UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD ECONOMÍA Y PLANIFICACIÓN

"DETECCIÓN DE OUTLIERS ESPACIALES UTILIZANDO EL DIAGRAMA DE DISPERSIÓN DE MORÁN Y EL VARIOGRAMA NUBE"

Presentada por:

MARITZA SARELA PALACIOS MOSQUERA

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima - Perú

2018

Dedicatoria

Dedico esta tesis a mis padres y mis abuelos quienes brindaron la educación, el apoyo y los consejos para seguir adelante en la vida, a mis hermanos que también me dieron su ayuda, a mis compañeros de estudio, a mis maestros, a mi pareja y amigos. Gracias a su ayuda he logrado escribir esta tesis.

Agradecimiento

Quiero agradecer en estas líneas a Dios y a todos aquellas personas que me brindaron su apoyo para la realización de la presente tesis, mi enorme agradecimiento para mi asesor de tesis el profesor Jorge Chue Gallardo por haberme brindado la oportunidad de recurrir a su gran conocimiento cientifico, asi como también por haberme tenido toda la paciencia del mundo para guiarme durante todo el desarrollo de la tesis.

Quiero agradecer a mi abuelo Perico, como le decía de cariño, por inculcarme el deseo de tener una profesión, aunque no este presente.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

RESUMEN

Uno de los problemas del análisis del datos es la presencia de outliers, esto puede afectar las medidas estadísticas que se desean estimar de una población. La presente investigación se enfoca a la detección de los outliers pero en un contexto geográfico; para ello se empleó los datos obtenidos de la encuesta nacional de los egresados universitarios peruanos en el 2014. Como variable de estudio se consideró el ingreso total de los egresados universitarios en las diferentes regiones del pais para observar si existen ingresos muy atípicos respecto a una región a otra, o si dentro de una región existen valores muy altos respecto a su alrededor, estos valores anormales o raros dentro de un contexto geográfico se consideran como outliers espaciales que es muy diferente a los outliers tradicionales, para poder identificar dichos valores atípicos espaciales se empleó dos técnicas gráficas exploratorias para la detección de outliers espaciales: el variograma nube y el diagrama de dispersión de Morán, que tienen la particularidad de ser muy sensibles a la presencia de outliers espaciales. Se utilizó las dos pruebas los datos del ingreso total de la encuesta, se consideró una muestra 250 datos para su óptimo procesamiento, luego se logró detectar los outliers espaciales de la variable de la investigación que fue de ingreso total de S/7’400, y donde el variograma nube fue más sensible a la presencia de outliers que el diagrama de dispersión de Morán.

Palabras claves: outliers espaciales, variograma nube, el diagrama de dipersión Morán, egresados universitarios.

ABSTRACT

One of the problems of data analysis is the presence of outliers, this may affect the statistical measures that are desired to estimate a population. The present investigation focuses on the detection of outliers but in a geographical context; For this purpose, the data obtained from the national survey of Peruvian university graduates was used in 2014. As a study variable, the total income of university graduates in the different regions of the country was considered to observe if there are very atypical income with respect to a region. to another, or if within a region there are very high values around it, these abnormal or rare values within a geographical context are considered as spatial outliers that is very different from traditional outliers, to be able to identify such outliers in space. he used two exploratory graphical techniques for the detection of spatial outliers: the variogram cloud and the Morán scatterplot, which have the peculiarity of being very sensitive to the presence of spatial outliers. The two tests were used the data of the total income of the survey, a sample was considered 250 data for its optimal processing, then it was possible to detect the spatial outliers of the research variable that was of total income of S/ 7'400, and where the variograma cloud was more sensitive to the presence of outliers than the Morán scatterplot.

Keywords: spatial outliers, variograma cloud, Morán scatterplot, university graduates.

I. INTRODUCCIÓN

Cuando se realiza un análisis de datos ocurre con frecuencia que existen valores que tienen un comportamiento distinto respecto a los demás datos, se consideran outliers o valores atípicos. Los métodos clásicos para la detección de outliers es mediante gráficos como el boxplot, o mediante técnicas basadas en aproximaciones con conceptos estadísticos como: densidad, desviación estándar, distancia y de alta dimensión.

Un outlier tradicional se diferencia de un outlier espacial, porque, el primero esta basado en comparaciones globales, lo que no ocurre con los outliers espaciales. Por este motivo, los outliers tradicionales son denominados “outliers globales” mientras que los outliers espaciales son denominados “outliers locales”. Un outlier espacial no necesariamente se desvía del resto de los datos (Chen, Tien Lu, Kou, & Chen, 2007). Los outliers tradicionales corresponden a números, caracteres y categorías, mientras que un outlier espacial está representando a datos más complejos como puntos, líneas, polígonos y objetos en 3D. Esta última diferencia hace que algunos outliers espaciales necesitan ser detectados mediante una medida de la autocorrelación espacial (Kou, Tien Lu, & Chen, 2006).

La presencia de un outlier espacial puede proporcionar una información útil acerca de los valores de la variable a analizar, como hallazgos relevantes sobre su comportamiento (Díaz Muñiz, García Nieto, Alonso Fernández, Martínez Torres & Taboada, 2012). La eliminación o no de un outlier es una decisión que deberá asumir el investigador teniendo en consideración el contexto del estudio (Osborne & Overbay, 2016).

La detección de los outliers espaciales es un problema en diferentes disciplinas como la medicina, ecología, meteorología, tráfico vehicular, entre otras más (Kou, Tien Lu, & Chen, 2006). En la investigación se tomó en cuenta los datos de la encuesta nacional de egresados universitarios del Perú (ENEUP-2014) ejecutada por el Instituto Nacional de Estadística e Informática (INEI) en el año 2014.

Se consideró en la investigacion la variable ingreso total mensual de los egresados (ITM) definido con la suma de los ingresos del trabajo principal, trabajo secundario, bonificaciones, horas extras, comisiones entre otros.

El objetivo de la tesis es utilizar las técnicas exploratoria espaciales como el diagrama de dispersión de Morán y el variograma nube para la detección de outliers espaciales correspondientes a la variable ITM de la ENEUP. El diagrama de dispersión de Morán es una gráfica bivariada que cuenta con dos valores: en el eje "x"con un valor del atributo normalizado y el eje "y" el promedio de la zona de dicho valor de atributo normalizado (Chang-Tien, Yufeng, Hongjun, & Dechang). El variograma nube es un gráfica bivariada que possee dos valores: en el eje "y" la disimilitud del promedio entre dos observaciones en un espacio geografico y en el "x" una distancia euclidiana de dichas observaciones. Existen otras técnicas de detección de outliers espaciales que pueden ser investigaciones futuras. (Haslett, Bradley, Craig, & Wills, 1991)

Objetivo General

Detectar los outliers espaciales mediante el diagrama de dispersión de Morán y el variograma nube en la encuesta nacional de egresados universitarios 2014.

Objetivos Específicos

a. Presentar la metodología del diagrama de dispersión de Morán y del variograma nube.
b. Ilustrar la metodología exploratoria espacial propuesta para el análisis del ingreso de los egresados universitarios en el 2014.

II. REVISIÓN DE LITERATURA

2.1. Datos espaciales

Según Longley (2001), los datos espaciales se definen como datos construidos a partir de elementos atómicos o hechos sobre el mundo geográfico; es decir, que están vinculados a la ubicación geográfica (lugar), un tiempo, y una propiedad descriptiva o atributo de la entidad (Fischer & Wang, 2011). La diferencia entre un dato "espacial" y un dato "normal" es la siguiente: un dato "normal" es el atributo o característica de un objeto, por ejemplo: la leche es fría (en este caso el dato es "fría"); en cambio, un dato espacial es un atributo del objeto que está asociado a una ubicación geográfica, por ejemplo: la leche esta fría en la refrigeradora (el atributo de la leche es que esta fría y su ubicación es que está en la refrigeradora). El atributo de un objeto espacial depende de su ubicación y está influenciado por otro objeto vecino (Koperski, Adhikary, & Han, 2007). Otro detalle que debe tenerse en cuenta de los datos espaciales a diferencia de los datos clásicos o normales, están en identificar dos dimensiones: latitud y longitud, en un espacio geográfico. En el cuadro N°1 se resumen sus características.

Por ejemplo, considérese la siguiente afirmación: "La temperatura de las dos de la tarde del día 24 de diciembre de 2010 en la latitud de 48º 15' norte y de longitud 16º21' 28s este, fue de 6.7ºC". En esta afirmación se describen algunos datos espaciales de la temperatura atmosférica, debido a que estos datos están vinculados a un lugar, un tiempo y un atributo; en este caso la temperatura en grados Celsius es el atributo (Fischer & Wang, 2011). Los atributos de datos espaciales se pueden clasificar según la ubicación (por ejemplo la temperatura atmosférica y los ingresos), mientras que otros lo clasifican según categorías tales como, las clases de uso de la tierra para diferentes cultivos, la tierra residencial y la industria (Fischer & Wang, 2011).

Cuadro 1.Resumen de los tipos de datos

Tipos de datos	Distribucción de datos	Representación de los datos
Datos clásicos	Distribuidos por filas de registros	Mediante medidas estadísticas : suma, frecuencia, media, etc.
Datos espaciales	Distribuidos en un espacio euclidiano(dos dimensiones) Distribuidos en una red espacial (Ejm:las redes de transporte y movilidad)	Mediante centroides, medoides, etc.
Datos espaciales temporales	Distribuidos en una red espacial según un intervalo de tiempo(Ejm:redes de comunicación en un tiempo determinado en una region)	Mediante K-corredores primarios, etc.

Fuente (Shekhar, y otros, 2015).

2.1.1. Tipos de datos espaciales

La presentación de la información de las entidades espaciales es realizada en cuatro tipos:

- Puntos: es una ubicación de un punto único, como una lectura GPS o una dirección geo-codificada.
- Línea: es un conjunto de puntos ordenados, conectados por segmentos de línea recta.
- Polígono: es un área marcada por una o más líneas envolventes, que posiblemente contiene agujeros.
- Grid: una colección de puntos o células rectangulares, organizadas en una red regular.

Los tres primeros tipos de datos son vectoriales y representan las entidades más exactamente posible (Bivand, Pebesma, & Gómez-Rubio, 2008). Mientras que el último, el dato grid está conformada en cada matriz de celdas (o píxeles) organizadas por filas y columnas (o una cuadrícula) que contiene en su interior información, como la temperatura. Los datos grid o rásteres son las fotografías digitales de áreas, imágenes de satélite, imágenes digitales o incluso mapas escaneados (ESRI, 2016). En la Figura Nº 1 se muestran los diferentes tipos de datos espaciales del volcán Maunga Whau de Auckland -Nueva Zelanda (Bivand, Pebesma, & Gómez-Rubio, 2008). En la Figura 1.a se observa un gran número de puntos en

una rejilla regular densa que contiene la altitud del atributo por cada punto para aproximar a la superficie y los tonos grises se utilizan para denotar las clases de puntos existentes.

En la Figura 1.b se observa que se están formando líneas de contorno que conectan puntos ordenados con la misma altura; estos se superponen en la misma figura. En este caso, las líneas de contorno se derivaron de los valores de puntos en la cuadrícula regular. En la Figura 1.c las líneas de contorno para altitudes superiores están cerradas y forman polígonos. Estos polígonos se han formado cuando un conjunto de segmentos de línea forma un objeto cerrado sin líneas que se cruzan. En la Figura 1.d se presenta un área sobre el contorno de 160 m.s.n representada por un polígono con un agujero cuyo centro es parte del cráter, que está por debajo de 160 m.

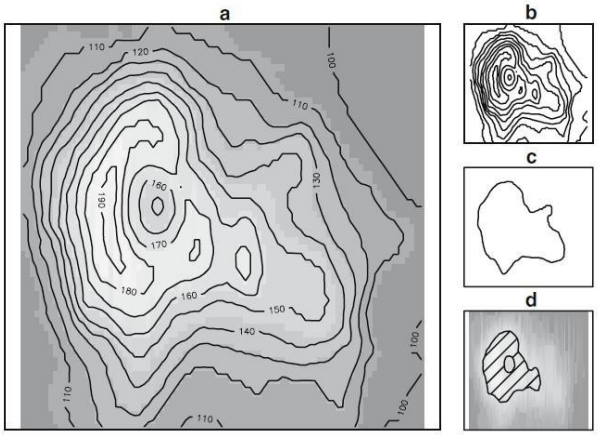

Figura 1. Tipo de datos espaciales: (a) Puntos,(b) Lineas, (c)Polígono y (d) Grid.

Fuente: **(Bivand, Pebesma, & Gómez-Rubio, 2008)**

2.2. Geoestadistica

La geoestadística es un conjunto de métodos estadísticos para el análisis y estimación de datos espaciales. Estos métodos incorporan las características espaciales de los datos reales en los procesos de estimación estadística. Un ejemplo de datos geoestadísticos espaciales es la calidad del agua que asume un gran número de valores dentro de una cuenca subterránea (Boniol & Toth, 1999). Los métodos clásicos de estimación estadística usualmente asumen que dichos datos recolectados son imparciales, es decir, que no están agrupados y que son independientes. En resumen que carecen de estructuras correlacionales, pero en la práctica, se espera que los datos de calidad del agua subterránea muestren un grado de estructura espacial que puedan agruparse alrededor de lugares críticos (Boniol & Toth, 1999).

En la Figura 2 se puede notar que ambos conjuntos de datos tienen las mismas medidas estadísticas sin importar las ubicaciones diferentes; por ejemplo, la altura media de los estudiantes en una aula no se modifica aunque los alumnos cambien de asientos en el salón; pero si se analiza dichos datos mediante una técnica espacial como variograma estos datos distintos debido a su distancia (Canchaya Moya, 2013).

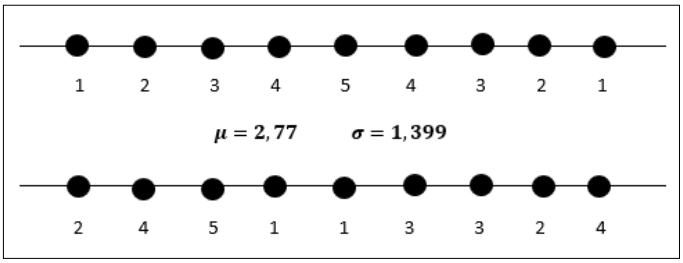

Figura 2. Dos distribuciones media $\boldsymbol{\mu}$ y la desviación $\boldsymbol{\sigma}$ iguales,pero con distancias diferentes.

Fuente **(Facultad de Ciencias Naturales y Museos, 2017)**.

Según Stefan (1934) *"los datos procedentes de unidades geográficas están atados, como racimos de uvas, no separados como las bolas en una urna"*, esto indica que el supuesto de la independencia de las observaciones que se emplea en los cálculos estadísticos tradicionales es poco realista en el marco espacial, debido a que las observaciones próximas tiene valores cercanos; mientras que si están más alejadas entre sí tiene menor relación entre ellas. Por lo tanto, no pueden ser consideradas como magnitudes aleatorias independientes sino como magnitudes relacionadas (Montero Lorenzo & Larraz Iribas, 2008). En otras palabras, una variable en una región o en un espacio no puede ser explicada únicamente por sus condiciones internas, se debe tomar en cuenta su valor con otras regiones vecinas, por esta razón no se puede suponer independencia entre las observaciones muestrales (Moreno Serrano & Vaya Valcarce, 2000). Por ejemplo, las áreas urbanas pobres tienden a estar más cerca a otras áreas pobres (Peña, 2006).

Los orígenes de la Geoestadística tienen sus inicios en la década de 1950, cuando el ingeniero de minería sudafricano Daniel G. Krige publicó por primera vez un método de interpolación basado en la dependencia espacial de las muestras. Posteriormente, en los años 60 y 70, el matemático francés George Matheron desarrolló la teoría de variables regionalizadas (están distribuidas en el espacio y en un tiempo determinado), que proporciona los fundamentos teóricos para los métodos más prácticos de Krige, con esta teoría contribuyó al análisis y estimación de variables espacialmente dependientes, que Matheron denominó

Geoestadística; y adicionalmente agregó el término kriging para la interpolación espacial por métodos geoestadísticos (Trauth, 2015).

La geoestadística reconoce estos factores y utiliza criterios bien definidos para proporcionar las herramientas estadísticas para (1) calcular las estimaciones más precisas, basadas en los resultados de la muestra, (2) cuantificar la exactitud de estas estimaciones y (3) seleccionar los lugares óptimos para ser muestreados (Boniol & Toth, 1999). La Geoestadística describe la autocorrelación de una o más variables en el espacio 1D, 2D o 3D, o incluso en espacio-tiempo 4D, con el fin de hacer predicciones para ubicaciones no observadas, obtener información sobre la exactitud de las predicciones, la propagación de la variabilidad espacial, y la incertidumbre. La forma, el alcance y la dirección de la dependencia espacial se detallan en la herramienta matemática de la geoestadística llamada variograma (Trauth, 2015).

2.3. Outlier

La definición de un outlier, según Douglas Hawkins (Hawkins, 1980), es una observación que se desvía mucho de otras observaciones y que se sospecha fue generada por un mecanismo diferente. Para Barnet y Lewis (Barnett & Lewis, 1994) es una observación periférica que aparece al desviarse notablemente de otros miembros de la muestra. En (Ben-Gal, Outlier detection, 2005) se menciona que Johnson (Wichern & Johnson, 1992) define un outlier como una observación en un conjunto de datos que parece ser incompatible con el resto de datos.

Existen tres tipos de outliers: puntos, contextuales (o condicionales) y colectivos:

- Los outliers de tipo punto son aquellos datos que son significativamente diferentes con el resto de los puntos de datos. Este es el tipo más simple de outlier y es el foco de la mayoría de la investigación sobre la detección de outlier. Como ejemplo de la vida real, se tiene la detección de fraudes de tarjetas de crédito utilizando un conjunto de datos correspondiente a las transacciones de una tarjeta de crédito de un individuo en que la característica de interés es la cantidad gastada. Una transacción para la cual la cantidad gastada es muy alta en comparación con el rango normal de gastos para esa persona será un punto "outlier" (Manoj, 2016). El inconveniente de su detección es definir la desviación para considerar al dato como un outlier (Pei, 2016).

- Los outliers de tipo contextuales (o condicionales) son los puntos de datos que son diferentes de los puntos de datos restantes en un contexto específico, pero no de otro modo (Kumar Singh, 2016). La noción de un contexto es inducida por la estructura del conjunto de datos y debe especificarse como parte de la formulación del problema; por ejemplo, suponga que un individuo por lo general tiene una factura semanal de compras de S/. 100 soles, excepto durante la semana de Navidad, cuando alcanza los S/. 1000 soles. Una nueva compra de S/.1000 en una semana de junio se considerará un outlier contextual, ya que no se ajusta al comportamiento normal del individuo en el contexto del tiempo (aunque la misma cantidad gastada durante la semana de Navidad se considera normal. (Manoj, 2016). Los atributos contextuales están definidos por el contexto; por ejemplo, en conjuntos de datos espaciales, la longitud y la latitud de una ubicación son los atributos contextuales. En cambio, los atributos conductuales están definidos por las características no contextuales de un dato; por ejemplo, en un conjunto de datos espaciales que describe la precipitación media de todo el mundo, la cantidad de lluvia en cualquier lugar es un atributo de comportamiento. El inconveniente en su detección es definir el contexto (Manoj, 2016).
- Los outliers de tipo colectivos se presentan cuando un subconjunto de datos se desvía colectivamente de todo el conjunto de datos, este subconjunto de datos deben estar relacionados; no se considera outlier de este tipo cuando son datos individuales. Por ejemplo, la detección de intrusos cuando un número de computadoras continúan enviando paquetes de denegación de servicio entre sí (Pei, 2016).

Para Osbone y Overbay (Moreno Castellanos, 2012) los aspectos negativos de un outlier son: el aumento de la varianza del error, la reducción del poder de las pruebas estadísticas, la desviación de la distribución normal y el incremento del sesgo en las estimaciones. El aspecto positivo de un outlier es que pueden proporcionar información útil sobre los datos. (Songwon, 2006).

La identificación de un outlier tiene una serie de aplicaciones prácticas en áreas tales como la detección de fraudes, cuellos de botella en la red telefónica, diagnostico médicos, defectos en una línea de producción, entradas inesperadas en las bases de datos entre otras (J. Hodge & Austin, 2004). Para detectar un outlier existen seis enfoques: 1) análisis estadístico, 2) basado en la profundidad, 3) basado en la desviación, 4) basado en la distancia, 5) basado en

la densidad y 6) alta dimensión (Kriegel, Kröger, & Zimek, Ludwig-Maximilians-Universität München, 2010).

2.4. Outlier espacial

En los últimos años, la aplicación de la información geográfica ha generado enormes cantidades de datos espaciales; este incremento ha dado lugar a una nueva rama de la minería de datos, conocida como la minería de datos espacial. La minería de datos espacial es un proceso de descubrimiento de patrones ocultos y es valioso en grandes conjuntos de datos espaciales. Las técnicas de minería de datos espaciales se pueden clasificar en cuatro categorías: clasificación, agrupación, análisis de asociación, y detección de outliers (Shekhar & Xiong, Encyclopedia of GIS, 2007). La minería de datos espacial posee muchas aplicaciones en el sistema de información geográfica tales como: transporte, ecología, seguridad pública, salud pública, climatología y en servicios basados en la localización. (Cao, Liu, Wang, & Zhang, 2013).

En el contexto espacial, un outlier espacial es un objeto referenciado cuyos valores de atributos no espaciales son significativamente diferentes de los otros objetos referenciados espacialmente en su entorno espacial. (Chen, Tien Lu, Kou, & Chen, 2007). Es decir, un outlier espacial es una inestabilidad local, o una observación extrema que se desvía significativamente de su vecindad espacial (Kumar Singh, 2016). Por ejemplo, una casa nueva en un viejo vecindario de un área metropolitana en crecimiento es un outlier espacial basado en la edad de la casa como un atributo no espacial (Kantardzic, 2011). Los outliers espaciales pueden ser de dos tipos: globales y locales:

- Los valores atípicos globales se caracterizan por ser un valor inconsistente en el conjuntos de datos o valores que se desvian de los demás datos; por ejemplo, si 99 de cada 100 puntos tienen un valor entre 300 y 400, pero el punto 100 tiene un valor 750, el punto 100 puede ser outlier global (Salah, 2009).
- Los outliers locales son aquellos que difieren de los demás datos respecto a la vecindad que tienen con los mismos (Shekhar, y otros, 2015).

La detección de los outlier espaciales tiene como objetivo descubrir las inestabilidades locales que rompen la autocorrelación espacial y la variabilidad (Shekhar & Xiong, Encyclopedia of GIS, 2007). Un ejemplo acerca de la detección de outliers espaciales fue la realizada por el meteorólogo británico Sir Gilbert Walker a principios de 1900 cuando

descubrió que las variaciones extremas de presión superficial sobre la línea ecuatorial cerca de Australia están correlacionados con lluvias y la sequía en la India y otras partes del mundo. Esta variación se capturó en una medida, que ahora se llama el Índice de Oscilación del Sur (IOS). Cuando el IOS alcanza valores atípicos (outliers), es decir, cuando se trata de dos o más desviaciones estándar de distancia de la media, la temperatura de la superficie del mar en el Océano Pacífico también se eleva y cae bruscamente. Por lo tanto, un IOS de dos desviaciones estándar por debajo de la media corresponde a un aumento de la temperatura de la superficie, esto producía el fenómeno climatológico llamado "El Niño" y cuando se producía lo opuesto se le denomina "La Niña". La detección de los outliers que descubrió en este estudio contribuyó a la compresión y predicción de dichos fenómenos climatológicos (Chawla & Sun, 2005).

Existen numerosos algoritmos para la detección de los outliers espaciales. Algunos se basan en su representación gráfica tales como: el variograma nube, pocket plots, diagrama de dispersión de Morán y scatter plot. Otros algoritmos aplican pruebas estadísticas para descubrir la inconsistencia local utilizando el enfoque del valor z y z-iterativo. Hay ciertos algoritmos que están diseñados para adaptarse a las propiedades especiales de los datos espaciales, por ejemplo Sherhar et al. (2001), introdujo un método para detectar valores atípicos espaciales en un conjunto de datos gráficos. Zhao et al. (Zhao, Lu, & Kou, 2003) propone un enfoque basado en ondas, Cheng Li (Cheng & Li, 2004) desarrolló un enfoque de escala múltiple para outliers espaciales temporales y Adam et al (Adam, Pursnani Janeja, & Atluri, 2004) propuso un algoritmo que considera la relación espacial y semántica entre los vecinos (Kou, Tien Lu, & Chen, 2006).

Otros autores como (Chen, Tien Lu, Kou, & Chen, 2007) han desarrollado un algoritmo para outliers espaciales con múltiples atributos basándose en la distancia de Mahalanobis Otra novedosa técnica desarrollada por Lijuan Cao et al (Cao, Liu, Wang, & Zhang, 2013) se sustenta en la gráfica del algoritmo del vecino más cercano-KKN Graph.

2.4.1. Detección de outliers espacial

La detección de outliers espaciales (también conocida como identificación de los outliers espaciales) es la búsqueda de datos con características espaciales distintivas de sus vecinos circundantes. Las pruebas bipartidas son métodos típicos de detección espacial multidimensional de valores atípicos. Las pruebas bipartidas utilizan los atributos espaciales

para caracterizar la localización, vecindad y distancia, y los atributos no espaciales para comparar un objeto espacialmente referenciado con sus vecinos.

Shekhar et al. presentó una definición unificada de un outlier espacial. Esta definición no sólo abstrae objetos espaciales a puntos aislados debido a los efectos de vecindad de las propiedades espaciales de objetos (como frontera, tamaño, volumen y ubicaciones) en muchas aplicaciones reales (como transporte, público, seguridad y ubicación- servicios basados).

En primer lugar, se consideró un marco espacial $SF = \langle S, NB \rangle$ donde $S = \{s_1, s_2, \dots, s_n\}$ es un conjunto de ubicaciones y $NB: S\text{x}S \rightarrow \{True, False\}$ es una relación de todo par de vecinos sobre S. Sea $N(x)$ una relación de vecindad de localización x en S haciendo referencia a NB, específicamente $N(x) = \{y | y \in S, NB(x, y) = True\}$.

El valor atípico espacial se define entonces como un objeto $O: S - outlier\left(f, f_{aggr}^{N}, F_{diff}, ST\right)$.

Esta definición es válida si $ST\left\{F_{diff}\left[f(x), f_{aggr}^{N}(f(x)\right]\right\}$ es verdadera, donde R es un conjunto de números reales, $f: S \rightarrow R$ es una función de atributo, $f_{aggr}^{N}: R^N \rightarrow R$ es una función de agregación para los valores de f sobre el vecindario, $F_{diff}: R\text{x}R \rightarrow R$ es una función de diferencia, y $ST : R \rightarrow \{True, False\}$ es un procedimiento de prueba estadística para determinar la significación estadística.

Como un ejemplo en el dominio de aplicación de tráfico de red, la función de agregado de vecindad $f_{aggr}^{N}(x) = E_{y \in N(x)}(f(y))$ es la función de valor de atributo promedio sobre el vecindario $N(x)$.

La función de diferencia $F_{diff}(x)$ se expresa nuevamente como $\mathcal{F}(x) = \left[f(x) - E_{y \in N(x)}(f(y))\right]$, que es la diferencia aritmética entre la función de atributo $f(x)$ y la nueva función agregada de vecindad $E_{y \in N(x)}(f(y))$.

Si $\mu_{\mathcal{F}(x)}$ y $\sigma_{\mathcal{F}(x)}$ son la media y la desviación estándar, respectivamente, de la nueva función de diferencia $\mathcal{F}(x)$, entonces la función de prueba de significación T se puede definir a $Z_{\mathcal{F}(x)} = \left|\frac{\mathcal{F}(x) - \mu_{\mathcal{F}(x)}}{\sigma_{\mathcal{F}(x)}}\right| > 0$.

Dada la definición matemática, existen varias herramientas o métodos estadísticos disponibles para la detección espacial de valores atípicos.

La literatura de estadística espacial proporciona dos tipos de pruebas multidimensionales bipartitas, a saber, pruebas gráficas y pruebas cuantitativas.

Las pruebas gráficas, que normalmente son el variograma nube, los diagramas de dispersión o los diagramas de dispersión de Morán, ilustran (visualizan) la distribución de la diferencia de vecindad en una figura e identifican puntos en porciones particulares de la Figura como valores atípicos espaciales.

- El variograma nube muestra los puntos de datos relacionados por la vecindad. Las ubicaciones cercanas entre sí pero con grandes diferencias de atributos podrían indicar un outlier espacial.
- Las gráficas de dispersión muestran los valores de los atributos en el eje X y el promedio de los valores de los atributos en la vecindad en el eje Y. Se utiliza una línea de regresión de mínimos cuadrados para identificar los valores atípicos espaciales.
- El diagrama de dispersión de Morán, trazar los valores de los atributos normalizados frente a los valores medios de vecindad de los atributos normalizados y los valores atípicos son los puntos rodeados por vecinos de alto o bajo valor inusualmente.

Sin embargo, las pruebas gráficas están limitadas por la falta de criterios precisos para distinguir el outlier espacial. Aunque comparten tecnologías comunes con las pruebas gráficas, las pruebas cuantitativas proporcionan una prueba más precisa para distinguir los valores atípicos espaciales. Los valores atípicos espaciales detectados por los diagramas de dispersión y el diagrama de dispersión de Morán son casos especiales (Shekhar, Evans, Kan, & Mohan, 2011) que tienen las siguientes caracteristicas:

- Un outlier de Morán es un caso especial de outlier espacial, y se detecta un punto localizado en el cuadrante superior izquierdo o inferior derecho del diagrama de dispersión de Morán.
- Un diagrama de dispersión es también un caso especial de un outlier espacial, y se define como un punto con un error residual estandarizado significativo de la línea de regresión de mínimos cuadrados en el diagrama de dispersión.

- El valor z se utiliza para detectar valores atípicos espaciales para un valor asignado normalmente distribuido $f(x)$.Para cada ubicación x con un valor de atributo
- $f(x)$, se detecta el valor atípico si $Z_{\mathcal{F}(x)} = \left|\frac{\mathcal{F}(x)-\mu_{\mathcal{F}(x)}}{\sigma_{\mathcal{F}(x)}}\right| > \Theta$, donde $\mathcal{F}(x)$ es la diferencia entre el valor del atributo en la posición xy el valor medio del atributo de $x's$ vecinos, $\mu_{\mathcal{F}(x)}$es el valor medio de $\mathcal{F}(x)$ y $\sigma_{\mathcal{F}(x)}$ es el valor de la desviación estándar de $\mathcal{F}(x)$en todas las estaciones.La elección de Θdepende de un nivel de confianza especificado, por ejemplo, $\Theta \approx 2$ dado un nivel de confianza del 95%.

Además, debido a los diversos formatos y semántica en los datos espaciales, los algoritmos espaciales de detección de valores atípicos están diseñados para acomodar las propiedades especiales de los datos espaciales dados.
Por ejemplo, Shekhar et al. introdujo un método de detección de valores atípicos espaciales para un conjunto de datos de gráficos, y Zhao et al propuso un método basado en la wavelet para detectar valores atípicos de la región en datos meteorológicos.

2.5. Variograma

El variograma es el eje fundamental de la geoestadística (Cressie & Hawkins, 1980). Un variograma explica la dependencia espacial de las observaciones referenciadas en un espacio unidimensional o multidimensional. El análisis de la dependencia espacial o autocorrelación espacial permite describir la distribución de los valores en el espacio, cuantificar las correlaciones o redundancias de información entre valores medidos en sitios diferentes, determinar el tamaño de la "zona de influencia" de una observación, así como detectar los cambios de direcciones de una variable que se encuentran distribuida en una región de manera continua o aleatoria (Emery, 2013)

Dado que el variograma de un proceso espacial suele ser desconocido, tiene que ser estimado a partir de las observaciones. Este procedimiento se denomina variografía. La variación se realiza calculando el variograma experimental a partir de los datos iniciales. En el siguiente paso, el variograma experimental se resume en el estimador del variograma. La variografía concluye con un modelo de variograma del estimador. El variograma experimental se calcula como las diferencias entre pares de valores observados y es dependiente del vector de separación h. El variograma experimental clásico se define por la semivariancia como se denota en el ecuación (1),

$$\gamma(h) = 0.5(z_x - z_{x+h})^2 \qquad (1)$$

Donde z_x es el valor observado en la posición x , y el z_{x+h} es el valor observado en otro punto a una distancia h de x. Estos valores se pueden trazar en función de la distancia de retraso espacial, y esta trama se denomina "variograma nube" que muestra la propagación de los valores según cada retraso (Bostan, 2017).La longitud del vector de separación h se llama la distancia de retraso, o simplemente el retraso. El término correcto para $\gamma(h)$ es el semivariograma (o semivariance), donde la expresión “semi” significa la mitad de la varianza en las diferencias entre z_x y z_{x+h}; pero la varianza por punto cuando los puntos se consideran en parejas. Convencionalmente, $\gamma(h)$ se denomina un variograma en lugar de un semivariograma (Trauth, 2015).

2.6. Variograma Nube

El variograma nube es un gráfico de dispersión que se utiliza para la exploración de datos en busca de outliers espaciales y también se emplea para evaluar la variabilidad con el aumento de la distancia (Kim, 2015). Para ello se trazan dos ejes x e y, donde el eje vertical es la mitad de las diferencias cuadráticas entre los valores de la variable de interes para todos los pares de ubicaciones espaciales llamada semivarianza,versus el eje horizontal es la distancia euclidiana entre los puntos asociados (Laurent, Ruiz-Gazen, & Thomas-Agnan, 2012). La ultima distancia de corte es elegida por el investigador (M. Fischer & Nijkamp, 2013). Las diferencias cuadráticas que se utilizan para crear en el variograma nube en el estudio permite determinar las tendencias espaciales en los datos que podrían no haber sido evidentes en las estadísticas descriptivas que se produjeron para el mismo conjunto de datos (Taylor, 2004).

Un variograma nube es construido para cualquier conjunto de datos, la única condición es que estén referenciados geográficamente; es decir, que posean una longitud y una latitud cada uno de ellos (Haslett, Bradley, Craig, & Wills, 1991). A partir del variograma nube se puede calcular el variograma. (Kanevski & Maignan, 2004).

Para identificar los outliers en un variograma nube, se reconoce por valores altos en el eje vertical de la grafica sin importar cuál sea la distancia, ya que afecta al cálculo

independientemente de que sea un valor atípico verdadero, un dato de asimetría o una observación incorrecta (Negreiros, 2004).

Según Ploner (1999), el variograma nube se utiliza para detectar outliers globales que se caracterizan porque sus distancias de las diferencias cuadráticas de los pares formados, serán valores significativmente mas grandes que el resto de valores en la nube. Por otro lado, los outliers locales difieren de sus valores vecinos ya que presentan diferencias cuadráticas altas pero distancia cercanas. Los outlier locales son más disimulados que globales debido a que se comportan forma normal en distancia medias a grandes (Kapageridis, 2015). El variograma nube muestra dependencia espacial; es decir, si las diferencias entre pares se harcen más variables a medida que aumenta la distancia entre pares de posiciones (Symanzik, Megretskaia, Majure, & Cook, 1996).

En la Figura N° 3, se muestra el variograma nube del estudio del nivel de las aguas subterráneas en la zona sur de Al Jabal Al Akhdar (Salah, 2009). En la Figura 3 se obsera los diferente tipos de outliers espaciales: globales y locales (Ploner, 1999).

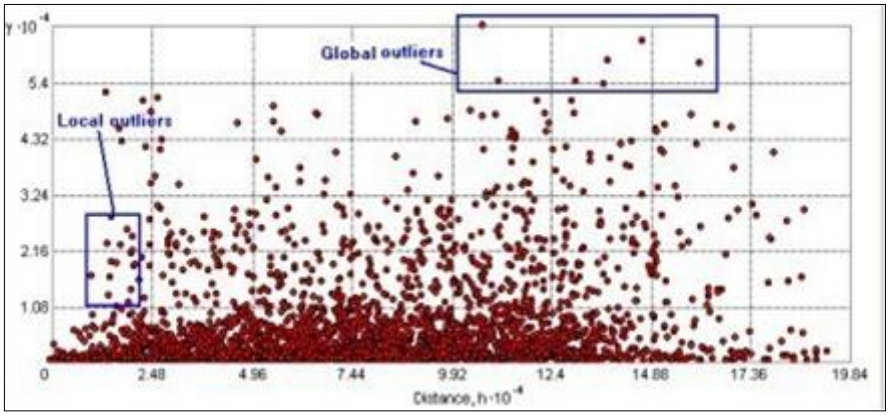

Figura 3. El variograma nube con outliers espaciales globales y locales .
Fuente (Salah, 2009).

Existe una limitación para trabajar graficar el variograma nube a medida que aumenta el número de observaciones en un conjunto de datos, el número de pares de datos se vuelve muy grande y la multitud de puntos difícilmente puede distinguirse en la parcela., por ende su cálculo se vuelven extremadamente lento; por esta razón, la creación de la gráfica se limita actualmente con menos de 5000 puntos (Haslett, Bradley, Craig, & Wills, 1991).

2.6.1. Construcción del variograma Nube

Para obtener el gráfico se debe medir la variabilidad de una variable regionalizada $z(x)$, $\{z(x): x \in D \subset R^n\}$, en diferentes escalas mediante el cálculo de la disimilitud entre pares de valores de datos $z_\alpha = z(x_\alpha)$ y $z_\beta = z(x_\beta)$ es decir, que se encuentra en los puntos de x_α y x_β donde $\{x_\alpha : \alpha = 1,..., N\}$ y $\{x_\beta : \beta = 1,..., N\}$ pertenece al dominio espacial D.

La medida de la disimilitud de los dos valores es definida por la ecuación (2):

$$\gamma^*_{\alpha,\beta} = \frac{(z_\alpha - z_\beta)^2}{2} \qquad (2)$$

La mitad del cuadrado de la diferencia entre los dos pares de valores. Los dos puntos x_α, x_β en el espacio geográfico pueden ser unidos por un vector $h_{\alpha,\beta} = x_\beta - x_\alpha$ como se muestra en la Figura 4.

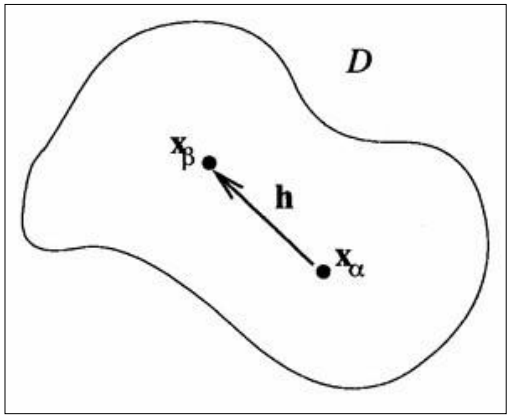

Figura 4. Un vector h que une dos puntos ($x_\alpha, x_\beta = x_\alpha + h$) en el espacio 2D.

Fuente: (Wackernagel, 2003)**.**

La disimilitud γ^* depende de la separación y de la orientación del par de los puntos descritos por el vector **h** esta denotado por la ecuación (3)**.**

$$\gamma^*(h_{\alpha,\beta}) = \frac{1}{2}(z(x_\alpha + h) - z(x_\alpha))^2 \qquad (3)$$

La disimilitud es simétrica con respecto a **h** denotado por la ecuación (4)**:**

$$\gamma^*(-h_{\alpha,\beta}) = \gamma^*(+h_{\alpha,\beta}) \qquad (4)$$

El uso de todos los pares de muestras en conjunto de datos (hasta la distancia de la mitad del diámetro de la región) genera un gráfico de las diferencias γ^* versus la separación espacial **h** originando el variograma nube; es decir, la disimilitud producen el variograma nube sobre las distancias entre las estaciones (Wilcke, Leuprecht, & Gobiet, 2012). Cuando los pares de puntos se presentan con una alta disimilitud geográfica a través de una pequeña distancia se consideran inicialmente como potenciales outliers (O'Leary, Reiners Jr, Xu, & D. Lemke, 2016). Un ejemplo esquemático se presenta en la figura 5.

Los puntos presentes en la gráfica del variograma nube se clasifican para estimar la influencia de los valores de la muestra individual sobre los valores experimentales del variograma con mayor precisión (Kresse & Danko, 2011).

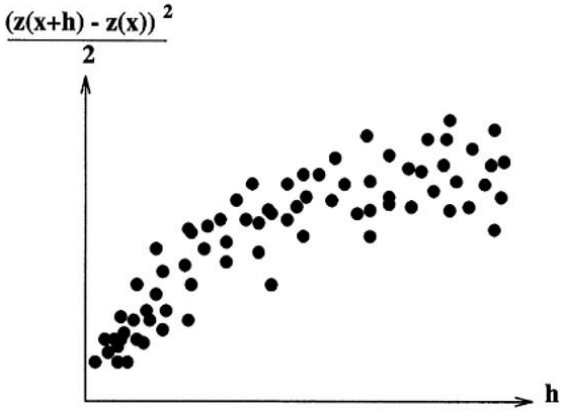

Figura 5.Gráfica de la disimilitud γ^* versus la separación espacial h de pares de muestras.

Fuente (Wackernagel, 2003)**.**

Según (Chilès & Delfiner, 2009), las coordenadas espaciales de un variograma nube está conformado por:

- Eje X: La distancia por parejas entre vecinos: $\mathrm{h}_{\alpha,\beta} = \left| x_\beta - x_\alpha \right|$
- Eje Y: La mitad del cuadrado de la diferencia entre los dos pares

El variograma nube puede mostrar comportamientos diferentes a lo largo de las diferentes direcciones de la separación $\mathrm{h}_{\alpha,\beta} = \left| x_\beta - x_\alpha \right|$; es decir, mostrar anisotropía (Chiles & Delfiner, 1999) que significa que los conjuntos de datos presentan una fuerte dependencia espacial, la varianza en las diferencias del atributo aumentará con el aumento de la distancia entre ubicaciones (Mahalik, 2012).

El variograma nube es un método de identificación de valores atípicos pero con un enfoque cualitativo porque, a diferencia del boxplot, no hay límites especificados dentro del gráfico para definir lo que constituye un valor atípico espacial. Por esta razón, el proceso de identificación de los valores atípicos espaciales en el variograma nube es subjetivo (O'LEARY, 2014).

La interpretación del variograma nube se explica con la autocorrelación espacial o la dependecia espacial en el conjunto de datos (Virrantaus, 2015). Dado que la estructura dependencia espacial es a menudo desconocida, debe ser estimada. En este caso, se considera el estimador propuesto por Cressie y Hawkins en 1980,como se muestra en ela ecuación(5):

$$\gamma(h) = \frac{1}{2(0.457 + \frac{0.494}{|N(h)|})} \left(\frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} |Z(s_i) - Z(s_j)|^{1/2}\right)^4 \qquad (5)$$

Donde $N(h)$ es el numero de pares, $N(h) = \{(i, j) : s_i - s_j = h\}$ (Meilán Vila, 2016).

En el cuadro N°2 se presentan los siguientes datos como ejemplo para calcular el variograma nube propuesto en el libro "Mathematical Methods for Engineers and Geoscientists" (Olga, 2008):

Cuadro 2. Datos para el variograma nube

x	0.0	1.0	2.0	3.0	4.0	5.0
z	7.0	3.0	6.0	7.0	0.0	3.0

Para poder usar el variograma nube, debe tener una medida de distancia, para ello se tomarán en cuenta estas cuatro distancias posibles: $h_{ij} = 0,1,2,3,4,5$; mediante la distancia de pares se obtendrá los valores de disimilitudes para construir el variograma, con los valores "x" y "z",donde "x" es el dato espacial y "z" es el atributo espacial de dicho dato.

En el cuadro 3 se muestra la distancia de los pares con sus respectivos disimilitudes.

Cuadro 3. La distancia entre los pares con sus respectivos disimilitudes

Distancia $h_{ij} = \|x_i - x_j\|$	**Disimilitud** $\gamma_{ij} = 0.5(z(x_i) - z(x_j))^2$
0	Seis pares llevan a 0
1	Cinco pares con las siguientes coordenadas (0,1),(1,2),(2,3),(3,4),(4,5) $0.5(z(0)-z(1))^2=0.5(7-3)^2=8$ $0.5(z(1)-z(2))^2=0.5(3-6)^2=4.5$ $0.5(z(2)-z(3))^2=0.5(6-7)^2=0.5$ $0.5(z(3)-z(4))^2=0.5(7-0)^2=24.5$ $0.5(z(4)-z(5))^2=0.5(0-3)^2=4.5$
2	Cuatro pares con las siguientes coordenadas (0,2),(1,3),(2,4),(3,5) $0.5(z(0)-z(2))^2=0.5(7-6)^2=0.5$ $0.5(z(1)-z(3))^2=0.5(3-7)^2=8$ $0.5(z(2)-z(4))^2=0.5(6-0)^2=18$ $0.5(z(3)-z(5))^2=0.5(7-3)^2=8$
3	Tres pares con las siguientes coordenadas (0,3),(1,4),(2,5) $0.5(z(0)-z(3))^2=0.5(7-7)^2=0$ $0.5(z(1)-z(4))^2=0.5(3-0)^2=4.5$ $0.5(z(2)-z(5))^2=0.5(6-3)^2=4.5$
4	Dos pares con las siguientes coordenadas (0,4) y (1,5) $0.5(z(0)-z(4))^2=0.5(7-0)^2=24.5$ $0.5(z(1)-z(5))^2=0.5(3-3)^2=0$
5	Un par de coordenada (0,5) $0.5(z(0)-z(5))^2=0.5(7-3)^2=8$

Luego de obtener los valores de disimilitudes de las cincos distancias h_{ij}, se procede a construir el variograma, donde el eje x son los valores de las distancias entre pares h_{ij} y en el eje y son los valores de disimilitudes de los pares.

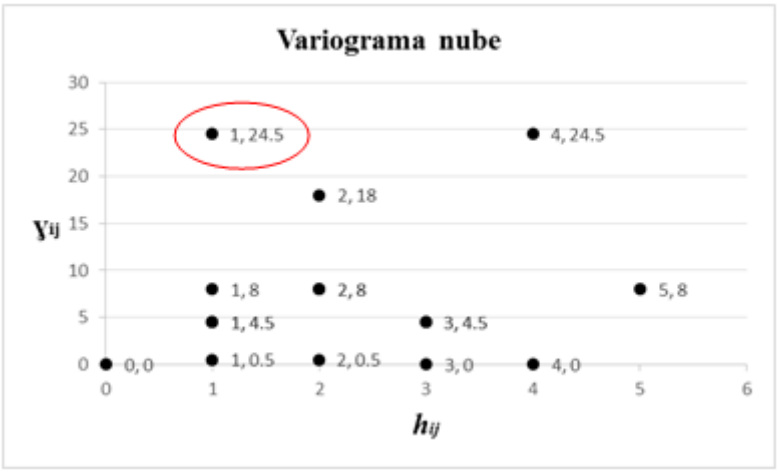

Figura 6. Gráfica del Variograma Nube.

Fuente elaboración propia

En la Figura 6 se puede considerar el punto (1, 24.5) como un outlier debido a que tiene un valor alto de disimilitud en una distancia tan corta. Debe tomar en cuenta que el variograma nube muestra pares de puntos. Los outliers espaciales serán aquellos pares de puntos que son diferentes, pero relativamente cercanos entre sí.

2.6.2. Variograma Nube en R

En R (R Core Team, 2016) existe una librería de llamada GeoXp que fue elaborado por Yves Aragon, Thibault Laurent, Lauriane Robidou, Anne Ruiz-Gazen, Christine Thomas-Agnan en el año 2015 para el análisis exploratorio espacial de los datos, en el caso de variograma nube la función que se utilizará es "variocloudmap" (R project, 2015).

2.7. Índice de Morán

El índice Morán (IM) es una medida de correlación no espacial en un contexto espacial y es uno de los indicadores más antiguos de la autocorrelación espacial.La autocorrelación espacial se basa en la primera ley de la geografía enunciada por Waldo Tobler que dice lo siguiente: las cosas cercanas son más similares que las más lejanas (Byun, Huh, Yu, & Kim, 2007). En otras palabras, el IM mide cómo un objeto es similar a los demás que lo rodean. Si los objetos son atraídos por cada otro, significa que las observaciones no son independientes (WordPress, 2016).

La autocorrelación espacial positiva se obtiene cuando las áreas vecinas son similares o iguales. El índice de Morán puede medir tanto la tendencia global mediante una tendencia local mediante del conjunto de datos (Byun, Huh, Yu, & Kim, 2007).

2.8. Diagrama de dispersión de Morán

El diagrama de dispersión de Morán nos permite visualizar el tipo y la fuerza de la autocorrelación espacial en una distribución de datos (Long, 2017).Este diagrama se basa en el índice de Morán,es una gráfica bivariada donde índice de Morán es representado como la pendiente de una línea de regresión, al trazar esta línea junto con los pares de coordenadas de la regresión,nos permite identificar los posibles outlier espaciales como también obtener una visión de la estructura local de los datos (Plant, 2012).Esta se debe porque la gráfica combina concepto de la regresión lineal clásico para detección de outlier espaciales (Vila, 2016).

En el DDM está representa por un plano cartesiano donde eje X esta representado por la variable previamente estandarizada y en el eje Y se representa el retardo espacial de dicha variable estandarizada.Se entiende por retardo espacial al promedio ponderado de los valores que adopta una variable en el subconjunto de observaciones vecinas a una dada. Por ejemplo, el retardo espacial de la variable renta per cápita de la provincia de Madrid podría obtenerse como una media aritmética simple de los valores de renta per cápita en las provincias limítrofes (Yrigoyen, 2017).

El DDM se forma cuando las variables se expresan en forma estandarizada (con media cero y desviación estándar igual a uno), esto permite una evaluación tanto de la asociación espacial global gracias a que la pendiente de la línea es el índice de Morán, como también de la asociación espacial local (tendencias locales en el diagrama de dispersión mediante los cuadrantes) (Sirgy, Phillips, & Rahtz, 2009).

En esta gráfica, se relaciona para cada observación, el valor de la variable en la misma y el valor promedio en sus correspondientes observaciones vecinas, la pendiente de la recta de regresión que es el valor del Indice de Morán de autocorrelación espacial globa;cuanto mayor sea el valor de este estadístico, es decir, el ángulo que forme la recta de regresión con el eje de abscisas, más fuerte será el grado de autocorrelación espacial en la variable, y

viceversa. En la Figura 7, la variable tasa de paro de los barrios madrileños tiene un mayor grado de dependencia espacial que la variable población (Yrigoyen, 2017).

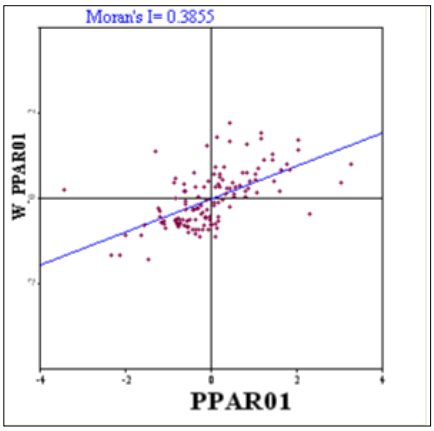

Figura 7. Diagrama de dispersión de Morán de los barrios madrileños respecto a su tasa de paro.
Fuente: (Yrigoyen, 2017)**.**

Este diagrama de dispersión suele dividir el tipo de asociación espacial en cuatro categorías: dos para autocorrelación espacial positiva (valores altos de una variable rodeados de valores altos o valores bajos rodeados de valores bajos) y dos para autocorrelación espacial negativa (valores altos rodeados por valores bajos, y viceversa) (Yrigoyen, 2017). Esto se puede representa en cuatro cuadrantes, los cuadrantes I y III corresponde la asociación espacial positiva y los dos restantes, II y IV, corresponde con asociación (Yrigoyen, 2017). En estos dos últimos cuadrantes se localizan los outliers (Sirgy, Phillips, & Rahtz, 2009).Esto se puede observar en la Figura 8:

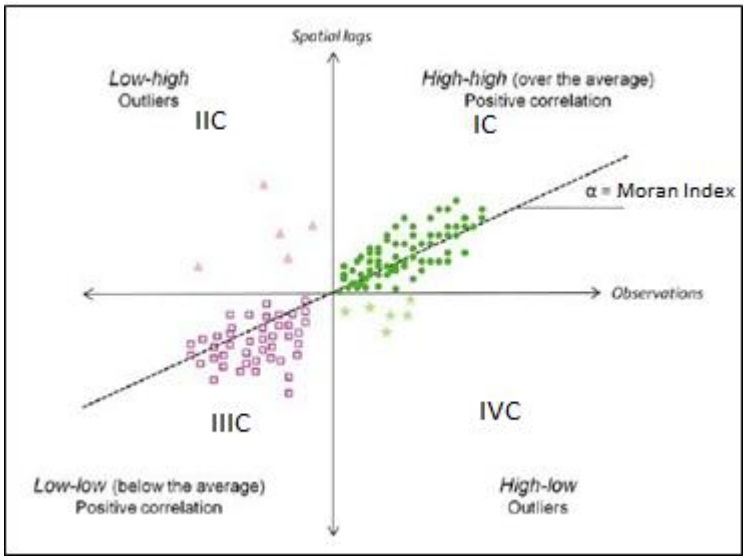

Figura 8. Diagrama de dispersión de Morán la distribución de los cuadrantes.
Fuente (Gómez, White, & Wulder, 2011)

Los puntos en el diagrama de dispersión que son extremos con respecto a la tendencia central reflejada por la pendiente de regresión pueden ser atípicos en el sentido de que no siguen el mismo proceso de dependencia espacial que la mayor parte de las otras observaciones. Por lo tanto, podrían considerarse focos de no estacionariedad local, especialmente si corresponden a lugares o puntos fronterizos espacialmente contiguos. La presencia de valores atípicos también puede indicar problemas con la especificación de la matriz de pesos espaciales o con la escala espacial en la que se registran las observaciones.

Del mismo modo, las observaciones que ejercen una gran influencia o influencia en la pendiente de regresión son de interés, una vez más, sobre todo si se agrupan espacialmente o corresponden a puntos fronterizos. Este último caso proporciona una forma de evaluar la influencia de los valores límite en la medida global de la asociación espacial (Fischer, Scholten, & Unwin, 2005).

Los DDM se pueden clasificar en dos categorías: el primero denominado “dinámico de Morán” ya que cambia dinámicamente cuando se incluyen diferentes observaciones en el cálculo y el segundo denominado “coropletico” qu es igual a un diagrama geográfico de coropletas (Ping, Zartman, Bronson, & Green, 2004). Un DDM coropletico es un mapa temático que representa la distribución espacial de datos mediante tonos de color en los que la intensidad de éste expresan los intervalos de datos en unidades territoriales por ejemplo los límites municipales (Sittón, 2017). Mediante sombreado, tintes o patrones que utilizan para mostrar las diferencias de un valor en proporción en una ubicación geográfica o región; esto le permite identificar rápidamente estas diferencias relativas con sombreados que va del claro (valores menos frecuentes o inferiores) a oscuro (más frecuentes o superiores). Los mapas coropléticos son útiles para mostrar información cuantitativa en un mapa, para mostrar las relaciones y patrones espaciales, cuando los datos están normalizados, cuando se trabaja con datos socioeconómicos y para obtener una visión general de la distribución en las ubicaciones geográficas (Microsoft, 2017).

2.8.1. Construcción del diagrama de dispersión de Morán

Para explicar DDM se denota las observaciones z_i es la variable estandarizada $(x_i - \bar{x})/S_x$, donde $\bar{x}$ y S_x corresponde a la media y desviación estándar de los valores de los atributos x_i, $i = 1, ..., n$ respectivamente.El DDM indica el grado de asociación lineal entre los valores observados estandarizados z_i y un promedio ponderado de los atributos normalizados vecinos, $Y_{w,i} = \sum_j w_i z_j, i = 1, ..., n$ donde $w_{ij} > 0$ y $\sum_j w_{ij} = 1$ (Meilán Vila, 2016).

Dado que z_i son observaciones esta desviando respecto a su media y $Y_{w,i}$ es el retraso espacial asociado,el diagrama de dispersión se centra (0,0) donde divide en cuatro cuadrantes donde representa las diferentes asociaciones entre z_i y w_i. Los cuadrantes superior e inferior izquierdo representan una asociación positiva en el sentido de que una observación en un lugar tiene valores similares a los de su vencidario.Para el cuadrante superior izquierdo e inferior derecho corresponden a la asociación negativa es decir, los valores bajos son redondeados por valores altos (superior izquierdo) y los valores altos están rodeados por valores bajos (inferior derecha). Como se ha mencionado, los pares $(Y_{w,i}, z_i)$ se dan para valores de estandarización, de modo que los valores atípicos pueden visualizarse fácilmente como puntos más allá de dos unidades fuera del origen(0,0) (Meilán Vila, 2016).

Los puntos del DDM que son extremos con respecto a la tendencia central que se puede ver reflejada por la pendiente pueden considerarse valores atípicos o outliers debido que en el sentido que tienen no siguen el mismo proceso de dependencia espacial que presenta la mayoría de las otras observaciones. Una observación intuitiva de los outliers puede basarse en los residuos normalizados de la regresión de Y_w en z (Meilán Vila, 2016).
Tomando en cuenta modelo de regresión lineal clásico se tendría la siguiente la ecuación (6)

$$Y_{w,i} = \beta_0 + \beta_1 z_i + \varepsilon_i \quad , i = 1, ..., n \qquad (6)$$

Donde $\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2)$ y indepediente. Por consiguiente, los valores ajustados $\hat{Y}_{w,i}$ se podría expresar en la ecuación (7):

$$\hat{Y}_{w,i} = \hat{\beta}_o + \hat{\beta}_1 z_i \ , \ i = 1,..,n \qquad (7)$$

Donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de respectivamente. En resumen, tendría los siguientes errores o residuos de predicción como se muestra en la ecuación (8):

$$\hat{\varepsilon}_i = Y_{w,i} - \hat{Y}_{w,i} = Y_{w,i} - \hat{\beta}_0 - \hat{\beta}_1 z_i \ \ , i = 1,..,n \qquad (8)$$

La idea es elegir los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ que brinden los residuos de regresión más pequeños. Se puede demostrar, usando mínimos cuadrados, como se muestra en la ecuación (9) y (10):

$$\hat{\beta}_0 = \bar{Y}_w - \hat{\beta}_1 \bar{z} \qquad (9),$$

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} (z_i - z)\, Y_{w,i}}{S_{zz}} \quad \text{donde } S_{zz} = \sum_{j=1}^{n} (z_j - \bar{z})^2 \qquad (10)$$

Por lo tanto, podemos expresarlo en la siguiente ecuación (11):

$$\hat{Y}_{w,i} = \bar{Y}_w + \hat{\beta}_1 (z_j - \bar{z}) = \frac{1}{n} \sum_{j=1}^{n} Y_{w,i} + \sum_{j=1}^{n} \frac{(z_j - \bar{z})}{S_{zz}} Y_{w,j} (z_i - \bar{z})$$

$$= \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{S_{zz}} \right] Y_{w,j} = \sum_{j=1}^{n} h_{ij} Y_{w,j} \quad \text{donde } h_{ij} = \frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{S_{zz}} \qquad (11)$$

Asi también, el valor de $\hat{Y}_{w,i}$ será un promedio ponderado de todas las respuesta $Y_{w,j}$con pesos h_{ij}. Entonces, el peso ejercido por un individuo en su propia predicción sería de esta manera como se muestra en la ecuación (12):

$$h_{ii} = \frac{1}{n} + \frac{(z_i - \bar{z})^2}{S_{zz}} \qquad (12)$$

Que si el valor h_{ii} es grande su abscisa estará lejos de la media.La cantidad h_{ii} se conoce como conjunto de datos de los i-ésimo datos. Depende solamente del valor de la variable explicativa, en este caso z_i, e indica el peso que $Y_{w,i}$ en su propio ajuste $\hat{Y}_{w,i}$. Obsérvese que h_{ij} son elementos de una matriz idempotente y simétrica H. Como consecuencia $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$ (Meilán Vila, 2016).

Para detectar los outliers espaciales, se emplearán los residuos de la regresión. Obsérvese que los residuos pueden escribirse de esta maneta como se muestra en la ecuación (13):

$$\hat{\varepsilon}_i = Y_{w,i} - \hat{Y}_{w,i} = \beta_0 + \beta_1 z_i + \varepsilon_i - \sum_{j=1}^{n} h_{ij} Y_{w,j} = \beta_o + \beta_1 z_i + \varepsilon_i - \sum_{j=1}^{n} h_{ij} (\beta_0 + \beta_1 z_j + \varepsilon_j)$$

$$= \beta_0 + \beta_1 z_i + \varepsilon_i - \beta_0 \sum_{j=1}^{n} hij - \beta_1 \sum_{j=1}^{n} h_{ij}\varepsilon_j = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j. \qquad (13)$$

Las expresiones $\sum_{j=1}^{n} h_{ij}$ y $\sum_{j=1}^{n} h_{ij} z_j$ se puede obtener considerado las dos siguientes ecuaciones (14) y (15):

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n}\left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{S_{zz}}\right] = \frac{n}{n} + \frac{(z_i - \bar{z})}{S_{zz}} \sum_{j=1}^{n} (z_j - \bar{z}) = 1. \qquad (14)$$

$$\begin{aligned} \sum_{j=1}^{n} h_{ij} z_j &= \sum_{j=1}^{n}\left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{S_{zz}}\right] z_j = \bar{z} + \frac{(z_i - \bar{z})}{S_{zz}} \sum_{j=1}^{n} (z_j - \bar{z}) z_j \\ &= \bar{z} + \frac{(z_i - z)}{S_{zz}} \sum_{j=1}^{n} (z_j - \bar{z})^2 - \bar{z}\frac{(z_i - \bar{z})}{S_{zz}} \sum_{j=1}^{n} (z_j - \bar{z}) = z_i \end{aligned} \qquad (15)$$

Bajo el supuesto de Anselin (Anselin, Spatial Econometrics: Methods and Models, 1998), dado que los ε_i son variables aleatorias independientes, con una media cero con varianza común σ^2. El valor esperado de los $\hat{\varepsilon}_i$ se puede expresar de la siguiente ecuación (16):

$$E(\hat{\varepsilon}_i) = E\left(\varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j\right) = E(\varepsilon_i) - \sum_{j=1}^{n} h_{ij} E(\varepsilon_j) = 0. \qquad (16)$$

y la varianza de los $\hat{\varepsilon}_i$ como se expresa en la ecuación (17) :

$$\begin{aligned} Var(\hat{\varepsilon}_i) &= Var(\varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j) = Var(\varepsilon_i) + Var(\sum_{j=1}^{n} h_{ij}\varepsilon_j) - 2Cov(\varepsilon_i, \sum_{j=1}^{n} h_{ij}\varepsilon_j) \\ &= \sigma^2 + \sum_{j=1}^{n} h_{ij}^2 \sigma^2 - 2h_{ii}\sigma^2 = \sigma^2(1 - h_{ii}). \end{aligned} \qquad (17)$$

Dado que cada residuo tiene una varianza diferente, dependiendo de su conjunto de datos, los residuos estandarizados se utilizarán para detectar los outliers. Dado que la varianza del error es a menudo desconocida, debe ser estimada. Un estimador natural sería la varianza de la muestra, que se denotará por σ^2. Por consiguiente, estos nuevos residuos de regresión se pueden expresar en la ecuación (18):

$$d_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad (18)$$

Por lo tanto, las observaciones con residuos estándar demasiado grandes (en valor absoluto) muestran que los datos son atípicos de alguna manera. Se puede considerar candidato para ser un outlier las observaciones que han estandarizado residuos mayores o menores que 2 o

-2, respectivamente, que serían aproximadamente los quantiles de una distribución normal estándar que contiene más del 95% de las observaciones (Meilán Vila, 2016).

En la Figura 9 se muestra cinco regiones en una disposición espacial de sus unidades (Kassel, 2017).

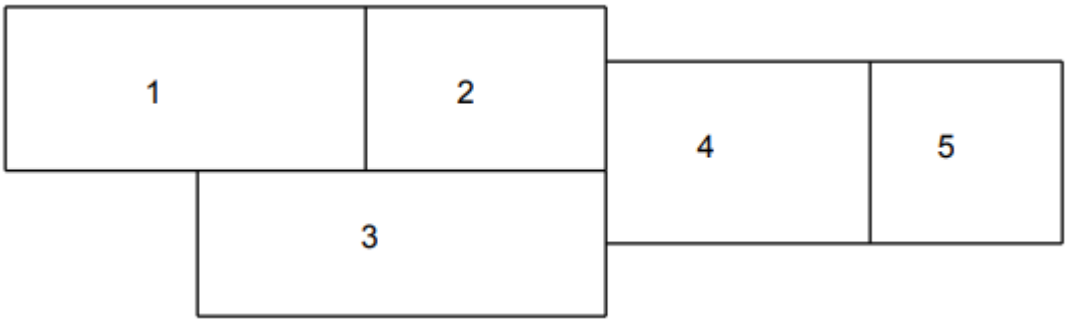

Figura 9.La disposición espacial de las cinco regiones.

Fuente : (Kassel, 2017).

El primer paso es establecer la matriz de pesos a partir de la cantidad de vecinos que tiene cada unidad espacial. Se utilizó la relación de contigüidad tipo Rook o torre, que es un modo de recuento que siguen los regla de movimientos de dicha Figura de ajedrez;es decir, considera vecinas a aquellas unidades espaciales que comparten una arista (Anguera Argilaga, 1999). Por ejemplo, la región 3 tiene tres vecinos a los que les corresponde un 1 en la matriz y 0 al resto.

En el cuadro 4 se muestra la matriz de los pesos $\boldsymbol{W_{ij}}$.

Cuadro 4.Matriz de los pesos

Región	1	2	3	4	5	Total
1	0	1	1	0	0	2
2	1	0	1	1	0	3
3	1	1	0	1	0	3
4	0	1	1	0	1	3
5	0	0	0	1	0	1
∑∑	Suma de pesos (# de pesos distintos de cero)					12

En el siguiente cuadro 5 se muestra la matriz estandarizada de los pesos:

Cuadro 5.Matriz estandarizada de los pesos

Región	1	2	3	4	5
1	0	1/2	1/2	0	0
2	1/3	0	1/3	1/3	0
3	1/3	1/3	0	1/3	0
4	0	1/3	1/3	0	1/3
5	0	0	0	1	0

Fuente (Kassel, 2017)

El siguiente paso es calcular la media $\bar{z}$ y desviación estándar S_z de los valores x_i,en el cuadro 6 se muestra las regiones con sus respectivos valores x_i,

Cuadro 6. Las regiones con sus respectivos valores x_i

Regiones	1	2	3	4	5
x_i	8	6	6	3	2

Fuente (Kassel, 2017)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(8+6+6+3+2) = \frac{1}{5}(25) = 5.$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(8-5)^2 + (6-5)^2 + (6-5)^2 + (3-5)^2 + (2-5)^2}{5} = 4.8$$

$$S_x = \sqrt{4.8} = 2.191.$$

En el cuadro 7 se muestra como se calcula la variable estandarizada z_i y retardo espacial $Y_{w,i}$.

Cuadro 7.Cálculo de la variable estandarizada y el retardo espacial

Variable estandarizada z_i	Retardo Espacial $Y_{w,i}$
z_1=(8-5)/2.191=1.369	$Y_{wij,1}=w_{1,2}z_2+w_{1,3}z_3$=(1/2)*0.456+(1/2)*0.456=0.456
z_2=(6-5)/2.191=0.456	$Y_{wij,2}=w_{2,1}z_1+w_{2,3}z_3+w_{2,4}z_4$=(1/3)*1.369+(1/3)*0.456+(1/3)*(-0.913)=0.304
z_3=(6-5)/2.191=0.456	$Y_{wij,3}=w_{3,1}z_1+w_{3,2}z_2+w_{3,4}z_4$=(1/3)*1.369+(1/3)*0.456+(1/3)*(-0.913)=0.304
z_4=(3-5)/2.191= -0.913	$Y_{wij,4}=w_{4,2}z_2+w_{4,3}z_3+w_{4,5}z_5$=(1/3)*0.456+(1/3)*0.456+(1/3)*(-1.369)= -0.152
z_5=(2-5)/2.191= -1.369	$Y_{wij,5}=w_{5,4}z_4$=1*(-0.913)= -0.913

Fuente (Kassel, 2017)

Finalmente se construye el diagrama de dispersión de Morán, como se muestra en la Figura 10:

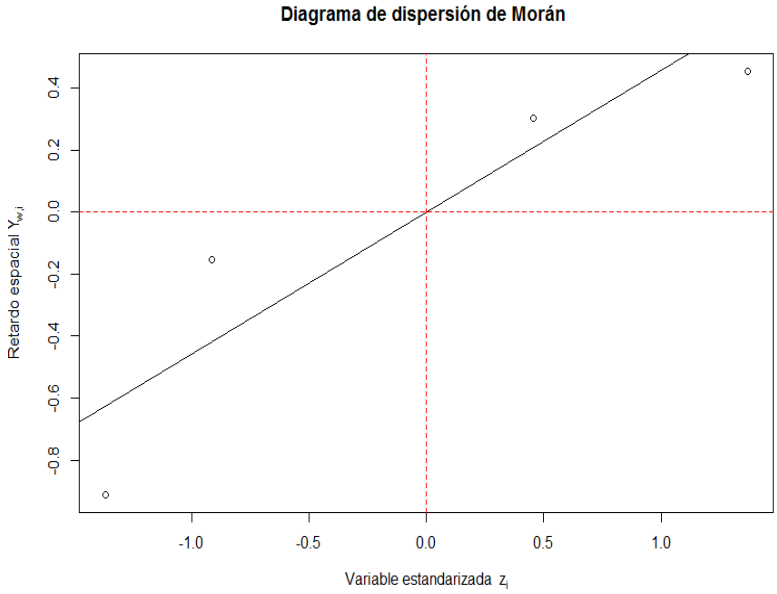

Figura 10. El diagrama de dipersión de Morán de las cinco regiones.

Fuente elaboración propia.

En la figura se puede observar que no hay presencia de outliers espaciales en los cuadrantes II y IV del diagrama de dispersión de Morán.

2.8.2. Diagrama de dispersión de Morán en R

En R (R Core Team, 2016) existe una librería de llamada “GeoXp” que se aplica para análisis de datos exploratorios espaciales interactivos este paquete fue creo en 2015 por los investigadores Yves Aragon, Thibault Laurent, Lauriane Robidou, Anne Ruiz-Gazen, Christine Thomas-Agnan; una función que se emplea poder graficar el diagrama de dispersión de Morán es “Moránplotmap” (R project, 2015).

2.9. Encuesta Nacional de Egresados y Universidad 2014

En el 2014 el Instituto Nacional de Estadística e Informática (INEI) ejecutó la encuesta nacional a egresados universitarios y universidades 2014, con el objetivo de obtener información sobre la inserción laboral y percepción de los servicios educativos recibidos por los egresados de las diferentes carreras de las universidades públicas como privadas que participantes en el Censo Nacional Universitario 2010 (INEI, 2014).

La encuesta se realizó del 20 de octubre hasta el 15 de diciembre de 2014, y se aplicó a 10564 egresados universitarios y también a 131 universidades públicas y privadas que estaban en funcionamiento al 30 de junio de 2014. Esta investigación se ejecutó en los 24 departamentos del país y en la Provincia Constitucional del Callao (INEI, 2014).

III. MATERIALES Y MÉTODOS

3.1. Materiales y equipos

Para la realización del presente trabajo de investigación se tomó en cuenta los siguientes materiales: Hardware, software y servicios.

a) Materiales:

- Fuentes primarias
- Funtes secundarias
- Materiales de escritorio:lapiceros, hojas A4 de 80gr

b) Hardware y software:

- Laptop (Core (TM) i7,2.40 GHz,12 GB RAM)
- Impresora HP Deskjet 2050
- Memoria USB
- R Versión 3.4.1. En este trabajo de investigación se ha utilizado las librerías: GeoXp, ggplot2, ggmap y RgoogleMaps

c) Servicios

- Asesoría Profesional

3.2 Método

3.2.1.Tipo de investigación

El presente trabajo de investigación tiene un enfoque exploratorio dado que se ha recurrido a investigaciones que en la mayoria de los casos no solo se centra en la detección de outliers espaciales, sino que también abarca otros temas relacionados; y que el único fin de haber realizado esa labor fue la de contar con un conocimiento básico para iniciar el estudio con información de manera suficiente y que contribuya con el avance de esta investigación.

3.2.2.Poblacion

La población está comprendida por las coordenadas geográficas (latitud y longitud) de todos los egresados de las universidades públicas y privadas del país donde fueron encuestados el año 2010 durante el II Censo Nacional Universitario (INEI, 2015).

3.2.3.Muestra

Se tomó una muestra de 250 egresados de 122 universidades públicas y privadas del país.

3.3.Hipótesis de la investigación

El variograma Nube es una técnica grafica más adecuada para detectar outlier a diferencia de diagrama de dispersión de moran (O'Leary, Reiners Jr, Xu, & Lemke, Identification and influence of spatio-temporal outliers in urban air quality measurements, 2016).

3.4. Diseño de la investigación

El diseño de la investigación que se empleó en esta tesis es no experimental porque se realizó sin manipulación de la variables se observaron los datos tales como se dando en la ENEUP-2014. La variable considerada para análisis fue “ingreso total” de los egresados universitarios mediante las gráficas descritas en la investigación, los datos tomado fueron transversales; es decir, que fueron tomados en un tiempo único.

3.5. Secuencia metodológica

La secuencia metodológica que se emplea en esta investigación consta de los siguientes pasos: determinar la latitud y longitud de las ubicaciones geográficas de los egresados, aplicación de la función jitter, construcción del variograma nube y el diagrama de dispersión de Morán, identificación de los outliers, análisis y conclusiones. Esta secuencia se expresa en el modelo metodológico que se presenta en la Figura 11.

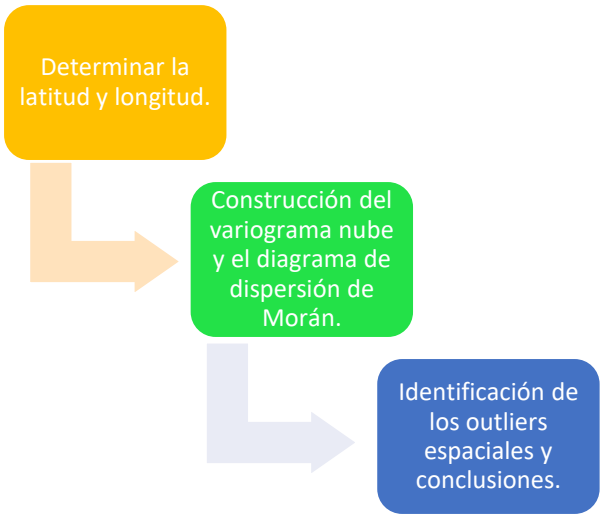

Figura 11. Modelo metodológico para la construcción del variograma nube y del diagrama de dispersión de Morán para el análisis exploratorio de los datos de los egresados de la ENEUP-2014.
Fuente elaboración propia.

- **Determinar la latitud y la longitud** : En primer lugar se debe calcular latitud y longitud para poder convertirlo en datos espaciales, debido que las técnicas descritas trabaja con datos espaciales.
- **Construccion del variograma nube y diagrama de dispersión de Morán:** Obtenido los datos espaciales se calcula los graficos de variograma nube y diagrama de dispersión de Morán, este ultimo antes ejecutar la función de Morán es necesario estandarizar los datos espaciales.
- **Identificacion de los outliers y conclusiones**: Para identificación se considera los criterios de detección de cada gráfica.

IV. RESULTADOS Y DISCUSIONES

4.1.Determinación de la latitud y la longitud

La muestra de la ENEUP-2014 consta de 10564 registros y 518 variables. Mediante la librería "ggmap" del software R-Project (R Core Team, 2016) dicha librería cuenta con la función "getGeoCode" que permite obtener la latitud y longitud de las regiones donde se realizó la encuesta de los egresados universitarios del Perú en 2014.

```
library(ggplot2)
library(ggmap)
library(RgoogleMaps)
fgr<-read.delim("clipboard")
head(fgr,10)
```

	Full	NOMBRECCDD	NOMBRECCPP
1	LIMA LIMA	LIMA	LIMA
2	CHICLAYO LAMBAYEQUE	LAMBAYEQUE	CHICLAYO
3	LIMA LIMA	LIMA	LIMA
4	LIMA LIMA	LIMA	LIMA
5	HUANCAYO JUNIN	JUNIN	HUANCAYO
6	LIMA LIMA	LIMA	LIMA
7	LIMA LIMA	LIMA	LIMA
8	HUANCAYO JUNIN	JUNIN	HUANCAYO
9	SANTA ANCASH	ANCASH	SANTA
10	SAN ROMAN PUNO	PUNO	SAN ROMAN

En la primera columna de la base cuyo nombre FULL se encuentra concatenada el nombre de la ciudad y provincia donde se realizó la encuesta,las dos columnas restantes NOMBRECCDD Y NOMBRECCPP, indica el nombre de la ciudad y la provincia del egresado encuestado respectivamente.

```
attach(fgr)
t(sapply(fgr$Full[1:10], getGeoCode))
```

	lat	lon
[1,]	-12.046374	-77.04279
[2,]	-6.776597	-79.84430
[3,]	-12.046374	-77.04279
[4,]	-12.046374	-77.04279
[5,]	-12.068636	-75.21030
[6,]	-12.046374	-77.04279
[7,]	-12.046374	-77.04279
[8,]	-12.068636	-75.21030
[9,]	-8.987736	-78.61419
[10,]	-15.499836	-70.12965

Luego de obtener los nombres concatenados de la provincia y ciudad donde se realizo la encuesta, se procede a emplear la función "getGeoCode" y se obtendrá la latitud y la longitud.

4.2.Aplicación de las técnicas gráficas del variograma nube y diagrama de dispersión de Morán

La librería "GeoXp" cuenta con dos funciones para ilustrar las técnicas gráficas, con la función "**Moránplotmap**" se empleó para el diagrama de dispersión de Morán y para el variograma nube se empleó la función "**variocloudmap**".

En la muestra para el análisis exploratorio cuenta con tres columnas: la latitud, la longitud y el ingreso total de los egresados proveniente de la base del INEI de la ENEUP-2014. Lo primero que se debe hacer es cargar los paquetes necesario para el adecuado funcionamiento de la librería "GeoXp".

```
library(quantreg)
library(SparseM)
library(sp)
library(Matrix)
library(spdep)
library(rgl)
library(GeoXp)
```

Estos son los paquetes necesario para cargar antes de usar la librería "GeoXp".

```
IngresoEgrs<-read.delim("clipboard")
 head(IngresoEgrs, 5 )
```

	latitude	longitude	Ingreso
1	-12.08202	-76.92823	10500
2	-12.11106	-77.03159	10000
3	-12.09728	-76.99510	10000
4	-12.07876	-77.06554	10000
5	-11.92998	-77.05354	10000

La información del ingreso total está expresado en soles, con la latitud y longitud correspondiente a la región donde ha sido encuestado los egresados.

```
 attach(IngresoEgrs)
 names(IngresoEgrs)
[1] "latitude" "longitude" "Ingreso"

 # IngresoEgrs es un objeto data.frame.
 class(IngresoEgrs)
[1] "data.frame"

 #Dimensiones de data IngresoEgrs
 dim(IngresoEgrs)
[1] 10564    3
```

- El conjunto de datos posee 10,564 registros de los cuales se puede genera 55’’793’766 pares para crear semivarianza o disimilitud del variograma nube, pero dicha cantidad de pares no son adecuado el software estadístico R-project dado que ocasiona un error como este : “`Error: cannot allocate vector of size 421.0 Mb`”. Por la falta de capacidad de procesamiento se opta por tomar una muestra aleatoria de 250 registros para lograr ilustrar el análisis exploratorio de los datos con las técnicas graficas de la investigación.

```
IngresoEgrs1 <- IngresoEgrs[sample(1:nrow(IngresoEgrs),
250, replace=FALSE),]
```

Se tomó la muestra aleatoria de 250 registros con la function “sample”

- Con los 250 datos obtenido es necesario crear los puntos espaciales con la función “**`SpatialPoints`**” para poder graficar el variograma nube mediante la longitud y latitud de las regiones encuestadas.

```
IngresoEgrs1.sp=
SpatialPoints(cbind(IngresoEgrs1$longitude,IngresoEgrs1$latitude))
```

- Obtenidos los puntos espaciales es necesario asignarle un **data.frame** mediante la función “**`SpatialPointsDataFrame`**”.

```
IngresoEgrs1.spdf =SpatialPointsDataFrame(IngresoEgrs1.sp,
IngresoEgrs1)
```

- Se obtuvo los puntos espaciales en una dataframe, para el calculó del variograma nube se empleó la función “**`variocloudmap`**”, donde el primer argumento es el marco de datos de los puntos espaciales y el segundo argumento corresponde al nombre de la variable de interés.

```
library(GeoXp)
 variocloudmap(IngresoEgrs1.spdf, "Ingreso")
```

En la Figura 12 se observa el variograma nube con sus respectivos ejes “x” e “y” representados por la distancia y semivariancia o disimilitud de los pares de datos. Los puntos de color rojo son outliers espaciales depediendo de la distancia entre los pares se pueden considerar outliers espaciales globales o locales, lo que posee una distancia corta pero con grandes valores de disimilitud se le considera outliers locales y los que se aleja de los demás puntos teniendo valores altos disimilitud se le considera outliers globales.

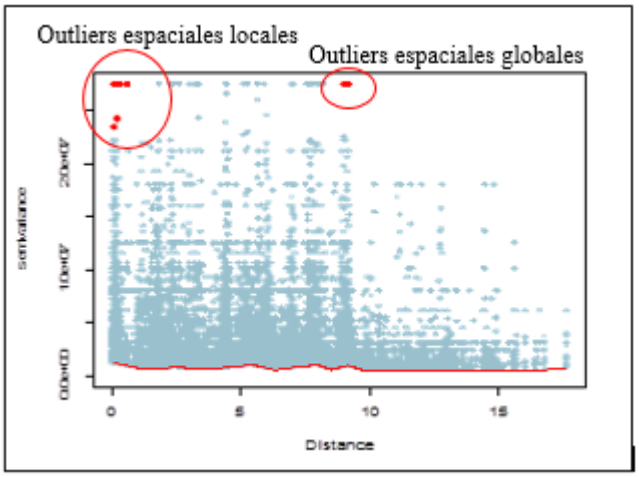

Figura 12.El variograma nube del ingreso de los egresados universitarios peruanos.

Fuente elaboración propia.

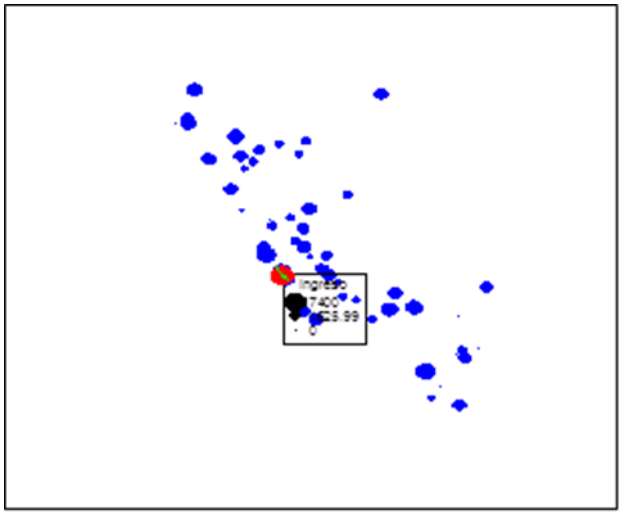

Figura 13.Ploteo de puntos espaciales del mapa del Perú.

Fuente elaboración propia.

Se puede destacar el valor s/.7400 como un valor atípico en alrededor de la región de Lima cuya ubicación geográfica es (0,1625.99).

- Para el caso de diagrama de dispersión de Morán al igual que variograma nube es necesario crear los datos espaciales a partir de la latitud y longitud de la muestra; para diagrama es necesario el pesos espaciales de los datos espaciales. Luego es necesario normalizar los datos para crear el diagrama de dispersión de Morán.

```
W.nb <- knn2nb(knearneigh(IngresoEgrs1.sp, k=4))
```

```
W.listw <- nb2listw(W.nb,style="W")
```

- Se estandariza los puntos espaciales con los respectivos valores de retardo espacial luego de obtener los pesos espaciales, se empleó la función "**Moránplotmap**", donde el primer argumento es el objeto espacial; el segundo argumento corresponde a la variable de interés y el tercero argumento son los pesos espaciales.

```
moranplotmap(IngresoEgrs1.spdf, "Ingreso", W.listw )
```

En la Figura 14 se muestra el diagrama de dispersión de Morán con los cuatros cuadrantes, los puntos rojos son los outliers espaciales ubicados en los cuadrantes IIC y IVC. Según la posición de los cuadrantes se distinguen outliers globales y locales,aquellos con ingresosbajos con retardo altos son outliers locales, en cambio, los que posee ingresos muy altos se considera outliers globales. Se puede observar que la coordenada espacial(0,1625.99) con ingreso de s/7400 es un outlier espacial situado en la región de lima.

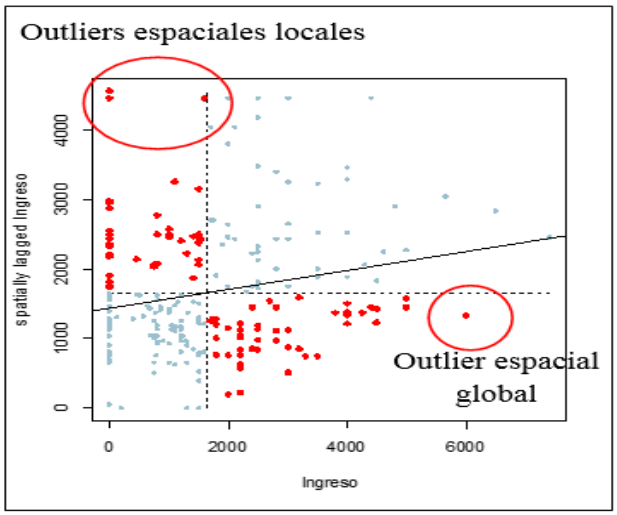

Figura 14. Diagrama de dispersión de Moran del ingreso de los egresados univerisitarios peruanos.

Fuente elaboración propia

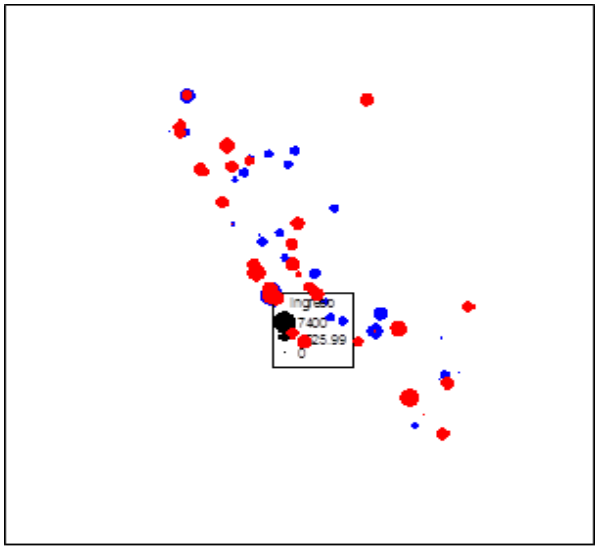

Figura 15.Ploteo de los puntos espaciales del mapa del Perú.

Fuente elaboración propia.

El valor s/.7400 es considerado como un outlier espacial ubicado alrededor de la región de Lima, cuya ubicación geográfica es (0,1625.9).

- Adicionalmente se puede obtener el índice Morán para saber si existe una relación negativa o positiva entre las zonas vecinas de los puntos geográficos de los egresados universitarios; es decir, si los egresados de una región tiene similar ingreso total o no.

```
# Indice de moran
moran.test(IngresoEgrs1.spdf$Ingreso,W.listw)
```

```
        Moran I test under randomisation

data:  IngresoEgrs1.spdf$Ingreso  |
weights: W.listw

Moran I statistic standard deviate = 3.4909, p-value = 0.0002407
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.137181613      -0.004016064       0.001636022
```

Según el índice de Morán se obtuvo un valor positivo, esto decir, que los egresados dentro de una zona geográfica tiene ingresos similares.

Si existeria dificultad para visualizar de los datos se puede optar por el siguiente paso:

Paso opcional : Aplicación de la función jitter.

Puede presentarse que existan datos que están muy cercanos lo que conlleva una visualización confusa de los mismo es necesario utilizar la función llamada “jitter” de la librería ggplot2 para añadir un pequeño ruido y de esta manera separar los puntos superpuestos. Este ruido es un valor aleatorio de la distribución uniforme en [-a,a] donde a es muy pequeño y mayor que cero. A continuación se presenta un ejemplo de algunos datos (donde la variable predictora es discreta y el resultado es continua), para observar los problemas con el trazado de estos tipos de datos utilizando los valores predeterminados de R, y luego observar la función “jitter” para dibujar una mejor dispersión.

```
set.seed(1)
x <- sample(1:10, 200, TRUE)
y <- 3 * x + rnorm(200, 0, 5)
plot(y ~ x, pch = 15)
```

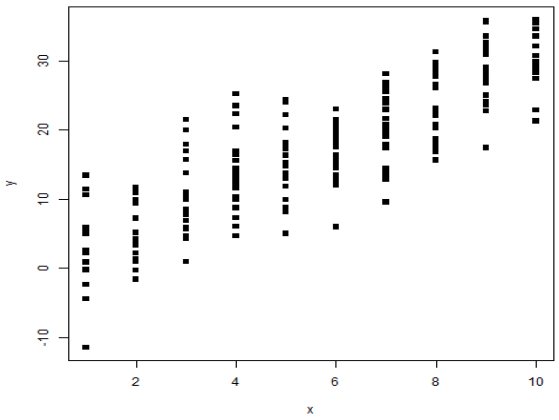

Figura12. Datos superpuestos.
Fuente: (thomasleeper, 2016)

Como se puede visualizar en la Figura 12 existen puntos superpuestos a otros y por tanto no se pueden visualizar correctamente los puntos. Luego se aplica la función jitter a los datos y se observa en la Figura 13 la nube de puntos con mayor claridad. La instrucción R para obtener la gráfica es:

```
plot(y ~ jitter(x, 1), pch = 15)
```

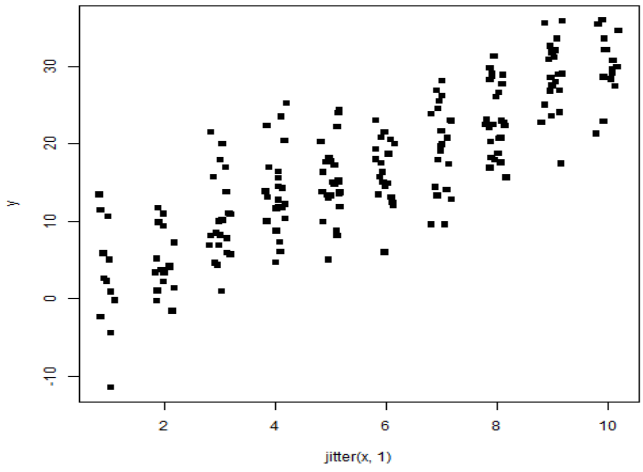

Figura 13. Datos claramente distinguible por la función jitter.
Fuente: (thomasleeper, 2016).

Si se aumenta el valor del parámetro en la distribución uniforme se visualiza con mayor claridad los puntos como se observa en la Figura 14. La instrucción R para obtener la gráfica es:

```
plot(y ~ jitter(x, 2), pch = 15)
```

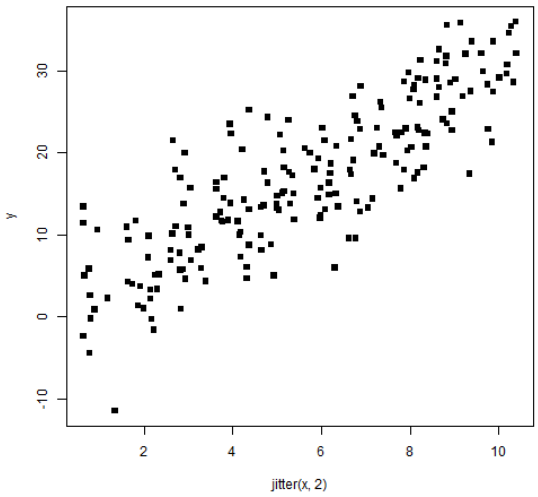

Figura 14.Datos con una claridad de 2 usando la función jitter.
Fuente: (thomasleeper, 2016).

4.3.Identificación de outliers

Respecto a los resultados obtenidos en las dos gráficas tanto el variograma nube y el diagrama de dispersión de Morán en el r-project tiene en común un punto geografico ubicado (0,1625.99) que significa un ingreso de s/.7400, donde se encontraron alrededor de 7 puntos outliers espaciales en el variograma nube con dicha coordenada y ingreso, de los cuales 5 de ellos son outliers espaciales locales debido a la distancia corta entre los pares de datos pero con valores altos disimilitudes y lo restante, 2 puntos son outliers espaciales globales debido a que se aleja considerablemente de los otros pares de datos. En el diagrama de dispersión de Morán se presentó 4 outliers espaciales debido que se encuentra en el II cuadrante donde se identifico 3 outliers espaciales locales debido a la distancia corta de los pares pero con valores de la variable estandarizada, y el cuadrante IV cuadrante se encontra 1 solo outliers es considerado como outlier espacial global. Respecto a las dos técnicas de detección de outliers espaciales el variograma nube fue la más sensible para la identificación.

V. CONCLUSIONES

La detección de outliers espaciales es un tema importante en el campo de la minería de datos espaciales. El objetivo de identificar tales outliers es descubrir el conocimiento oculto pero potencialmente útil que tiene dichos valores, por ejemplo, si un egresado de la ciudad de lima gana 7400 soles en su ingreso total, al igual que un egresado de la ciudad Iquitos, esto se puede dar debido que sus profesiones son muy rentables en dichas ciudades o puede que exista otros factores a dichas anomalías respecto a la región geográfica.

La conclusiones de la investigación son las siguientes :

- El diagrama de dispersión de Morán como en el variograma nube brindaron un valor en común de s/.7400 en el ingreso total de egresados situado aproximadamente en la región de Lima considerado como outlier espacial respecto a entorno geográfico de coordenada (0,1625.9); pero el variograma nube fue más sensible el variograma nube a encontrar más valores outliers en dicho punto que el diagrama de dispersión de Morán.
- Respecto a la metodología del diagrama de dispersión de Morán y del variograma nube no es adecuado aplicarlo cuando se presenta más de 250 registros debido a la dificultad de visualizar y detectar los outliers espaciales.
- Mediante las funciones “Moránplotmap” y “variocloudmap” de la librería “GeoXp” se logró ilustrar la metodología de las graficas, pero teniendo una limitación en procesamiento de grandes datos, porque se genera una gran cantidad de pares que ocasiona que el software no responda adecuadamente.

VI. RECOMENDACIONES

Las recomendaciones para emplear las dos gráficas exploratorias para identificación de outliers espaciales son:

- Las técnicas graficas del variograma nube y diagrama de dispersión de Morán presentan dificultad tanto como metodología junto con el software R-project para el procesamiento de más de 250 datos debido que se genera muchos pares de datos que no es posible procesarlo y su visualiza se dificultad considerablemente.
- Si se da el caso de una investigación con este mismo enfoque y presenta una gran número de datos, lo recomendable es tomar una o varias muestras aleatorias que no supere la cantidad de 250 datos.
- En caso que se tenga una muestra de 250 datos y no se pueda visualizar los outlieres espaciales se puede emplear la función “jitter”del software R-project, esto no se puede para muestra mayor 250.
- Para poder realizar las dos técnicas gráficas de esta investigación se debe emplear para una sola variable, porque si se usa dos o más variables se dificultaría la visualización y identificación de los posibles outliers espaciales.

VII. REFERENCIAS BIBLIOGRAFICAS

- Ben-Gal, I. (2005). Kluwer Academic Publishers. Obtenido de OUTLIER DETECTION:http://www.eng.tau.ac.il/~bengal/outlier.pdf

- O'Leary, B., Reiners Jr, J., Xu, X., & D. Lemke, L. (2016). *Identification and influence of spatio-temporal outliers in urban air.* Obtenido de Science of the Total Environment 573 (2016) 55–65: http://ac.els-cdn.com/S0048969716317235/1-s2.0-S0048969716317235-main.pdf?_tid=2504d04e-8c11-11e6-9f20-00000aacb35d&acdnat=1475791774_van der Loo, M. (20 de 01 de 2016). *R-project,CRAN*. Obtenido de https://cran.r-project.org/web/packages/extremevalues/extremevalues.pdf
- Cao, L., Liu, X., Wang, Z., & Zhang, Z. (Diciembre de 2013). *The Spatial Outlier Mining Algorithm based on the KNN Graph.* Obtenido de JOURNAL OF SOFTWARE: http://www.jsoftware.us/vol8/jsw0812-24.pdf
- Chang-Tien, L., Yufeng, K., Hongjun, W., & Dechang, C. (s.f.). *MapView - A Visualization Tool for Spatial Outlier Detection.* Obtenido de https://pdfs.semanticscholar.org/1429/ed3000fe7e350562853c981847f12e6a14b9.pdf
- Chawla, S., & Sun, P. (2005). *SLOM: a new measure for local spatial outliers.* Obtenido de Knowledge and Information Systems: https://pdfs.semanticscholar.org/996a/8247c9f6710e07e978fc83dc828fb2fe222c.pdf
- Chen, D., Tien Lu, C., Kou, Y., & Chen, F. (2007). On Detecting Spatial Outliers. *Geoinformatica*, 455–475.
- Chilès, J.-P., & Delfiner, P. (2009). Geostatistics: Modeling Spatial Uncertainty. Wiley Interscience publication. Obtenido de https://books.google.com.pe/books?id=tZl07WdjYHgC&pg=PA35&lpg=PA35&dq=variogram+cloud+bastin%2Bexample&source=bl&ots=kKGFVS9JOh&sig=9fKyvH2iDNg3l1ttg-

qkJUtoYa8&hl=es&sa=X&ved=0ahUKEwiZ_eSo4cvPAhXMsh4KHVnOA8gQ6AEIGjAA#v=onepage&q=variogram%20cloud%20bastin

- Chue Gallardo, J. (2017). *Análisis del Empleo y de los Ingresos de los Egresados Universitarios del Perú utilizando la Ciencia de Datos Espacial.*. Tesis doctorado: Universidad Nacional del Santa.
- Haslett, J., Bradley, R., Craig, P., & Wills, G. (1991). *BioMedware*. Obtenido de https://www.biomedware.com/files/documentation/spacestat/interface/Views/Variogram_Cloud.htm
- Hernandez Sampieri, R., Fernandez Collado, C., & Baptista Lucio, P. (2006). Metodologia de la investigacion. Mexico: McGRAWHILLIINTERAMERICMA EDITORES, SA.
- Hernández-Hernández, V. (2012). *Análisis geoespacial de las elecciones presidenciales en México.* Obtenido de Researchgate: https://www.researchgate.net/publication/272030241_Análisis_geoespacial_de_las_elecciones_presidenciales_en_Mexico_2012?_sg=GQ1b32mqXopn3FCExchN3jx4zu5dAl9Tj_MueuolldxNV4pBbpxrxKhUdiAF1iP8fcFPy9_2FfcBsk6Fi7l63g
- INEI. (24 de 10 de 2014). Obtenido de https://www.inei.gob.pe/prensa/noticias/inei-ejecuta-encuesta-a-egresados-universitarios-y-universidades-7820/
- INEI. (Nov de 2015). Obtenido de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1298/Libro.pdf
- J. Hodge, V., & Austin, J. (2004). *A Survey of Outlier Detection Methodologies.* Obtenido de White Rose Consortium ePrints Repository : http://eprints.whiterose.ac.uk/767/1/hodgevj4.pdf
- Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. wiley. Obtenido de https://books.google.com.pe/books?id=4IyrNiNvx0gC&pg=PA360&dq=variogram+cloud+example&hl=es&sa=X&ved=0ahUKEwjSo8KbocnPAhXRMx4KHRnGACoQ6AEIPjAD#v=onepage&q=variogram%20cloud%20example&f=false
- Kapageridis, I. (03 de 2015). *Variable Lag Variography Using k-means Clustering.* Obtenido de Computers & Geosciences 85 ·:

https://www.researchgate.net/publication/274896521_Variable_Lag_Variography_Using_k-means_Clustering

- Kou, Y., Tien Lu, C., & Chen, D. (2006). Spatial Weighted Outlier Detection. *SDM*, 614-618.
- Kriegel, H.-P., Kröger, P., & Zimek, A. (17 de Setiembre de 2016). *Outlier Detection Techniques.* Obtenido de https://www.siam.org/meetings/sdm10/tutorial3.pdf
- Laurent, T., Ruiz-Gazen, A., & Thomas-Agnan, C. (Abril de 2014). *Journal of Statistical Software*. Obtenido de Volume 47, Issue 2: https://www.jstatsoft.org/article/view/v047i02/v47i02.pdf
- Maimon, O., & Rokach, L. (2006). Data Mining and Knowledge Discovery Handbook. En *https://books.google.com.pe/books?id=S-XvEQWABeUC&pg=PA839&lpg=PA839&dq=criteria+to+distinguish+spatial+outliers+variogram+cloud&source=bl&ots=LCUo8jyz2P&sig=8OLIwyB5pZB0jw8XVUMAumrm7f8&hl=es&sa=X&ved=0ahUKEwiniJG6r7bPAhWI0h4KHRWvABUQ6AEIGjAA#v=onepage&q=* (págs. 838-839). Springer Science & Business.
- Oregon, D. o.-U. (2006). *Plots of spatial statistics (Variograms)*. Obtenido de http://geog.uoregon.edu/bartlein/old_courses/geog414s05/topics/variograms.htm
- Shekhar, S., & Xiong, H. (2007). Encyclopedia of GIS. USA: Springer Science & Business Media.
- Songwon, S. (26 de April de 2006). *A Review and Comparison of Methods for Detecting Outliers.* Obtenido de Thesis of University of Pittsburgh: http://d-scholarship.pitt.edu/7948/1/Seo.pdf
- Songwon Seo, M. (2006). Obtenido de A Review and Comparison of Methods for Detecting Outliers: http://d-scholarship.pitt.edu/7948/1/Seo.pdf
- Thomasleeper. (20 de 10 de 2016). Obtenido de http://thomasleeper.com/Rcourse/Tutorials/jitter.html
- Wackernagel, H. (2013). Multivariate Geostatistics:An Introduction with Applications. Springer Science & Business Media.
- *zevross*. (22 de 10 de 2016). Obtenido de Technical Tidbits From Spatial Analysis & Data Science: http://zevross.com/blog/2014/05/05/unhide-hidden-data-using-jitter-in-the-r-package-ggplot2/