Anales Científicos, 77 (2): 329-337 (2016)
ISSN 0255-0407 (Versión impresa)
ISSN 2519-7398 (Versión electrónica)
© Universidad Nacional Agraria La Molina, Lima - Perú
DOI: http://dx.doi.org/10.21704/ac.v77i2.810

Presentado:05/12/2016
Aceptado: 27/12/2016

ANÁLISIS EXPLORATORIO ESPACIAL DEL INGRESO DE LOS EGRESADOS UNIVERSITARIOS DEL PERÚ

EXPLORATORY SPATIAL ANALYSIS OF THE INCOME OF UNIVERSITY GRADUATES OF PERU

[1]Jorge Chue Gallardo

Resumen

La Encuesta Nacional a Egresados Universitarios del Perú 2014 registró información de 10564 egresados y 518 variables. El informe publicado por el INEI se limita a tablas, gráficos de barras y gráficos circulares. Hay un vacío de conocimiento debido a la ausencia de análisis exploratorio y modelamiento de los datos desde la perspectiva de la estadística espacial. En este trabajo, se realiza lo correspondiente al análisis exploratorio de las coordenadas geográficas y la ubicación espacial de los ingresos de los egresados aplicando limpieza de datos, cálculo de estadísticas descriptivas como centro promedio, centroides, mediana, desviación estándar, visualización, intensidad de puntos, autocorrelación espacial con el índice de Morán, y la función K de Ripley para identificar patrones de puntos. El software estadístico R fue utilizado para los cálculos y gráficas respectivas.

Palabras claves: Patrón espacial, visualización, autocorrelación espacial, índice de Morán, función K de Ripley.

Abstract

The National Survey of University Graduates of Peru 2014 recorded information on 10564 graduates and 518 variables. The report published by INEI is limited to tables, bar graphs and pie charts. There is a knowledge gap due to the absence of exploratory analysis and modeling of data from the perspective of spatial statistics. In this work, we perform the exploratory analysis of the geographical coordinates and the spatial location of the income of graduates applying data cleaning, calculation of descriptive statistics such as average center, centroids, median, standard deviation, visualization, intensity of points, Moran index spatial autocorrelation, and Ripley's K function to identify point patterns. The statistical software R was used for the respective calculations and graphs.

Palabras claves: Spatial pattern, visualization, spatial autocorrelation, Moran index, Ripley function K.

1. **Introducción**

El ingreso según la (International Labour Organization, 2015) es el precio de la fuerza de trabajo o costo de la remuneración del trabajo equivalente a la cantidad de dinero en efectivo que las personas reciben durante un periodo de tiempo por las actividades económicas que realizan. En el Perú, de acuerdo al tipo de renta, los trabajadores se clasifican en dependientes e independientes. Un trabajador es independiente si percibe rentas de cuarta categoría por el ejercicio individual e independiente de su profesión, arte, ciencia u oficio. En el caso de los trabajadores dependientes las rentas son de cuarta categoría (SUNAT, 2015). Asimismo, en el Decreto Supremo Nº166-2013-EF se indica que un trabajador independiente es el sujeto que percibe ingresos de cuarta y/o quinta categoría (Ministerio de Trabajo, 2013). Para el (INEI, 2015), el ingreso es un derecho constitucional que proporciona bienestar material y espiritual. Las variables relacionadas al empleo que se analizan en la ENEU-2014 son: ingreso promedio mensual clasificado por género, grupo étnico-racial, tipo de universidad, grupo ocupacional, categoría ocupacional, actividad económica, tamaño de empresa, tipo de contrato y campo de educación.

El ingreso de los egresados universitarios peruanos han sido motivos de estudio desde diferentes perspectivas. En (Rodríguez & Montoro, 2013) se señala que habría un importante subempleo entre las personas con educación superior. En cuanto a los ingresos, (Yamada, 2007) sostiene que el retorno de la inversión en una educación privada es muy superior a la pública y que esta diferencia podría aumentar por el deterioro de la educación pública en los últimos años. Los investigadores (Calónico & Ñopo, 2007) también afirman que los retornos son mayores para aquellos que asistieron a instituciones privadas en comparación a las públicas. En la investigación realizada por (Lavado, Martínez, & Yamada, 2014) se afirma que cuatro de cada diez profesionales universitarios

Departamento Académico de Estadística e Informática, Facultad de Economía y Planificación, Universidad Nacional Agraria La Molina, Apartado postal 12-056 - La Molina, Lima, Perú. Email: jchue@lamolina.edu.pe

al 2012 son sobre-educados realizando actividades no profesionales y sub remuneradas.

A nivel internacional, las investigaciones acerca del ingreso de los egresados universitarios son extensas y variadas. En un estudio realizado en Chipre por (Eliophotou, Pashourtidou, Polycarpou, & Pashardes, 2012) se encontró que los estudios de postgrado, el lugar de residencia y la profesión estudiada eran los factores que afectan la probabilidad de empleo de los recién graduados. En (Corbett & Hill, 2012) se señala que en Estados Unidos de Norteamérica en el año 2009 las mujeres ganaban el 82% de lo que ganaban los hombres de similares características laborales y académicas. En un estudio acerca de los ingresos de los egresados universitarios en Canadá realizado por (Oreopoulos & Petronijevic, 2013) se menciona que antes de tomar la decisión de estudiar en una institución superior los estudiantes deben considerar lo siguiente: institución educativa, especialidad a estudiar, futuros ingresos de la carrera, posibilidades de completar la carrera, costos y valor del préstamo estudiantil.

Las investigaciones antes mencionadas, a nivel nacional e internacional, son una muestra pequeña de la gran cantidad de trabajos realizados acerca del ingreso de los egresados universitarios. En el caso de las investigaciones a nivel nacional las técnicas estadísticas utilizadas no consideran las dimensiones del tiempo y del espacio, es decir, ignoran "dónde" y "cuándo" ocurrieron los eventos. Este vacío es una oportunidad para mejorar el conocimiento del ingreso de los egresados universitarios del Perú utilizando la Ciencia de Datos Espacial (CDE) porque en opinión de (Cressie & Wikle, 2011) la no inclusión del tiempo y lugar en el análisis impide estudiar las relaciones de causa-efecto y el efecto del lugar en los datos. Más aún, si se sabe que en la naturaleza las variables físicas y biológicas presentan heterogeneidad espacial que no puede dejarse de lado en el proceso de análisis como lo sostiene (Moral García, 2004).

La Ciencia de Datos Espacial es definida por (Shi, 2015) como la ciencia del descubrimiento del conocimiento y explicación espacial de las actividades humanas y naturales basadas en Big Data espacial; es decir, conjuntos de datos espaciales que son grandes, complejos y diversos (Gartner, 2013). La CDE abarca conocimientos de ciencia de datos, ciencia de información geográfica, ciencias de computación, estadística espacial, aprendizaje automático y data mining espacial. En un programa de maestría de investigadores en Ciencia de Datos Espacial & Visualización se enseñan cursos de técnicas innovadoras de análisis de datos, minería de datos, modelamiento y visualización (CASA, 2015).

Algunas investigaciones utilizando datos espaciales han sido realizadas en las ciencias ambientales (Moral García, 2004), sociología (Moreno, 2011), biología marina (Quiroz Cornejo, 2011), econometría (Herrera Gómez, Cid, & Paz, 2012), medicina (Cuartas, Ariza, Pachajoa, & Méndez, 2011), (Lisset & Fuentes, 2014), (Hernández, Rodriguez, & Antón, 2013), epidemiología (Borroto & Martínez-Piedra, 2000), minería de datos (Shekhar, Zhang, Huang, & Raju Vatsavai, 2009) y ecología de plantas (Law, y otros, 2009).

En el Perú, el 2 de noviembre del 2007 fue creado con R.M. Nº 325-2007-PCM el Comité Coordinador Permanente de la Infraestructura de Datos Espaciales del Perú (CCIDEP) con el objetivo de promover y coordinar el desarrollo, intercambio y el uso de datos y servicios de información espacial entre todos los niveles de gobierno, sector privado, organizaciones sin fines de lucro, instituciones académicas y de investigación (ONGEI, 2010). En (Moreno, 2011) se analizó la distribución y concentración espacial de la población peruana en los Estados Unidos de Norteamérica concluyéndose que los peruanos comparten el espacio con hispanos nacidos fuera de los Estados Unidos. En la tesis de (Curatola Fernández, 2009) se estudiaron los patrones de la distribución espacial del árbol Triplaris Americana en Tambopata, Perú.

La importancia del análisis exploratorio espacial se sustenta en que los beneficios que se adquieran con su utilización podrá reducir los errores en la toma de decisiones y por tanto los costos de los mismos. Los procesos que se puedan desarrollar e implementar con esta metodología permitirán automatizar y desarrollar sistemas de información que respondan al más breve plazo y en línea a los requerimientos de los responsables de la toma de decisiones. Por ejemplo, con los datos de la Encuesta Nacional a Egresados Universitarios (ENEU-2014) se pueden identificar patrones de los niveles de empleo de los egresados universitarios en las diferentes regiones del Perú. Estos patrones serán información valiosa para las políticas de desarrollo social, económico, industrial entre otras, que el gobierno central e inversionistas decidan aplicar.

En este trabajo de investigación, ante el vacío de conocimiento del comportamiento espacial del ingreso de los egresados universitarios del Perú, se plantea la utilización del análisis exploratorio espacial para cubrir parcialmente este vacío.

El alcance del presente proyecto de tesis es el análisis exploratorio espacial de los datos de los egresados correspondientes al ingreso total mensual monetario (ITMM). Se eligió ITMM a que es la suma de los ingresos que reciben los egresados por fuente de trabajo principal, fuente de trabajo secundario, horas extras, movilidad, comisiones, bonificaciones entre otros ingresos. Esta información se encuentra registrada en la ENEU-2014 (INEI, 2014).

El objetivo de la presente investigación es realizar el análisis exploratorio espacial del ITMM de los egresados universitarios del Perú registrados en la encuesta nacional de egresados universitarios 2014 mediante estadísticas espaciales descriptivas, cálculo de la autocorrelación del índice de Morán (Moran, 1950), la intensidad espacial, la función K de Ripley (Ripley, 2001).

2. Materiales y métodos.

La población objetivo estuvo formada por todos los egresados de las universidades públicas y privadas (según el II Censo Nacional Universitario del año 2010) que terminaron antes del 15/12/2014. La población muestreada estuvo formada por todos los egresados residentes en el país de las universidades públicas y privadas registradas en el Censo Nacional Universitario 2010 (INEI, 2014). La ausencia de un padrón de egresados de las universidades del país se resolvió con el marco muestral del II CENAUN (Censo Nacional Universitario) acotado de acuerdo con los siguientes criterios: egresados con al menos un año de egreso al momento de la encuesta, egresados con menos de 3 años de atraso de estudios en el curso de su carrera y egresados que en su último año de estudios no tengan más de 25, 26 y 27 años según la extensión de su carrera (5, 6 y 7 años respectivamente). El marco muestral para la encuesta de egresados estuvo compuesto por 213,370 alumnos de 92 universidades censados por el CENAUN 2010 (INEI, 2014). La unidad de análisis fue el egresado de cada universidad con al menos 40 alumnos egresados entre todas las carreras ofrecidas. La muestra fue probabilística, estratificada, de lista, de una etapa e independiente en cada carrera universitaria. El tamaño de la muestra fue de 10564 egresados. El periodo de aplicación de la encuesta fue del 20/10/2014 al 15/12/2014. La cobertura de la encuesta fue a nivel nacional incluyendo la Provincia Constitucional del Callao. Los datos se encuentran disponibles en http://iinei.inei.gob.pe/microdatos/ (INEI, 2014)

Limpieza de datos. La limpieza de los datos de la ENEU-2014 se realizó aplicando los siguientes criterios sugeridos por (Silipo, 2016): remover las variables que tienen más del 40% de datos faltantes, remover las variables con coeficiente de variabilidad menor a 10% porque los valores son mu homogéneos y contribuyen muy poco al análisis, remover las variables que tienen alta correlación considerándose sólo una de ellas en el análisis. Se utilizaron las formulas siguientes:

Coeficiente de correlación de Pearson para variables

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Chi-cuadrado de Pearson para independencia para variables cualitativas.

$$\chi^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Remover los outliers con la regla de Tukey de aplicar el intervalo Q1 – 1,5IQR y Q3 + 1,5IQR para los outliers leves y el intervalo Q1 – 3IQR and Q3 + 3IQR para los outliers extremos (Dawson, 2011). La estimación de valores perdidos o imputación múltiple para la ENEU-2014 será considerada en trabajos futuros.

Análisis exploratorio de datos. Este conjunto de técnicas fue utilizado para identificar estructuras, extraer variables significativas, detectar outliers y detectar patrones de asociación espacial. Las técnicas a utilizarse son:

a) Cálculo de estadísticas descriptivas de las coordenadas de los datos de la variables ITMM utilizando el software gratuito R (R Core Team, 2016). Las estadísticas descriptivas a ser calculadas son:

- Centro promedio de las coordenadas de las variables ITMM. Las fórmulas para calcular el centro promedio son:

$$\bar{x} = \frac{\sum_{i=1}^{n}x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n}y_i}{n} \quad (3)$$

- Centroide es el equivalente a un centro promedio de un polígono. Es el centro de gravedad. En esta tesis se calculará el centroide de los datos de los diferentes departamentos del Perú incluidos en la muestra.
- Centro promedio ponderado que se obtiene de los centroides de los departamentos.

$$\bar{x} = \frac{\sum_{i=1}^{n}w_i x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n}w_i y_i}{n} \quad (4)$$

- Centro mediana es el punto que minimiza la suma de las distancias entre sí misma y todos los otros puntos. No tiene una expresión analítica única y es obtenida por aproximación.
- Desviación estándar de las distancias que representa la dispersión de los puntos con respecto al centro promedio. Su fórmula es:

$$S_D = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_c)^2 + \sum_{i=1}^{n}(Y_i - Y_c)^2}{N}} \quad (5)$$

b) Visualización espacial de los datos de ITMM. Para esta visualización se utilizarán las librerías geoXp, spatial, spatstat, sp, maptools, maps, mapdata, rgdal, Rgooglemaps, entre otros relacionados a datos espaciales del software R.

c) Medición de la intensidad de puntos (número esperado de puntos por unidad de área). Para (Baddeley, 2010), la intensidad espacial λ(x) mide la distribución de la variable ITMM en la región bajo estudio. La intensidad espacial puede ser constante (homogéneo) o asumir diferentes valores de una ubicación a otra (no homogéneo o heterogéneo). En el presente trabajo

sólo se considera el caso homogéneo. La fórmula para calcular la intensidad cuando el proceso puntual X es homogéneo es $\hat{\lambda} = \frac{n(x)}{área(W)}$ donde es el estimador máximo verosímil de λ(x).

d). Prueba de autocorrelación espacial de la variable ITMM. El índice de Morán es una extensión del coeficiente de correlación de Pearson a una variable. La intuición del índice de Morán es que observaciones cercanas son probablemente similares a diferencia de aquellas que se encuentran más lejos (Paradis, 2014). La fórmula para calcular el índice de Morán es:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

Donde y_i es la i-ésima observación de ITMM e su promedio. Los valores w_{ij} es el peso espacial de la vinculación o enlace entre i y j. Utilizar el promedio como punto de referencia es equivalente a afirmar que el modelo correcto es constante y que cualquier patrón existente es originado por las relaciones espaciales expresadas por los pesos espaciales w_{ij} (Bivand, Pebesma, & Gómez-Rubio, 2008). El índice de Morán permite probar la hipótesis nula de que la autocorrelación espacial es cero versus la hipótesis alternante que es diferente de cero.

e) Función K. La función K fue construida especialmente para patrones de puntos espaciales (Ripley, 2001). Esta función K permite capturar la dependencia espacial entre los diferentes puntos muestrales del área bajo estudio. Según (Tarpey, 2012), la estimación de la función K es obtenida con la siguiente formula

$$\hat{K}(h) = \frac{\text{Promedio de los eventos dentro de una distancia h uno del otro}}{\hat{\lambda}} \quad (7)$$

Las librerías GeoXp, ouliers, extremevalues, aspace, spatstat del software R (R Core Team, 2016) fueron utilizadas para calcular lo mencionado en los párrafos anteriores.

3. Resultados y discusión.

La limpieza de datos fue realizada según las pautas indicadas en la sección anterior obteniéndose finalmente 128 variables. En esta investigación se analizarán inicialmente los ingresos posteriormente se incorporarán las demás variables. El ITMM de los egresados por empleo (ocupados/desocupados) se presenta en la tabla 1. La figura 1 presenta el histograma de los ITMM y la distribución de los egresados por empleo.

Tabla 1. Estadísticas descriptivas del ITMM de los egresados universitarios 2014

	Media	Desv. Estándar	Mediana	Mínimo	Máximo
Ocupados	1764,3	1264,4	1500	0	10500
Desocupados	299	754,5	0	0	7000

Las funciones chisq.out.test y outlier de las librerías outliers (Komsta, 2015) y extremevalues (Van Der Loo, 2016) del software R (R Core Team, 2016) fueron aplicadas a la variable ITMM encontrándose que el ingreso con valor 10500 era un outlier. Este valor se registró en el distrito de Los Olivos, provincia de Lima y departamento de Lima. Esto fue confirmado al utilizar la función outlierplot para graficar los valores observados versus los valores pronosticados con una distribución exponencial como parámetro (ver figura 2). Este valor outlier de 10500 no fue eliminado porque el objetivo no es el valor mismo sino dónde está ubicado.

Las coordenadas geográficas fueron obtenidas utilizando el Batch Geocoding desarrollado por (Zwiefelhofer, 2016). La entrada de datos a ser geocodificados fue con la siguiente estructura "distrito, provincia, departamento, país" obteniendo como resultado la latitud y longitud de residencia del egresado.

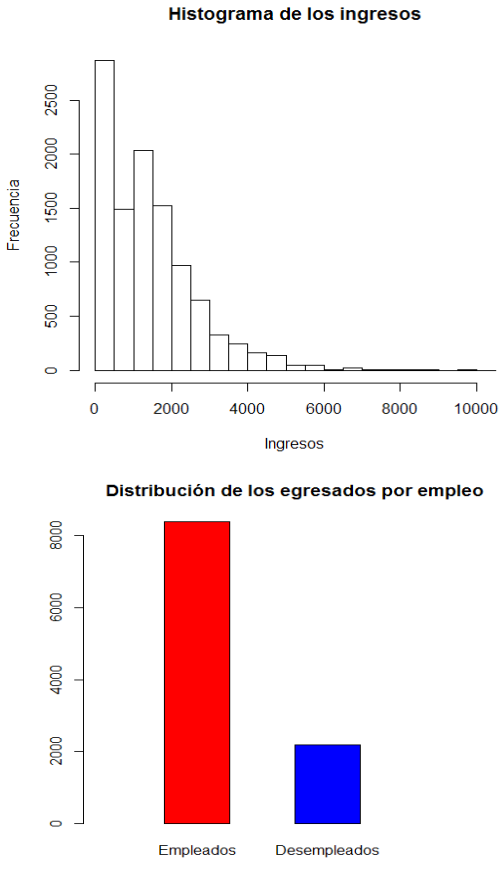

Figura 1. Distribución de los ITMM y de los empleos de los egresados universitarios 2014

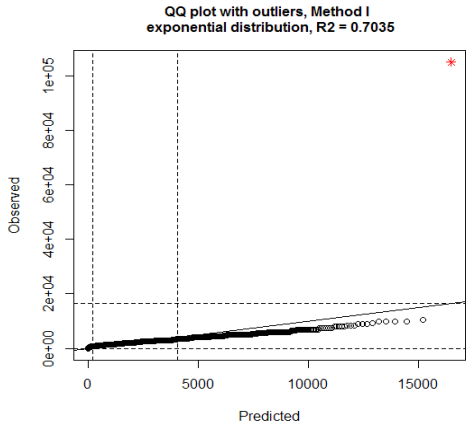

Figura 2. Gráfica de los ITMM observados versus los pronosticados de los egresados universitarios 2014. Nótese el outlier en el extremo superior derecho con un asterisco rojo.

Un análisis específico de los valores de ITMM de los egresados en el intervalo de 3100 a 4200 soles utilizando la librería GeoXp (Laurent, Ruiz-Gazen, & Thomas-Agnan, 2012) de R reveló la ubicación de los egresados que se muestran en la figura 3 conjuntamente con el histograma y los valores de ITMM.

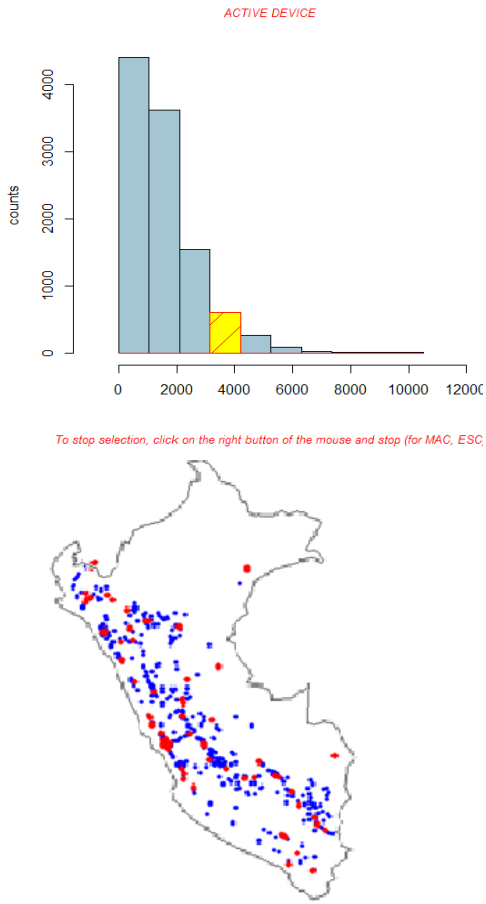

Figura 3. Histograma y gráfica de los valores de ITMM

Nótese en la figura 3 que los ingresos en el intervalo de 3100 a 4200 soles (puntos rojos) se encuentran distribuidos con mayor incidencia en el centro y norte de la costa, en todo el centro y bastante disperso en la selva. En el sur del país hay muy pocos egresados con estos ingresos.

La figura 4 presenta la visualización de los 10564 egresados en el territorio nacional del Perú con el centro promedio (burbuja de color azul) y coordenadas geográficas con longitud= -76,01924 y latitud= -11,27021.

En la figura 4 se observa que hay ausencia de egresados en el interior de Piura, en las alturas de Arequipa y al sur de Ica. En la selva es notaria la ausencia de egresados excepto en San Martín y Amazonas.

Figura 4. Visualización de los egresados en el territorio nacional y el centro promedio

Los valores de los centroides de cada departamento se presentan en la tabla 2 y su representación gráfica en la figura 5. Los centroides de la figura 5 indican la separación que tienen los egresados por departamento. Nótese la separación del centroide de Ica con respecto a Arequipa y éste a su vez con los centroides de Madre de Dios y Cuzco. Este resultado serviría para incentivar la presencia de egresados en estos lugares que ayuden a su desarrollo económico y social.

Los valores de la media y la mediana fueron calculados y graficados con la función driftmap() de GeoXp que se presenta en la figura 6. La primera grafica corresponde a un diagrama de dispersión con las coordenadas de las ubicaciones. La segunda y tercera gráfica corresponden a la media y la mediana de los valores de ITMM para las filas y las columnas de las coordenadas. La cuarta gráfica corresponde a la leyenda de los datos y una brújula indicando la ubicación de los datos. Estos gráficos nos indican que no hay diferencias entre los valores de la media y la mediana en forma vertical y horizontal. También se observa claramente que los ingresos en Lima son los más altos mientras que en Tacna son los más bajos.

Tabla 2. Centroides por departamentos de los egresados universitarios 2014.

Departamento	N° egresados	Latitude	Longitude
Amazonas	122	-6,1702	-77,917
Ancash	416	-9,2235	-78,18
Apurimac	270	-13,662	-73,097
Arequipa	558	-16,406	-71,558
Ayacucho	149	-13,178	-74,201
Cajamarca	258	-7,0037	-78,548
Callao	186	-12,031	-77,12
Cusco	366	-13,564	-71,932
Huancavelica	188	-12,821	-74,867
Huánuco	329	-9,6001	-76,153
Ica	258	-13,94	-75,828
Junín	454	-11,938	-75,26
La Libertad	488	-8,079	-79,017
Lambayeque	466	-6,7578	-79,842
Lima	3872	-12,021	-77,023
Loreto	228	-3,9639	-73,438
Madre de Dios	104	-12,592	-69,194
Moquegua	142	-17,309	-71,035
Pasco	101	-10,681	-76,129
Piura	463	-5,0859	-80,665
Callao	1	-12,051	-77,126
Puno	438	-15,587	-70,072
San Martin	193	-6,5086	-76,572
Tacna	259	-17,988	-70,248
Tumbes	120	-3,5808	-80,445
Ucayali	135	-8,4572	-74,531

Figura 5. Visualización de los centroides por departamento en el territorio nacional

El cálculo de la desviación estándar fue realizado con el paquete aspace (Bui, Buliung, & Remmel, 2012) dando como resultado los siguientes valores: $CENTRE.x= -11,26993, $CENTRE.y= -76,01872, $SDD.radius= 4,398142 y $SDD.area= 60,76989. Los dos primeros valores corresponden al centroide de todos los datos, el tercer valor al radio del círculo para cuantificar la dispersión y el cuarto valor a la desviación estándar de las ubicaciones con respecto al centroide. La figura 7 presenta la dispersión de las coordenadas. Esta dispersión indica aproximadamente la separación promedio con respecto al centroide de todos los datos. Los puntos con un color rojo más intenso indican una mayor presencia de egresados en estos lugares. Se observa una fuerte dispersión de los egresados.

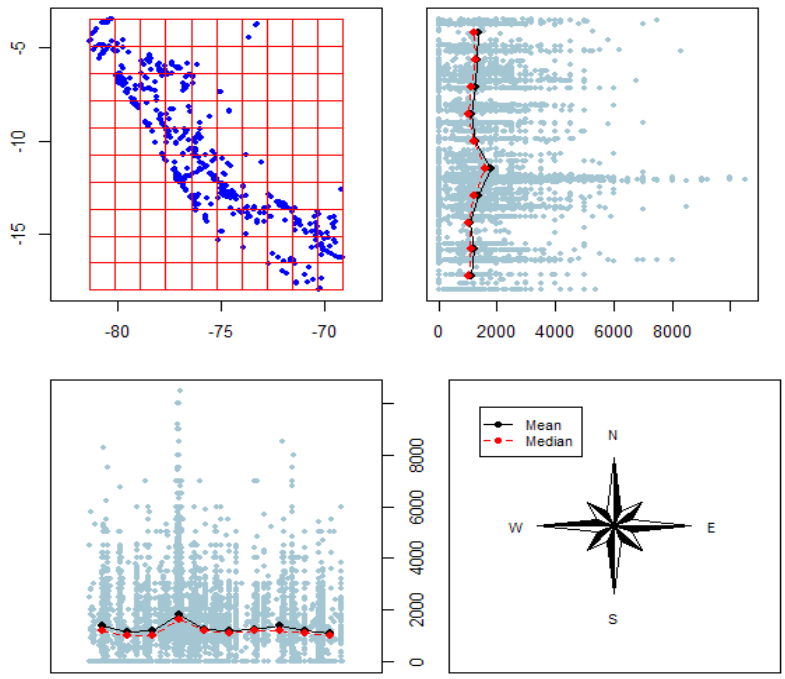

Figura 6. Valores de la media y mediana de ITMM de los egresados universitarios 2014 por fila y columna de las coordenadas geográficas.

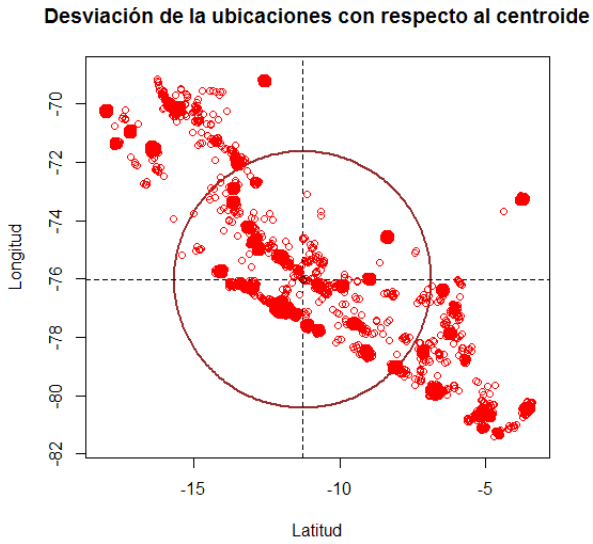

Figura 7. Dispersión de las coordenadas con respecto al centroide

Los cálculos realizados anteriormente están basados únicamente en las latitudes y longitudes que definen la ubicación de los egresados. A continuación se presentan otros tipos de análisis exploratorio que incluyen la variable ITMM y que fueron realizados con el software spatstat (Baddeley, 2010). Spatstat aplica el concepto de "window" para definir la muestra; es decir, dado una región espacial, window representa la muestra extraida de dicha región. Para manejar los datos de la muestra se deben crear objetos tipo ppp que permitirán realizar los cálculos correspondientes a la intensidad de los puntos, la autocorrelación espacial de Morán y la función K que mide la dependencia espacial entre los valores de ITMM.

La creación de un objeto tipo ppp requiere la aplicación de la función jitter con la finalidad de separar las observaciones que tienen las mismas coordenadas y de esta manera construir el objeto ppp que sea reconocible por el paquete spatstat. La grafica de las coordenadas ligeramente corregidas con jitter se presenta en la figura 8. Nótese en esta gráfica que, aunque no es muy claro, el contorno del Perú aparece ligeramente delineado.

Con la suposición de que el proceso puntual de valores observados de ITMM tiene una distribución homogénea de Poisson se calculó la intensidad con spatstat y la función intensity.ppp resultando un valor de $\hat{\lambda}$ = 10564/238= 44,3865 puntos por unidad cuadrada (la unidad cuadrada se refiere a (111 Km)^2 porque la distancia aproximada de separación entre dos latitudes es aproximadamente 111 km, siendo este mismo valor aproximadamente el mismo para la separación entre dos longitudes). El valor de 238 es el área de la "window" utilizada como muestra en función de los máximos y mínimos de las latitudes y longitudes registradas en la ENEU-2014. El valor de 44,3865 indica la cantidad de egresado por cada 12321 km^2. El error estándar estimado de $\hat{\lambda}$ según ([illegible]atial point patterns in R, 2010) es $\sqrt{\hat{\lambda}/Área\ (Window)}$ = 0,431854.

La autocorrelación espacial de Morán fue calculada con el software ape desarrollada por (Paradis, Blomberg, Bolker, & Claude, 2016) en la suite de R. La hipótesis nula es que los valores de ITMM no tienen autocorrelación espacial o que la autocorrelación es cero, versus la hipótesis alternante de que la autocorrelación es diferente de cero. El índice de Morán resultó ser 0,05076 lo que indica que la autocorrelación espacial de la variable ITMM es positiva y que la hipótesis nula de que la autocorrelación es cero es rechazada porque el p-value resultó ser un valor cercano a cero. El valor esperado del índice de Morán y su respectiva desviación estándar fueron -9,467e-05 y 0,0007094, respectivamente. La autocorrelación positiva de 0,05076 significa que los valores altos de ITMM tienden a estar próximos entre sí, de igual manera los valores pequeños.

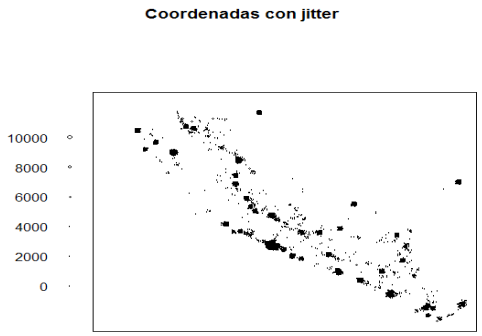

Figura 8. Gráfica de las coordenadas ligeramente corregidas con jitter.

La función K de (Ripley, 2001) fue obtenida con el software spatstat de (Baddeley, Turner, & Rubak, 2016). La figura 9 presenta la gráfica de la función K de Ripley. Se puede observar que las funciones observadas y corregidas $\hat{K}_{iso}(r)$, $\hat{K}_{trans}(r)$, y $\hat{K}_{bord}(r)$ difieren significativamente de la función teórica $K_{pois}(r)$ basada en la distribución de Poisson. Adicionalmente, se observa que las funciones $\hat{K}_{iso}(r)$, $\hat{K}_{trans}(r)$, y $\hat{K}_{bord}(r)$ están por encima de $K_{pois}(r)$ lo que indica que los valores de ITMM tienden a formar grupos (clusters). Esto confirma lo obtenido con el índice de Morán que resultó ser diferente de cero. La función K posee una pronunciada curvatura lo que indica que los ingresos de los egresados tienden a form

ar grupos o clusters.

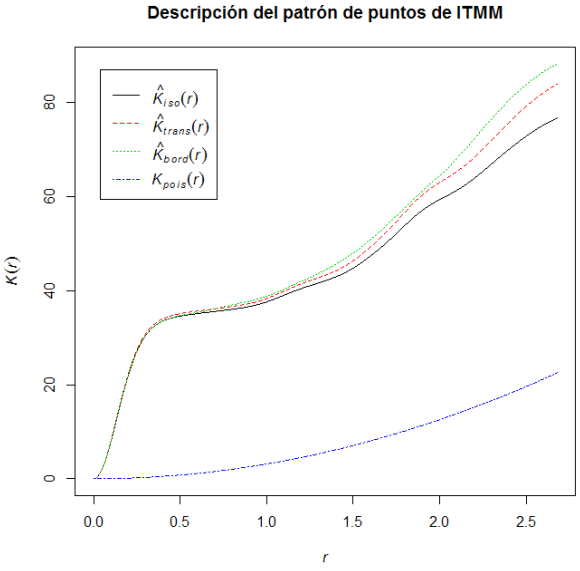

Figura 9. Gráfica de la función K con diferentes correcciones (iso, trans, bord) y gráfica de la función K teórica considerando la distribución de Poisson.

4. Conclusiones

Las conclusiones del presente trabajo de investigación son las siguientes:

El centro promedio tiene coordenadas geográficas con longitud= -76,01924 y latitud= -11,27021 que corresponde a un lugar situado a aproximadamente 40 km al noroeste de la ciudad de Tarma en el departamento de Junín.

La dispersión de los egresados con un círculo de radio 4,398142 con respecto al punto (-11,26993, -76,01872) tiene un área de 60,76989 unidades (cada unidad es 12321 km^2)

Hay poca presencia de egresados en las alturas de Arequipa, sur de Ica e interior de Piura. En la selva, es notoria la ausencia de egresados.

De acuerdo a la "window" de datos seleccionados, la intensidad espacial es de 44,3865 egresados por cada 12321 km^2.

La autocorrelación positiva calculada con el índice de Morán fue 0,05076 indica que los valores altos de ITMM tienden a estar próximos entre sí, de igual manera los valores pequeños. Este índice de Morán resultó ser significativo con un p-value aproximadamente igual a cero.

Los ingresos de los egresados tienden a formar grupos o cluster según lo señala la función K.

5. Literatura citada

Baddeley, A. 2010. Analysing spatial point patterns in R. Recuperado de https://www.researchgate.net/publication/228768037_Analysing_spatial_point_patterns_in_R

Baddeley, A., Turner, R., & Rubak, E. 2016. Recuperado de https://cran.r-project.org/web/packages/spatstat/spatstat.pdf

Bivand, R. S., Pebesma, E. J. & Gómez-Rubio, V. 2008. Applied Spatial Data Analysis with R. New York: Springer.

Borroto, R. & Martínez-Piedra, R. 2000. Geographical patterns of cholera in Mexico, 1991–1996. International Journal of Epidemiology, 764-772.

Bui, R., Buliung, R. & Remmel, T. 2012. aspace: A collection of functions for estimating centrographic statistics and computational geometries for spatial point patterns. Recuperado de https://rdrr.io/cran/aspace/

Calónico, S. & Ñopo, H. 2007. Returns to Private Education in Peru. Inter-American Development Bank.

CASA. (Setiembre de 2015). Centre for Advanced Spatial Analysis . Obtenido de MRes Spatial Dat Science & Visualisation: www.casa.ucl.ac.uk.

Corbett, C. & Hill, C. 2012. Graduating to a Pay Gap. The earnings of Women and Men One year after College Graduation. Washington: The American Association of University Women .

Cressie, N. & Wikle, C. K. 2011. Statistics for Spatio-Temporal Data. New Jersey: Wiley.

Cuartas, D. E., Ariza, Y., Pachajoa, H. & Méndez, F. 2011. Análisis de la distribución espacial y temporal de los defectos congénitos registrados entre 2004 y 2008 en un hospital de tercer nivel en Cali, Colombia . Colombia médica / Universidad del Valle, Facultad de Salud., 9-16.

Curatola Fernández, G. 2009. Patrones de distribución espacial de Triplaris americana en Tambopata, Perú. Recuperado de http://tesis.pucp.edu.pe/repositorio/ bitstream/handle/123456789/454/CURATOLA_FERNANDEZ_GIULIA_PATRONES_DISTRIBUCION.pdf?sequence=1

Dawson, R. 2011. How Significant Is A Boxplot Outlier? Journal of Statistics Education, 1-13. Recuperado de http://www.amstat.org/publications/jse/v19n2/dawson.pdf

Eliophotou, M., Pashourtidou, N., Polycarpou, A. & Pashardes, P. 2012. Students' expectations about earnings and employment and the experience of recent university graduates: Evidence from Cyprus. International Journal of Educational Development, 32(6): 805-813.

Gartner. 2013. IT Glossary. Obtenido de http://www.gartner.com/it-glossary/big-data

Hernández, N., Rodriguez, M. & Antón, O. 2013. Análisis espacial de la morbimortalidad del cáncer de mama y cérvix. Villa Clara. Cuba. 2004-2009. Rev Esp Salud Pública, 49-57.

Herrera Gómez, M., Cid, J. C. & Paz, J. A. 2012. Introduction to Spatial Econometrics: An application to the study of fertility in Argentina using R. MPRA Munich Personal RePEc Archive.

INEI. 2014. Encuesta nacional a egresados universitarios y universidades-Ficha técnica. Lima. Recuperado de http://iinei.inei.gob.pe/iinei/srienaho /Descarga/ FichaTecnica/454-Ficha.pdf

INEI. 2014. INEI ejecuta Encuesta a Egresados Universitarios y Universidades. Recuperado de https://www.inei.gob.pe/prensa/noticias/inei-ejecuta-encuesta-a-egresados-universitarios-y-universidades-7820/

INEI. 2015. Sistema de Documentación Virtual de Investigaciones Estadísticas. Recuperado de http://webinei.inei.gob.pe/anda_inei/index.php/catalog #_r=&sort_by=proddate&sort_order=desc

International Labour Organization. 2015. Promoting Jobs, Promoting People. Recuperado de http://www.ilo.org/global/statistics-and-databases/ statistics-overview-and-topics/status-in-employment/current-guidelines/lang--en/index.htm

Komsta, L. 2015. Package'outliers' . Recuperado de https://cran.r-project.org/web/packages/outliers/outliers.pdf

Laurent, T., Ruiz-Gazen, A. & Thomas-Agnan, C. 2012. GeoXp: An R Package for Exploratory Spatial Data Analysis. Journal of Statistical Software.

Lavado, P., Martínez, J. J. & Yamada, G. 2014. ¿Una promesa incumplida? La calidad de la educación superior universitaria y el subempleo profesional en el Perú. Asociaciòn Peruana de Economía, No. 23. Recuperado de http://perueconomics.org/wp-content/uploads/2014/01/WP-23.pdf

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C. & Gunatilleke, I. 2009. Ecological information from spatial patterns of plants: insights from point process theory. Journal of Ecology, 616-628.

Lisset, E. & Fuentes, S. 2014. "Distribución espacial de la Diabetes Mellitus, el Asma Bronquial y la Hipertensión Arterial en Cuba. Génetica Comunitaria.

Ministerio de Trabajo. 2013. DECRETO SUPREMO N° 166-2013-EF . Recuperado de http://www.trabajo.gob.pe/boletin/documentos/boletin_31/doc_boletin_31_02.pdf

Moral García, F. 2004. Aplicación de la geoestadística en las ciencias ambientales . ecosistemas, 78-86.

Moran, P. 1950. Notes on Continuous Stochastic Phenomena. Biometrika, 17-23. Recuperado de http://dds.cepal.org/infancia/guia-para-estimar-la-pobreza-infantil/ bibliografia/capitulo-IV/Moran%20Patrick%20A%20P%20(1950)%20Notes%20on%20continuous%20stochastic%20phenomena.pdf

Moreno, M. 2011. La distribución espacial de las comunidades peruanas en los Estados Unidos. Debates en Sociología, 27-55.

ONGEI. 2010. INFORMACION TERRITORIAL OFICINA NACIONAL DE GOBIERNO ELECTRONICO E INFORMATICA. Perú. Recuperado de http://es.slideshare.net/agroredperu/oficina-nacional-de-gobierno-electronico-e-informatica

Oreopoulos, P. & Petronijevic, U. 2013. Making College Worth It: A Review of the Returns to Higher Education. www.futureofchildren.org, 23(1). Recuperado de http://files.eric.ed.gov/fulltext/EJ1015240.pdf

Paradis, E. 2014. Moran's Autocorrelation Coefficient in Comparative Methods. Recuperado de http://star-www.st-andrews.ac.uk/cran/web/packages/ape/ vignettes/ MoranI .pdf

Paradis, E., Blomberg, S., Bolker, B. & Claude, J. 2016. Analysis of Phylogenetics and Evolution. Recuperado de http://ape-package.ird.fr/

Quiroz Cornejo, Z. J. 2011. Estudio de la distribución de aves marinas en relación con sus presas y la profundidad del límite superior de la zona de mínimo oxígeno en el sistema de la corriente de Humboldt utilizando análisis de procesos puntuales y modelos lineales generalizados. Lima, Perú.

R Core Team. 2016. R: A language and environment for statistical computing. Computing, R Foundation for Statistical. Vienna, Austria. Recuperado de https://www.R-project.org/

Ripley, B. 2001. Spatial Statistics in R. Rnews, 14-15.

Rodríguez, J. & Montoro, L. 2013. La educación superior en el Perú: situación actual y perspectivas.

Shekhar, S., Zhang, P., Huang, Y. & Raju Vatsavai, R. 2009. Spatial Data Mining. En Encyclopedia of Database Systems, 2695-2698. Springer US.

Shi, W. 2015. Towards Spatial Data Science. Recuperado de Center for Geographic Analysis: http://www.gis.harvard.edu/events/seminar-series/towards-spatial-data-science

Silipo, R. 2016. Seven Techniques for Data Dimensionality Reduction. (KDnuggets, Editor) Recuperado de http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html

SUNAT. 2015. Guía Tributaria. Recuperado de http://guiatributaria.sunat.gob .pe /component/rssearch/?option=com_rssearch&search=dependiente&x=0&y=0&view=results&layout=default&module_id=210

Tarpey, T. 2012. Spatial Statistics. Recuperado de http://www.wright.edu /~thaddeus.tarpey/es714spatial.pdf

Van Der Loo, M. 2016. Package 'extremevalues'.

Yamada, G. 2007. Retornos a la Educación Superior en el Mercado Laboral: ¿Vale la pena el esfuerzo?

Zwiefelhofer, D. 2016. Batch Geocoding. Recuperado de http://www.findlatitudeandlongitude.com/batch-geocode/#.WDyJ7IWcGCQ