.

Nakaya, T., Fotheringham, S., Charlton, M., Brunsdon, C., 2009. Semiparametric geographically weighted generalised linear modelling in GWR 4.0.

Pou, S.A., Niclis, C., Tumas, N., Díaz, M.P., 2019. Disparidades en los patrones espacio-temporales de mortalidad por cáncer de mama y cérvix en Argentina, 1996-2015. Revista de la Facultad de Ciencias Médicas de Córdoba.

Puranik, A., Shreenidhi, S.M., Rai, S.N., 2020. Spatial evaluation of prevalence, pattern and predictors of cervical cancer screening in India. Public health 178, 124–136.

Ramos-Herrera, I.M., Reyna-Sevilla, A., González Castañeda, M.E., Robles-Pastrana, J.D., Herrera-Echauri, D.D., González-Rivera, C.A., Ramos-Herrera, I.M., Reyna-Sevilla, A., González Castañeda, M.E., Robles-Pastrana, J.D., Herrera-Echauri, D.D., González-Rivera, C.A., 2020. Cáncer de mama en Jalisco. Análisis espacial de la mortalidad en 2010-2017. Gaceta médica de México 156, 542–548. https://doi.org/10.24875/gmm.20005546

Robles, S.C., Galanis, E., 2002. El cáncer de mama en América Latina y el Caribe. Revista panamericana de salud pública 12, 141–143.

Santamaría Ulloa, C., 2003. El análisis espacial como herramienta para evaluar alarmas por cáncer. Población y Salud en Mesoamérica. Revista Electrónica, Vol 1 (1), artículo 1.

Souris, M., 2019. Epidemiology and geography: principles, methods and tools of spatial analysis. John Wiley & Sons.

Tanca-Campozano, J., Puga-Peña, G., Quinto-Briones, R., Real-Cotto, J., Jaramillo-Feijoo, L., 2019. Mortalidad y años de vida potencialmente perdidos en cáncer de mama y cérvix en Guayaquil. INSPILIP. Revista Ecuatoriana de Ciencia, Tecnología e Innovacion en Salud Pública 3.

Tejedor Flores, N.D., 2018. Desarrollo sostenible y nexo agua-energía-alimentos: una perpsectiva multivariante.

Vicente-Villardón, J.L., 2021. MultBiplotR: MULTivariate Analysis Using Biplots 2021. R Package Version 1, 30.

Vicente-Villardón, J.L., 2013. MultBiplotR: MULTivariate analysis using biplots. Departamento de Estadística. Universidad de Salamanca.

Vicente-Villardón, J.L., Galindo-Villardón, M.P., Blázquez-Zaballos, A., 2006. Logistic biplots. Multiple Correspondence Analysis and related methods 503–521.

Zymaroieva, A., Zhukov, O., Fedonyuk, T., Pinkin, A., 2019. Application of geographically weighted principal components analysis based on soybean yield spatial variation for agro-ecological zoning of the territory. https://doi.org/10.15159/ar.19.208

Conflicto de interés: los autores declaran no tener conflicto de interés y el contenido del manuscrito no ha sido publicado previamente.

Consentimiento informado: La información para estudio utilizó bases libres del INEC.

Fuente de financiamiento: propio de los autores.

BIBLIOGRAFÍA

Álvarez, F. J. D., & Villardon, P. G. (2015). A proposal for spatio-temporal analysis of traffic matrices using HJ-biplot. *2015 IEEE International Workshop on Measurements & Networking (M&N)*, 1-6.

Amaro, I. R., Vicente-Villardón, J. L., & Galindo-Villardón, M. P. (2004). Manova Biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, *29*(1), 26-32.

Amor-Esteban, V., Galindo-Villardón, M.-P., & García-Sánchez, I.-M. (2018). Industry mimetic isomorphism and sustainable development based on the X-STATIS and HJ-biplot methods. *Environmental Science and Pollution Research*, *25*, 26192-26208.

Amor-Esteban, V., García-Sánchez, I.-M., & Galindo-Villardón, M.-P. (2018). Analysing the effect of legal system on corporate social responsibility (CSR) at the country level, from a multivariate perspective. *Social Indicators Research*, *140*, 435-452.

Andrade-Sánchez, A. I., Galindo-Villardón, M. P., & Cuevas Romo, J. (2015). Análisis multivariante del perfil psicológico de los deportistas universitarios: Aplicación del CPRD en México. *Educación Física y Ciencia*, *17*(2), 00-00.

Beltrán, J. A. O., & Martínez, O. M. V. (2021). Caracterización clínica epidemiológica del cáncer de mama en mujeres mayores de 20 años en El Salvador. *Alerta, Revista científica del Instituto Nacional de Salud*, *4*(3), Article 3. https://doi.org/10.5377/alerta.v4i3.10952

Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, environment and urban systems*, *31*(1), 52-75.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (2002). Geographically weighted summary statistics—A framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, *26*(6), 501-524.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical analysis*, *28*(4), 281-298.

Caballero-Juliá, D., Villardón, M. P. G., & García, M.-C. (2017). JK-Meta-Biplot y STATIS Dual como herramientas de análisis de tablas textuales múltiples. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *25*, 18-33.

Cabrera, J. G., Martínez, M. F., Mateos, E. M., & Tavera, S. V. (2006). Study of the evolution of air pollution in Salamanca (Spain) along a five-year period (1994–1998) using HJ-Biplot simultaneous representation analysis. *Environmental Modelling & Software*, *21*(1), 61-68.

Cañizares, J. F. R., Abarca, E. F. G., Naranjo, D. N. C., Vicente-Villardón, J. L., & Demey, J. (2016). Caracterización de germoplasma de maíz local a través de marcadores SSR asistido por biplot logístico externo (BLE). *Proceedings of the XXVI Simposio Internacional de Estadística*, 4.

Carrasco, G., Molina, J.-L., Patino-Alonso, M.-C., Castillo, M. D. C., Vicente-Galindo, M.-P., & Galindo-Villardón, M.-P. (2019). Water quality evaluation through a multivariate statistical HJ-Biplot approach. *Journal of Hydrology*, *577*, 123993.

Charlton, M., Brunsdon, C., Demšar, U., Harris, P., & Fotheringham, S. (2010). *Principal components analysis: From global to local*.

Cisneros, J. T. C., Babici, V. R., Guerrero, C. A. R., & Villardón, J. L. V. (2020). Análisis multivariado HJ-Biplot de la ocurrencia de Helicobacter pylori como riesgo para cáncer gástrico, en la ciudadela el Cristo de Consuelo, Milagro Ecuador. *Boletín de Malariología y Salud Ambiental*, *60*(2).

Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., & Wang, Y. (2020). The GWR route map: A guide to the informed application of Geographically Weighted Regression. *arXiv preprint arXiv:2004.06070*.

Correa Londoño, G., Lavalett Oñate, L. L., Galindo Villardón, M. P., & Afanador Kafuri, L. (2007). Uso de métodos multivariantes para la agrupación de aislamientos de Colletotrichum spp. Con base en características morfológicas y culturales. *Revista Facultad Nacional de Agronomia Medellin*, *60*(1), 3671-3690.

Cubilla-Montilla, M., Nieto-Librero, A. B., Galindo-Villardón, M. P., & Torres-Cubilla, C. A. (2021). Sparse HJ biplot: A new methodology via elastic net. *Mathematics*, *9*(11), 1298.

de Noronha Vaz, T., Galindo, P. V., de Noronha Vaz, E., & Nijkamp, P. (2015). Innovative firms behind the regions: Analysis of regional innovation performance in Portugal by external logistic biplots. *European Urban and Regional Studies*, *22*(3), 329-344.

Delchambre, L. (2015). Weighted principal component analysis: A weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, *446*(4), 3545-3555.

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008a). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, *24*(24), 2832-2838.

Demey, Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008b). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, *24*(24), 2832-2838. https://doi.org/10.1093/bioinformatics/btn552

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: An overview. *Annals of the Association of American Geographers*, *103*(1), 106-128.

Díaz-Faes, A. A., González-Albo, B., Galindo, M. P., & Bordons, M. (2013). *HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos*.

Drápela, K., & Drápelová, I. (2011). Application of Mann-Kendall test and the Sen's slope estimates for trend detection in deposition data from Bílỳ

Kříž (Beskydy Mts., the Czech Republic) 1997-2010. *Beskydy*, *4*(2), 133-146.

Fathian, F., Dehghan, Z., Bazrkar, M. H., & Eslamian, S. (2016). Trends in hydrological and climatic variables affected by four variations of the Mann-Kendall approach in Urmia Lake basin, Iran. *Hydrological Sciences Journal*, *61*(5), 892-904. https://doi.org/10.1080/02626667.2014.932911

Fernández Gómez, M. J., Galindo Villardón, M. P., Barrera Mellado, I., Vicente Villardón, J. L., & Martín Casado, A. M. (1996). Alternativa al análisis canónico de correspondencias basada en los métodos Biblot. *Mediterránea. Serie de Estudios Biológicos, Época II, n. 15 (1996); pp. 63-71*.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography: Perspectives on spatial data analysis*. Sage.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: The analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, A. S., Crespo, R., & Yao, J. (2015a). Exploring, modelling and predicting spatiotemporal variations in house prices. *The Annals of Regional Science*, *54*, 417-436.

Fotheringham, A. S., Crespo, R., & Yao, J. (2015b). Geographical and Temporal Weighted Regression (GTWR). *Geographical Analysis*, *47*(4), 431-452. https://doi.org/10.1111/gean.12071

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, *107*(6), 1247-1265.

Frutos Bernal, E., Martín del Rey, A., & Galindo Villardón, P. (2020). Analysis of madrid metro network: From structural to HJ-biplot perspective. *Applied Sciences*, *10*(16), 5689.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

Galindo, P. V., De Noronha Vaz, M. T., & de Noronha Vaz, E. (2011). *Analysis of Regional Innovation Performance in Portugal-Results from an External Logistic Biplot Method*.

Galindo Villardón, M. P. G. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestiió: quaderns d'estadística i investigació operativa*, 13-23.

Gallego-Álvarez, I., Galindo-Villardón, M. P., & Rodríguez-Rosa, M. (2015). Analysis of the sustainable society index worldwide: A study from the biplot perspective. *Social Indicators Research*, *120*, 29-65.

Gallego-Álvarez, I., Vicente-Galindo, M. P., Galindo-Villardón, M. P., & Rodríguez-Rosa, M. (2014). Environmental performance in countries worldwide: Determinant factors and multivariate analysis. *Sustainability*, *6*(11), 7807-7832.

Gallego-Álvarez, I., & Vicente-Villardón, J. L. (2012). Analysis of environmental indicators in international companies by applying the logistic biplot. *Ecological Indicators*, *23*, 250-261.

García, A. B. S., & Villardón, P. G. (2018). Uso e integración de las TIC en el aula y dificultades del profesorado en activo de cara a su integración. *Profesorado, Revista de Currículum y Formación del Profesorado*, *22*(3), 341-358.

García-Garizábal, I. (2017). Rainfall variability and trend analysis in coastal arid Ecuador. *International Journal of Climatology*, *37*(13), 4620-4630.

Globocan - IARC, 2020. (s. f.). *Cancer today*. Recuperado 12 de mayo de 2023, de http://gco.iarc.fr/today/home

González, S. H., & Villardón, M. P. G. (2013). BIPROB: UN MÉTODO PARA OBTENER UN BIPLOT ROBUSTO. *Investigación Operacional*, *27*(3), 287-299.

González-García, N., Nieto-Librero, A. B., & Galindo-Villardón, P. (2021). C enet Biplot: A new proposal of sparse and orthogonal biplots methods by means of elastic net CSVD. *Advances in Data Analysis and Classification*, 1-15.

González-García, N., Sánchez-García, A. B., Nieto-Librero, A. B., & Galindo-Villardón, M. P. (2019). Actitud y enfoques de aprendizaje en el estudio de la Didáctica General. Una visión multivariante. *Revista de Psicodidáctica*, *24*(2), 154-162.

Grassi, J. E. A., Mantilla, H. A. T., Hernández, J. Y. L., & Polanco, M. R. (2019). Análisis espacio temporal de la homogeneidad de estaciones de precipitación en una zona árida y semiárida del Centro Occidente de Venezuela. *Ciencia e Ingeniería*, *40*(2), 185-194.

Haining, R. P. (2003). *Spatial data analysis: Theory and practice*. Cambridge university press.

Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1), 182-196. https://doi.org/10.1016/S0022-1694(97)00125-X

Han, J., Kang, X., Yang, Y., & Zhang, Y. (2022). Geographically and temporally weighted principal component analysis: A new approach for exploring air pollution non-stationarity in China, 2015–2019. *Journal of Spatial Science*, 1-18.

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717-1736.

Hernández Suárez, M., Molina Pérez, D., Rodríguez-Rodríguez, E. M., Díaz Romero, C., Espinosa Borreguero, F., & Galindo-Villardón, P. (2016). The compositional HJ-biplot—A new approach to identifying the links among bioactive compounds of tomatoes. *International Journal of Molecular Sciences*, *17*(11), 1828.

Hernández-Sánchez, J. C., & Vicente-Villardón, J. L. (2017). Logistic biplot for nominal data. *Advances in Data Analysis and Classification*, *11*, 307-326.

Herrera, H. L., Pedrosa, I., Galindo, M. P. V., Suárez-Álvarez, J., Villardón, M. P. G., & García-Cueto, E. (2014). Multivariate analysis of burnout syndrome in Latin-American priests. *Psicothema*, 227-234.

Hipel, K. W., McLeod, A. J., & Weller, R. R. (1988). Data Analysis of Water Quality Time Series in Lake Erie1. *JAWRA Journal of the American Water Resources Association*, *24*(3), 533-544. https://doi.org/10.1111/j.1752-1688.1988.tb00903.x

IARC, Cáncer. (s. f.). *Temas de cáncer – IARC*. Recuperado 31 de mayo de 2023, de https://www.iarc.who.int/cancer-topics/

Isensee, L. J., Detzel, D. H. M., Pinheiro, A., & Piazza, G. A. (2023). Extreme streamflow time series analysis: Trends, record length, and persistence. *Journal of Applied Water Engineering and Research*, *11*(1), 40-53. https://doi.org/10.1080/23249676.2022.2030254

Jaramillo-Feijoo, L. E., Galindo-Villardon, M. P., & Real-Cotto, J. J. (2020). Análisis clúster para big data: Una aplicación con variables demográficas en provincias del Ecuador. *J. health med. sci.(Print)*, 45-50.

Jaramillo-Feijoo, L. E., Galindo-Villardon, M. P., Real-Cotto, J. J., González-Rugel, J. L., & Idrovo-Madezco, S. E. (2020). Clúster espacial de mortalidad por cáncer de mama en Ecuador. *J. health med. sci. (Print)*, 29-36.

Karmeshu, N. (2012). *Trend detection in annual temperature & precipitation using the Mann Kendall test–a case study to assess climate change on select states in the northeastern United States*.

Kendall, M. G. (1975). Rank Correlation Methods, Charles Griffin, London (1975). *Google Sch*.

Lalangui, K., Rivadeneira Maya, K., Sánchez-Carrillo, C., Sosa Cortéz, G., & Quentin, E. (2022). The spatio-temporal dynamics of infant mortality in Ecuador from 2010 to 2019. *BMC Public Health*, *22*(1), 1841. https://doi.org/10.1186/s12889-022-14242-1

Li, Z., Cheng, J., & Wu, Q. (2016). Analyzing regional economic development patterns in a fast developing province of China through geographically weighted principal component analysis. *Letters in Spatial and Resource Sciences*, *9*, 233-245.

Libório, M. P., Martinuci, O. da S., Machado, A. M. C., Ekel, P. I., Abreu, J. F. de, & Laudares, S. (2022). Representing Multidimensional Phenomena of Geographic Interest: Benefit of the Doubt or Principal Component Analysis? *The Professional Geographer*, *74*(4), 758-771.

Lloyd, C. D. (2010). Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. *Computers, Environment and Urban Systems*, *34*(5), 389-399.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, 245-259.

Martínez-Hidalgo, P., Galindo-Villardón, P., Trujillo, M. E., Igual, J. M., & Martínez-Molina, E. (2014). Micromonospora from nitrogen fixing nodules of alfalfa (Medicago sativa L.). A new promising Plant Probiotic Bacteria. *Scientific Reports*, *4*(1), 6389.

Martínez-Regalado, J. A., Murillo-Avalos, C. L., Vicente-Galindo, P., Jiménez-Hernández, M., & Vicente-Villardón, J. L. (2021). Using HJ-Biplot and external logistic biplot as machine learning methods for corporate social responsibility practices for sustainable development. *Mathematics*, *9*(20), 2572.

Martín-Rodríguez, J., Galindo-Villardón, M. P., & Vicente-Villardón, J. L. (2002). Comparison and integration of subspaces from a biplot perspective. *Journal of Statistical Planning and Inference*, *102*(2), 411-423.

Medina-Hernández, E. J., Guzmán-Aguilar, D. S., Muñiz-Olite, J. L., & Siado-Castañeda, L. R. (2023). The current status of the sustainable development goals in the world. *Development Studies Research*, *10*(1), 2163677.

Mendes, S., Fernández-Gómez, M. J., Galindo-Villardón, M. P., Morgado, F., Maranhão, P., Azeiteiro, U. M., & Bacelar-Nicolau, P. (2009). The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-biplot method. *ARQUIPÉLAGO-Life and Marine Sciences*, 25-35.

Miranda, A. R., Scotta, A. V., Cortez, M. V., González-García, N., Galindo-Villardón, M. P., & Soria, E. A. (2022). Association of Dietary Intake of Polyphenols with an Adequate Nutritional Profile in Postpartum Women from Argentina. *Preventive Nutrition and Food Science*, *27*(1), 20.

Mondal, A., Kundu, S., & Mukhopadhyay, A. (2012). Rainfall trend analysis by Mann-Kendall test: A case study of north-eastern part of Cuttack district, Orissa. *International Journal of Geology, Earth and Environmental Sciences*, *2*(1), 70-78.

Montes Escobar, K. (2022). *Métodos multivariantes para evaluar tumores neuroendocrinos*.

Montilla, M. C., Rodríguez, C., & Ortega, E. (2015). Análisis de indicadores de sostenibilidad del Global Reporting Initiative. Una mirada desde el biplt logístico. *Revista Científica Centros*, *4*(Especial), 96-114.

Murillo, C. L. (2015). Estudio de la sostenibilidad de las empresas mexicanas utilizando el Biplot Logístico Externo. *Universidad de Salamanca*.

Nakaya, T., Fotheringham, S., Charlton, M., & Brunsdon, C. (2009). *Semiparametric geographically weighted generalised linear modelling in GWR 4.0*.

Neeti, N., & Eastman, J. R. (2011). A Contextual Mann-Kendall Approach for the Assessment of Trend Significance in Image Time Series. *Transactions in GIS*, *15*(5), 599-611. https://doi.org/10.1111/j.1467-9671.2011.01280.x

Nieto-Librero, A. B., Sierra, C., Vicente-Galindo, M. P., Ruíz-Barzola, O., & Galindo-Villardón, M. P. (2017). Clustering Disjoint HJ-Biplot: A new tool for identifying pollution patterns in geochemical studies. *Chemosphere*, *176*, 389-396.

Núñez González, S., Calle Celi, D., Pilco, J., & Simancas Racines, D. (2018). Cambios en la tendencia temporal de mortalidad por cáncer de mama en Ecuador 2001-2016. *Revista Ecuatoriana de Medicina y Ciencias Biológicas: REMCB*, *39*(2), 159-167.

Núñez-González, S., Aulestia-Ortiz, S., Borja-Villacrés, E., & Simancas-Racine, D. (2018). Mortalidad por enfermedades isquémicas del corazón en Ecuador, 2001-2016: Estudio de tendencias. *Revista médica de Chile*, *146*(8), 850-856. https://doi.org/10.4067/s0034-98872018000800850

Núñez-González, S., Delgado-Ron, A., & Simancas-Racines, D. (2020). Tendencias y análisis espacio-temporal de la mortalidad por diabetes mellitus en Ecuador, 2001-2016. *Rev. cuba. salud pública*, e1314-e1314.

Núñez-González, S., Delgado-Ron, J. A., Gault, C., & Simancas-Racines, D. (2018). Trends and Spatial Patterns of Oral Cancer Mortality in Ecuador, 2001–2016. *International Journal of Dentistry*, *2018*, e6086595. https://doi.org/10.1155/2018/6086595

Núñez-González, S., Duplat, A., & Simancas, D. (s. f.). *Mortalidad por enfermedades cerebrovasculares en Ecuador 2001- 2015: Estudio de tendencias, aplicación del modelo de regresión joinpoint.*

Núñez-González, S., Gault, C., & Simancas-Racines, D. (2019). Spatial analysis of dengue, cysticercosis and Chagas disease mortality in Ecuador, 2011–2016. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, *113*(1), 44-47.

OMS, Cáncer de mama. (s. f.). *Cáncer de mama*. Recuperado 27 de mayo de 2023, de https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

Ozocak, M., Akay, A. O., Esin, A. İ., Yurtseven, H., & Akgul, M. (2023). *A New Framework to Spatial and Temporal Drought Analysis for 1990–2020 Period with Mann–Kendall and Innovative Trend Analysis Methods in Turkey*.

Pedraz, C., & Galindo, P. (1986). Study of socio-cultural factors influencing the decision to breast-feed instead of bottle-feed. *Arch. Pediat*, *36*, 469-477.

Perera, S., Allali, M., Linstead, E., & El-Askary, H. (2022). Deriving Drought Vulnerability Index using Geographically Weighted Principal

Component Analysis (GWPCA) and K-Means Clustering for Nile Basin. *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 3187-3190.

Pérez-Mellado, V., & Galindo, M. P. (1986). Sistemática de Podarcis (Sauria, Lacertidae) ibéricas y norteafricanas mediante técnicas multidimensionales. *Salamanca: Serie Manuales Universitarios, Ediciones Universidad de Salamanca*.

Pilacuan-Bonete, L., Galindo-Villardón, P., & Delgado-Álvarez, F. (2022). HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics*, *10*(14), 2529.

Pou, S. A., Niclis, C., Tumas, N., & Díaz, M. P. (2019). Disparidades en los patrones espacio-temporales de mortalidad por cáncer de mama y cérvix en Argentina, 1996-2015. *Revista de la Facultad de Ciencias Médicas de Córdoba*. https://revistas.unc.edu.ar/index.php/med/article/view/25836

Que, X., Ma, X., Ma, C., & Chen, Q. (2020). A spatiotemporal weighted regression model (STWR v1. 0) for analyzing local nonstationarity in space and time. *Geoscientific Model Development*, *13*(12), 6149-6164.

Ramos-Herrera, I. M., Reyna-Sevilla, A., González Castañeda, M. E., Robles-Pastrana, J. D., Herrera-Echauri, D. D., González-Rivera, C. A., Ramos-Herrera, I. M., Reyna-Sevilla, A., González Castañeda, M. E., Robles-Pastrana, J. D., Herrera-Echauri, D. D., & González-Rivera, C. A. (2020). Cáncer de mama en Jalisco. Análisis espacial de la mortalidad en 2010-2017. *Gaceta médica de México*, *156*(6), 542-548. https://doi.org/10.24875/gmm.20005546

Riera-Segura, L., Tapia-Riera, G., Amaro, I. R., Infante, S., & Marin-Calispa, H. (2022). HJ-biplot and clustering to analyze the COVID-19 vaccination process of American and European countries. *Smart Technologies, Systems and Applications: Second International*

Conference, SmartTech-IC 2021, Quito, Ecuador, December 1–3, 2021, Revised Selected Papers, 383-397.

Robles, S. C., & Galanis, E. (2002). El cáncer de mama en América Latina y el Caribe. *Revista panamericana de salud pública*, *12*(2), 141-143.

Rodríguez, C. C., Cubilla, M. I., & Ortega-Gómez, E. (2014). Caracterización multivariante de los delitos en Panamá a través del método Hj-Biplot. *Revista Colón Ciencias, Tecnologia y Negocios*, *1*(2), 18-29.

Ruiz-Toledo, M., Ruff-Escobar, C., Benites, L., González, J. A., & Galindo-Villardón, M.-P. (2021). The Place of Latin American Universities in International University Rankings. A Multivariate Statistical Analysis. En *Perspectives and Trends in Education and Technology: Selected Papers from ICITED 2021* (pp. 163-181). Springer.

Sa'adi, Z., Shahid, S., Ismail, T., Chung, E.-S., & Wang, X.-J. (2019). Trends analysis of rainfall and rainfall extremes in Sarawak, Malaysia using modified Mann–Kendall test. *Meteorology and Atmospheric Physics*, *131*(3), 263-277. https://doi.org/10.1007/s00703-017-0564-3

Santos, C., Munoz, S. S., Gutierrez, Y., Hebrero, E., Vicente, J. L., Galindo, P., & Rivas, J. C. (1991). Characterization of young red wines by application of HJ biplot analysis to anthocyanin profiles. *Journal of Agricultural and food chemistry*, *39*(6), 1086-1090.

Santos, V. C. dos, Silva, R. A. e, & Maciel, G. F. (2023). AVALIAÇÃO DE TENDÊNCIA DO INÍCIO, FIM, DURAÇÃO E TOTAL DE PRECIPITAÇÃO DA ESTAÇÃO CHUVOSA DE PALMAS - TO. *DESAFIOS - Revista Interdisciplinar da Universidade Federal do Tocantins*, *2*(1), Article 1. https://doi.org/10.20873/pibic2022_10

SEOM: Cáncer de mama. (s. f.). *Cancer de mama—SEOM: Sociedad Española de Oncología Médica*. Recuperado 30 de mayo de 2023, de https://seom.org/info-sobre-el-cancer/cancer-de-mama

Sodagari, H. R., & Varga, C. (2023). Evaluating Antimicrobial Resistance Trends in Commensal Escherichia coli Isolated from Cecal Samples of Swine at Slaughter in the United States, 2013–2019. *Microorganisms*, *11*(4), 1033.

Souris, M. (2019). *Epidemiology and geography: Principles, methods and tools of spatial analysis*. John Wiley & Sons.

Tanca-Campozano, J., Puga-Peña, G., Quinto-Briones, R., Real-Cotto, J., & Jaramillo-Feijoo, L. (2019). Mortalidad y años de vida potencialmente perdidos en cáncer de mama y cérvix en Guayaquil. *INSPILIP. Revista Ecuatoriana de Ciencia, Tecnología e Innovacion en Salud Pública*, *3*(1).

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

Tsutsumida, N., Murakami, D., Yoshida, T., Nakaya, T., Lu, B., Harris, P., & Comber, A. (2022). A Comparison of Geographically Weighted Principal Components Analysis Methodologies (Short Paper). *15th International Conference on Spatial Information Theory (COSIT 2022)*.

Urresti Estala, B., Carrasco Cantos, F., Fernández Ruíz, L., & Jiménez Gavilán, P. (2012). *Evaluación de tendencias de contaminantes en la masa de agua Bajo Guadalhorce (sur de España): Aplicación del test estadístico de Mann-Kendall*. https://doi.org/10.13039/501100008737

Vairinhos, V. M., Pereira, L. A., Matos, F., Nunes, H., Patino, C., & Galindo-Villardón, P. (2022). Framework for Classroom Student Grading with Open-Ended Questions: A Text-Mining Approach. *Mathematics*, *10*(21), 4152.

Valenzuela-Cobos, J. D., Guevara-Viejó, F., Vicente-Galindo, P., & Galindo-Villardón, P. (2022). Food Sustainability Study in Ecuador: Using PCA Biplot and GGE Biplot. *Sustainability*, *14*(20), 13033.

Valenzuela-Cobos, J. D., Guevara-Viejó, F., Vicente-Galindo, P., & Galindo-Villardón, P. (2023). Eco-Friendly Biocontrol of Moniliasis in Ecuadorian Cocoa Using Biplot Techniques. *Sustainability*, *15*(5), 4223.

Vázquez-Pérez, J. P., Vicente-Galindo, P., & Galindo-Villardón, M. P. (2011). Variables que inciden en la seguridad de las escuelas públicas de los Estados Unidos. *Revista de Educación de Puerto Rico (REduca)*, *44*(1), 141-165.

Vicente Galindo, P., Vaz, E., & De Noronha, T. (2015). How corporations deal with reporting sustainability: Assessment using the multicriteria logistic biplot approach. *Systems*, *3*(1), 6-26.

Vicente, J. L., Galindo, P., Polanco, A. M., Hebrero, E., Rivas-Gonzalo, J. C., Gutiérrez, Y., & Santos-Buelga, C. (1993). Biplot analysis applied to enological parameters in the geographical classification of young red wines. *American journal of enology and viticulture*, *44*(3), 302-308.

Vicente-Villardón, J. L. (2013). MultBiplotR: MULTivariate analysis using biplots. *Departamento de Estadística. Universidad de Salamanca*.

Vicente-Villardón, J. L. (2021). MultBiplotR: MULTivariate Analysis Using Biplots 2021. *R Package Version*, *1*, 30.

Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zaballos, A. (2006a). Logistic biplots. *Multiple Correspondence Analysis and related methods*, 503-521.

Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zaballos, A. (2006b). Logistic biplots. *Multiple Correspondence Analysis and related methods*, 503-521.

Vicente-Villardón, J. L., & Hernández-Sánchez, J. C. (2020). External logistic biplots for mixed types of data. *Advanced Studies in Classification and Data Science*, 169-183.

Vicente-Villardón, J. L., & Sánchez, J. C. H. (2014). Logistic Biplots for Ordinal Data with an Application to Job Satisfaction of Doctorate Degree Holders in Spain. *arXiv preprint arXiv:1405.0294*.

Villardón, J. L. V. (1992). *Una alternativa a las técnicas factoriales clásicas basada en una generalización de los métodos Biplot* [PhD Thesis]. Universidad de Salamanca.

Villerias Salinas, S., Nochebuena, G., & Uriostegui Flores, A. (2020). *Análisis geográfico del COVID-19 Análisis espacial de vulnerabilidad y riesgo en salud por COVID-19 en el estado de Guerrero, México.* http://ri.uagro.mx/handle/uagro/1421

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley & Sons.

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, *259*(1), 254-271. https://doi.org/10.1016/S0022-1694(01)00594-7

Yue, S., & Wang, C. (2004). The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. *Water Resources Management*, *18*(3), 201-218. https://doi.org/10.1023/B:WARM.0000043140.61082.60

Zymaroieva, A., Zhukov, O., Fedonyuk, T., & Pinkin, A. (2019). *Application of geographically weighted principal components analysis based on soybean yield spatial variation for agro-ecological zoning of the territory*. https://doi.org/10.15159/ar.19.208