PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XIX N°2, pp. 50-58, Julio–Diciembre 2016

Métodos Factoriales en el Análisis de Datos Espaciales. Una Aplicación a los Datos del Censo Agropecuario 2012 para la Caracterización de las Provincias del Perú

Emma Cambillo[1] *Ysela Agüero*[2] *María del Pilar Alvarez*[3]
Alicia Riojas[4]

(Recibido: 28/09/2016 - Aceptado: 10/10/2016)

Resumen: Las provincias peruanas fueron tipificadas por métodos factoriales con la incorporación de datos georreferenciados con información del Censo Nacional Agropecuario de 2012. Se determinaron tres indicadores: La intensidad de la actividad agrícola, la comercialización de la producción agrícola, y el uso de fuentes de agua para el riego; El uso del Índice de Moran ayudó a identificar las provincias con características agrícolas similares. Los resultados podrían ser utilizados para encaminar algunos objetivos del plan estratégico del sector agrícola, y supervisar las políticas de desarrollo institucional en el sector. La información también podría ser útil para el diseño de planes de desarrollo que respondan a las necesidades del sector agrícola y para tener una idea clara de las características agrícolas de todas las provincias del Perú, con el objeto de orientar la inversión en el país.

Palabras Claves: Análisis exploratorio de datos espaciales, Índice de Moran, Métodos factoriales, Análisis espacial, censo agropecuario.

Factorial methods in the analysis of spatial data. An application to the 2012 agricultural census data for the characterization of the provinces of Peru.

Abstract: The Peruvian provinces were tipified by factorial methods with incorporating geo-referenced data with information from the 2012 National Agricultural Census. Were determined three indicators : Intensity of agricultural activity, commercialization of agricultural production, and use of water sources for irrigation; The use of the Moran Index helped identify provinces with similar agricultural characteristics. The results could be used to route some objectives of the strategic plan of the agricultural sector, and monitor policies of institutional development in the sector. The informatión also could be useful for designing development plans that meet the needs of the agricultural sector and to have a clear idea of the agricultural characteristics of all the provinces of Peru, with the object to orient the investment in the country.

Key Words: Exploratory spatial data analysis, Moran Index, Factorial Methods, Spatial Analysis, Agricultural census.

[1]UNMSM, Facultad de Ciencias Matemáticas, e-mail: ecambillom@unmsm.edu.pe
[2]UNMSM, Facultad de Ciencias Matemática, e-mail: yaguerop@unmsm.edu.pe
[3]UNMSM, Facultad de Ciencias Matemática, e-mail: malvarezr1@unmsm.edu.pe
[4]UNMSM, Facultad de Ciencias Matemática, e-mail: aliriojas@hotmail.edu.pe

1 Introducción

Los Censos Agropecuarios en el Perú, permiten conocer la estructura del sector agropecuario, de manera actualizada y desagregada a nivel de cada uno de los distritos del país que realizan actividades agropecuarias.

El IV Censo Nacional Agropecuario 2012, ejecutado por elInstituto Nacional de Estadística e Informática (INEI), es la investigación estadística más importante del Sector Agrario, para proporcionar datos que permitan conocer la base productiva agropecuaria mediante el recojo de las declaraciones de todos los productores agropecuarios del país. Esta información se reporta a nivel descriptivo sin tomar en cuenta el carácter multidimensional y espacial de estos datos.

El informe con los resultados definitivos del IV Censo Nacional agropecuario, contiene información estadística sobre la estructura del sector agropecuario, la cantidad de unidades agropecuarias y los productores agropecuarios que las conducen, según su condición jurídica y el régimen de tenencia, proporciona información sobre la cantidad de parcelas y el tamaño de las unidades agropecuarias estudiadas de acuerdo con el uso actual de la tierra, lo que ha permitido conocer la superficie total agropecuaria, la superficie agrícola y sus componentes, el tipo de agricultura y los sistemas utilizados para irrigar las tierras. Adicionalmente, se determinó la superficie cultivada, las prácticas agrícolas y el uso de energía al momento de la entrevista. (INEI, 2013).

Alza y Cambillo (2010), con datos del Censo Agropecuario 1994; incorporaron información georeferencial, para obtener mapas de variabilidad espacial, con este fin utilizaron los escores factoriales del análisis de correspondencia, identificando tres grupos de provincias con características agropecuarias similares, de acuerdo con la intensidad de actividad agropecuaria y el nivel tecnológico. Pero, no incorporaron la autocorrelación espacial que permitiria detectar relaciones subyacentes.

El objetivo de esta investigación fue caracterizar las unidades agropecuarias de las provincias del Perú, mediante el uso de las técnicas estadísticas de análisis factorial y espacial reduciendo la dimensión y estimar las correlaciones espaciales para la caracterización de las provincias del Perú a partir de un gran número de variables relacionadas obtenidas del Censo Nacional Agropecuario.

2 Metodología

La investigación es de tipo descriptivo correlacional con un diseño observacional ecológico. (Hernández y colaboradores, 2014; Landero y González, 2006).

Los datos utilizados corresponden al IV Censo Nacional Agropecuario realizado por el INEI entre el 15 de octubre y el 15 de noviembre de 2012. Este censo contiene información sobre el número de unidades agropecuarias y los productores agropecuarios que las conducen, según su condición jurídica y el régimen de tenencia; proporciona información sobre la cantidad de parcelas y el tamaño de las unidades agropecuarias que se ha estudiado acorde con el uso actual de la tierra lo que ha permitido conocer la superficie total o agropecuaria, la superficie agrícola y sus componentes, el tipo de agricultura y los sistemas utilizados para irrigar las tierras, la superficie con cultivos al momento de la entrevista, las prácticas agrícolas y el uso de energía.

Para el estudio se seleccionaron un total de 150 variables correspondientes a:

- Estructura del espacio agropecuario: Unidades agropecuarias, superficies agropecuarias, superficie agrícola, superficie cultivada, parcelas y tenencia de la tierra.
- Uso de la tierra: cultivos, árboles diversos y siembras.

- Riego: Procedencia del agua, pozos, canales, apreciaciones sobre el agua y derecho de uso.
- Prácticas agrícolas: insumos, buenas prácticas, uso de plaguicidas agrícolas.
- Uso de energía eléctrica, animal y mecánica (tractores).
- Equipos de energía animal, mecánica y otras.
- Estructura de la actividad pecuaria: población pecuaria y prácticas pecuarias.
- Condiciones de la actividad agropecuaria: Capacitación, crédito, infraestructura.
- Características del productor agropecuario: Perfil del productor, características de su hogar.

En una primera etapa, se realizó un análisis exploratorio y descriptivo de los datos. En la segunda etapa, mediante el análisis factorial se redujo el número de variables formando un total de tres nuevos factores los cuales fueron analizados utilizando técnicas de análisis de datos espaciales.

El procesamiento de datos se realizó utilizando los programas GeoDa versión 1.6.0, SPSS versión 21 y Excel 2013. (Buzai y Baxendal, 2009, GeoSIG, 2014).

2.1 Análisis factorial

Una técnica que ocupa un lugar primordial entre los métodos de análisis de datos estadisticos, es el Análisis Factorial (AF), que consiste en estudiar la estructura de una nube de puntos, resultante de un conjunto de variables observadas para cada objeto; permitiendo condensar la información contenida en los datos reduciendo la dimensión con una mínima pérdida de información. La obtención de estos factores está íntimamente ligada a la matriz de varianza-covarianza o la matriz de correlaciones.

Los objetivos que se persiguen al usar el AF son determinar si existe un conjunto menor de variables no relacionadas que expliquen las relaciones existentes entre las variables observadas; determinar el número de variables subyacentes; evaluar los objetos o individuos sobre estas nuevas variables, interpretar y usarlas en análisis posteriores. (Johnson y Wichern, 2007; Catena y colaboradores, 2003; Levy y Varela, 2003; Cea D'Ancona, 2002).

El método factorial parte de un conjunto amplio de variables que presentan interrelaciones importantes. Se asume que las relaciones existen porque las variables son manifestaciones comunes de factores no "observables" de forma directa y se pretende identificar esos factores resumiendo y clasificando las relaciones entre ellas con la menor pérdida de información posible. Es un método de interdependencia y no hay una selección a priori de una variable dependiente.

2.2 Modelo Factorial Ortogonal

Considere el modelo:

$$ {}_p(X-\mu)_1 = {}_p\Lambda_{mm} F_1 + {}_p\varepsilon_1 \tag{1} $$

donde

${}_pX_1$: vector de variables;

${}_p\mu_1$: vector de medias de ${}_pX_1$;

${}_p\Lambda_m$: matriz de pesos o cargas factoriales;

${}_mF_1$: vector de factores comunes;

${}_p\varepsilon_1$: vector de factores específicos.

El modelo (1) es similar al modelo de regresión lineal múltiple, con la diferencia que F_1 es no observable.

(i) $E(F) = 0$

(ii) $E(\varepsilon) = 0$

(iii) $\text{Cov}(F) = E(FF') = I$

(iv) $\text{Cov}(\varepsilon) = E(\varepsilon\varepsilon') = \Psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$

(v) $\text{Cov}(\varepsilon F') = E(\varepsilon F') = 0$

Resumiendo, $F \approx (0, I), \varepsilon \approx (0, \psi)$.

Propiedades

i) La matriz de pesos Λ contiene las covarianzas entre los factores y las variables observadas, los cuales pasarán a ser los coeficientes del modelo cuando se explica las variables observadas por los factores

$$\text{Cov}(X, F) = \Lambda,\ \text{Cov}\left(X_j, F_{kj}\right) = \lambda_{jk},\ (j \neq k = 1, 2, \ldots, p)$$

ii) La matriz de covarianzas admite una descomposición como la suma de dos matrices

$$\sum = \Lambda\Lambda' + \Psi$$

donde la primera matriz contiene la parte común al conjunto de las variables, la comunalidad y depende de las covarianzas entre las variables y los factores y la segunda es una matriz diagonal que contiene la parte específica de cada variable que es independiente de las restantes, denominada la especificidad.

En el caso de las variables estandarizadas se utiliza la matriz de correlaciones $\mathbf{J}$, en lugar de la matriz de varianza-covarianza Σ. Luego la descomposición es de la forma $\mathbf{J} = \Lambda\Lambda' + \psi$.

El propósito del análisis factorial es encontrar los mejores valores λ, los cuales representan las correlaciones entre los factores y los pesos y que permitan reproducir los valores de x_{ik} más próximos a los observados indicando claramente cuales variables pertenecen a los factores identificados.

Existen diferentes métodos de estimación de estos pesos, entre ellos podemos citar los métodos de componentes principales, de ejes principales, de máxima verosimilitud, el método alfa y otros. El método más comúnmente usado es el de componentes principales, en el cual los pesos de

los factores se obtienen mediante la descomposición espectral de la matriz de covarianzas o de correlaciones.

Para la determinación del número de variables, también existen varios métodos, tales como, el porcentaje de variación total explicada, los criterios de Catell o Kaiser y una prueba de hipótesis basado en una distribución normal multivariada. Una buena práctica es retener aquellos factores asociados a los autovalores mayores de 1.

Como el problema es encontrar la matriz de los pesos que permita identificar cuales variables están relacionadas a cada uno de los factores, es posible mejorar la solución inicial mediante rotaciones de los factores, de manera que, los pesos se aproximen a uno o a cero. La rotación mantiene la información total pero la reasigna a través de los factores y facilita la interpretación de estos. Para la obtención de esta estructura, existen varios tipos de rotación, por ejemplo las rotaciones Varimax o Quartimax (Johnson y Wichern, 2007).

2.3 Análisis espacial

En muchas situaciones, la información disponible tiene una componente geográfica, que dada su complejidad, requiere de técnicas especiales para integrarla y poder realizar así un análisis conjunto de todas las características del fenómeno en estudio.

El análisis espacial se centra en el estudio de las componentes geográficas, definiendo los elementos que la constituyen y la manera como éstos se comportan bajo ciertas condiciones, permitiendo describir y explicar el fenómeno integrando las relaciones espaciales. (Melo E. y Aidar de Freitas T. 2010).

Cuando se trabaja con datos espaciales deben considerarse explícitamente las características propias de esta información como la georreferenciación, multidireccionalidad y multidimensionalidad. Estas peculiaridades de los datos geográficos suelen producir en los mismos los fenómenos conocidos como efectos espaciales, de dependencia o autocorrelación y heterogeneidad espacial. La dependencia espacial podría ser definida como la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro punto (Moreno y Vayá, 2000). En cuanto a la heterogeneidad espacial, se trata de un efecto relacionado con la diferenciación espacial o regional y viene definido por la ausencia de estabilidad espacial del comportamiento humano o de otras relaciones. (Martori 2008).

Esta peculiaridad de los datos espaciales motiva el nacimiento de una subdisciplina del análisis exploratorio de datos denominada Análisis Exploratorio de Datos Espaciales (AEDE), diseñada para el tratamiento específico de los datos geográficos. (Chasco, 2003).

El AEDE es una sub-disciplina del Análisis Exploratorio de Datos (AED), que consiste en el tratamiento y comparación de una batería de variables utilizando técnicas que identifican formas estables espacialmente en ellas. Según Tukey (1977), autor que hizo posible la extensión de este tipo de análisis multivariado, el AED podría definirse como "...el conjunto de herramientas gráficas y descriptivas utilizadas para el descubrimiento de patrones de comportamiento en los datos y el establecimiento de hipótesis con la menor estructura posible".

El AEDE, según Anselin (1999), puede definirse como el conjunto de técnicas que describen y visualizan las distribuciones espaciales, identifican localizaciones atípicas o "atípicos espaciales", descubren esquemas de asociación espacial, agrupamientos o puntos calientes y sugieren estructuras espaciales u otras formas de heterogeneidad espacial. Por tanto, estos métodos tienen un carácter descriptivo más que confirmatorio, aunque la detección de estructuras espaciales en las variables geográficas, hace posible la formulación de hipótesis previas para la modelización y, en

su caso, posible predicción espacial de nuevos datos.

Para dar una medida resumen de la intensidad de la autocorrelacion entre los territorios considerados se utiliza el Índice de Moran global (Moran, 1950). Este indice mide la autocorrelación espacial basada en las ubicaciones y en los valores de las entidades simultáneamente. Dado un conjunto de entidades y un atributo asociado, evalúa si el patrón expresado está agrupado, disperso o es aleatorio. (Martori, 2008, Goodchild, 1986). El Indice de Moran toma valores entre $(-1, 1)$. Si los valores tienden a agruparse espacialmente (los valores altos se agrupan cerca de otros valores altos; los valores bajos se agrupan cerca de otros valores bajos), el Índice de Moran será positivo. Cuando los valores altos rechazan otros valores altos y tienden a estar cerca de valores bajos, el Índice será negativo. Si los valores positivos de los productos cruzados equilibran los valores negativos de los productos cruzados, el Índice será cercano a cero. (Griffith, 1987, Cambillo,2013).

3 Resultados y Discusión

Para la caracterización agropecuaria de las provincias del Perú se han utilizado los indicadores basados en el resultado del análisis factorial, los cuales fueron denominados como:

Indicador 1: Actividad agropecuaria: Mide la intensidad de actividad agropecuaria de las unidades agropecuarias en las provincias del Perú.

Indicador 2: Comercialización de productos: Mide la cantidad de productos de las unidades agropecuarias destinadas a venta en las provincias del Perú.

Indicador 3: Fuente de agua que utilizan para riego: Mide la cantidad de unidades agropecuarias que utilizan fuentes de agua convencionales en las provincias del Perú.

En la Figura 1, se puede observar que la mayor actividad agropecuaria se realiza en las provincias de la costa y sierra. El índice de Moran calculado para la actividad agropecuaria (0,230), es bajo, lo cual indicaría una débil autocorrelación espacial, esto es, las provincias con mayor intensidad de actividad agropecuaria se encuentran distantes con relación a otras provincias del país con la misma intensidad de actividad agropecuaria.

Las provincias de la costa y parte de selva destinan su producción para la comercialización, mientras que las provincias de la Sierra, lo destinan principalmente para el autoconsumo (Figura 2). Esto también se puede apreciar con el Índice de Moran (0,603) que muestra una mayor intensidad de agrupamiento espacial en relación con la comercialización de productos agrícolas.

En relación a la fuente de agua para el riego de las áreas cultivadas (Figura 3), se observa una intensidad relativamente alta de agrupamiento espacial (Índice de Moran=0,641). Las provincias de la costa utilizan el agua de rios, pozos y canales, como fuentes de agua para riego a diferencia de las provincias de la Sierra y la Selva, en las que como es natural, las lluvias son la principal fuente de agua para el riego de los cultivos.

Figura 1: Provincias del Perú según la intensidad de la Actividad Agropecuaria. Censo Nacional Agropecuario - 2012.

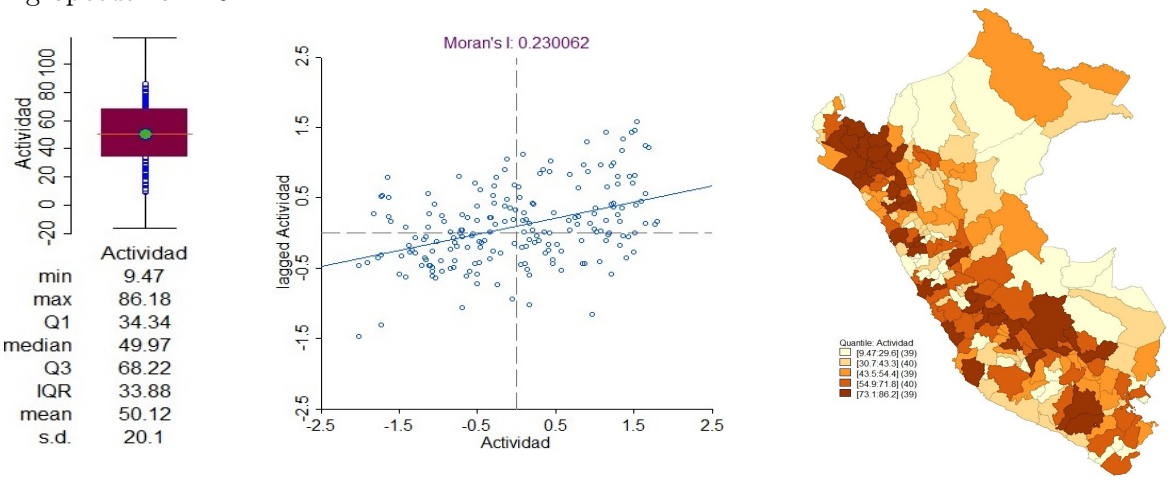


Figura 2: Provincias del Perú según la intensidad de la Actividad de comercialización de productos agropecuarios. Censo Nacional Agropecuario - 2012.

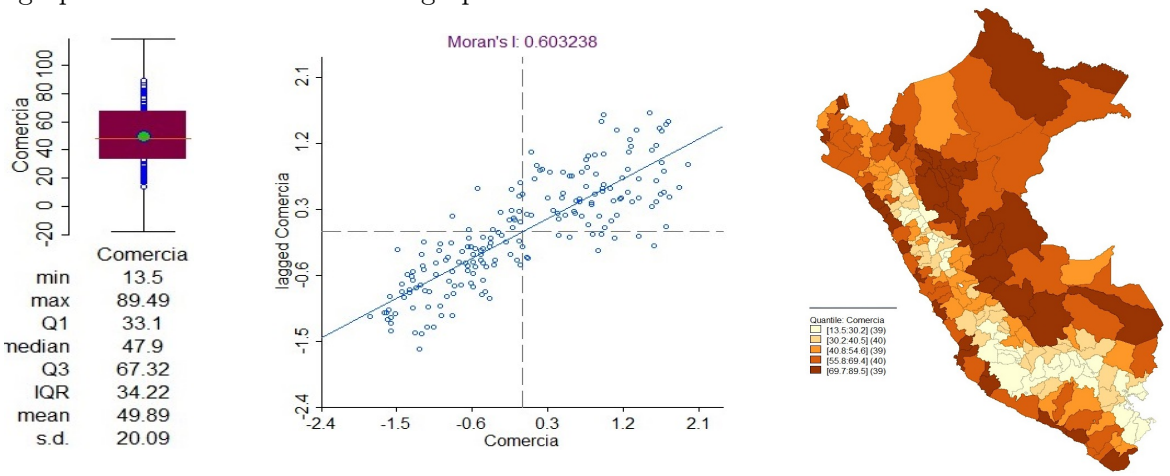
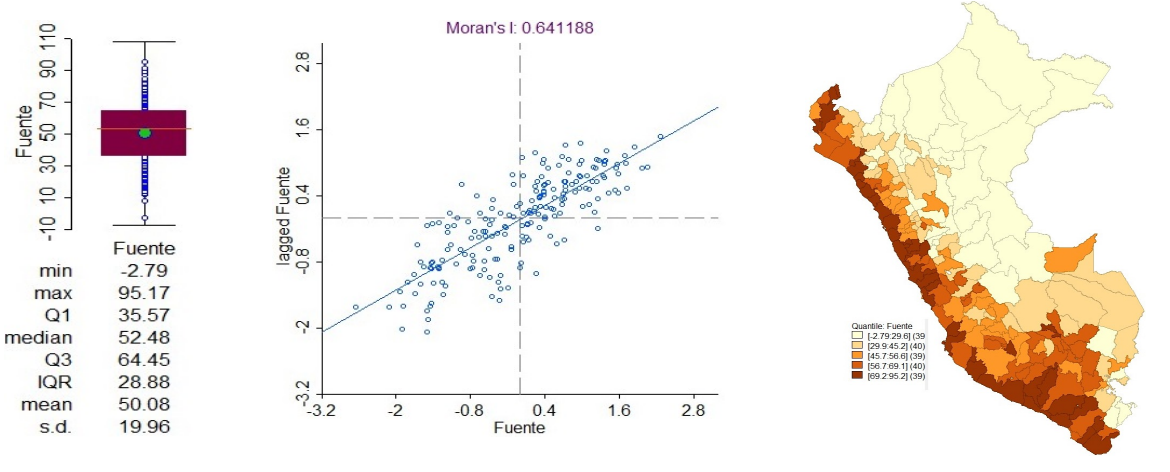


Figura 3: Provincias del Perú según la distribución espacial del tipo de fuente de agua utilizada para el riego de las parcelas cultivadas.



4 Conclusiones

- La información proveniente de los censos agropecuarios es eminentemente georreferenciada, es decir, que se trata de información que tiene sentido de acuerdo a su posición geográfica.
- El uso de los métodos factoriales permite la reducción de los datos con una mínima perdida de información y si los indicadores resultantes del análisis son ubicados espacialmente, se obtiene una mayor riqueza de información.
- En la presente investigación se utilizaron 150 variables obtenidas del Censo Agropecuario de 2012, las cuales se redujeron a tres indicadores mediante la aplicación del análisis factorial.
- El georeferenciamiento de los tres indicadores resultantes y el cálculo del índice de Moran permitieron ubicar las provincias de acuerdo a la intensidad de su actividad agrícola
- Por otro lado, existe una correlación espacial relativamente alta en relación a la fuente de agua utilizada diferenciando las provincias ubicadas en la costa con relación a la sierra y selva.
- Finalmente, se observa que para un número de provincias de la costa y la selva la producción agrícola tiene fines de comercialización, mientras que la mayoría de provincias de la sierra producen mayormente para el autoconsumo, esto último podría explicarse por los suelos accidentados de la sierra lo que impide el uso de fuentes de energía mecánicas y la fuerte dependencia de las lluvias para el desarrollo de la actividad agrícola.

4.1 Recomendaciones

La necesidad de incoporar el georeferenciamiento a los indicadores estadísticos, crea la necesidad de que los profesionales de esta área desarrollen y utilicen técnicas de análisis de datos espaciales, por lo que es recomendable que las Escuelas de Estadística modernizen sus planes de estudios incoporando cursos en los que se pueda aprender las técnicas de análisis de datos espaciales y de georeferenciamiento.

5 Agradecimientos

Al Consejo Superior de Investigaciones del Vicerrectorado de Investigación de la UNMSM, por el soporte financiero para la ejecución del proyecto de investigación otorgado el año 2015 cuyo resultado se plasma en la presente publicación.

Referencias Bibliográficas

[1] Alza J., Cambillo E. (2010). *Caracterización agropecuaria de las provincias del Perú.* Magistri et Doctores, Vol 59 pg. 133-150 2.

[2] Anselin, L. (1999). *The future of spatial analysis in the social sciences.* Geographic Information Sciences, 5(2); pp. 67-76.

[3] Buzai, G.D.; Baxendale, C.A. (2009). *Análisis exploratorio de datos espaciales.* Universidad Nacional de Luján. Sección software y metodología, Vol.1, N° 1, pp. 1-11.

[4] Cambillo E. (2013). *Índice de Moran en el análisis exploratorio de datos espaciales.* Primer Seminario de Análisis Espacial. UNMSM - Colegio de Matemáticos del Perú.

[5] Catena A., Ramos M., Trujillo H. (2003). *Análisis multivariado un manual para investigadores.* Ed. Biblioteca Nueva.

[6] Cea D' Ancona. (2002). *Análisis multivariable.* Teoría y Práctica en la investigación social. Ed. Síntesis.

[7] Chasco Y. (2003). *Métodos gráficos para el análisis exploratorio de datos espaciales.* Revisado el 01 de diciembre de 2014.

[8] GeoSIG. (2014). *GeoDAS Software Manual.* Revisado el 15 de marzo de 2014.

[9] Goodchild F. (1986). *Spatial Autocorrelation.* Catmog 47, Geo Books.

[10] Griffith, D. (1987). *Spatial Autocorrelation: A Primer. Resource Publications in Geography.* Association of American Geographers.

[11] Hernández R., Fernández C., Baptista P. (2014). *Metodología de la investigación.* 6ta Ed.Mc Graw Hill - México D.F.

[12] INEI. (2013). *Resultados definitivos IV. Censo Nacional Agropecuario.* INEI. Ministerio de Agricultura y Riego. Revisado el 15 de Julio de 2015.

[13] Johnson R. A., Wichern D. W. (2007). *Applied Multivariate Statisitical Analysis*, 6° ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

[14] Landero R., González M. (2006). *Estadística con SPSS y metodología de la investigación.* Ed. Trillas . México D. F.

[15] Levy J., Varela, J. (2003). *Análisis multivariado para las ciencias sociales.* Pearson.

[16] Martori J. (2008). *La incorporación del espacio en los métodos estadísticos: autocorrelación espacial y segregación.* X Coloquio Internacional de Geocrítica.- Univ de Barcelona.

[17] Melo E, Aidar de Freitas T. (2010). *Distribución y autocorrelación espacial de indicadores de la salud de la mujer y del niño en el estado de Paraná.* Brasil Enfermagem. 18(6)10 pp. 1-10.

[18] Moreno, R., Vayá E. (2000). *Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial.* Edicions Universitat de Barcelona, colección UB 44, manuals.

[19] Moran, P.A.P. (1950). *Notes on continuous stochastic phenomena.* Biometrika 37, pp17-23.

[20] Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading: Addison-Wesley.