CONTRIBUCIONES DE LAS MATEMÁTICAS A LAS CIENCIAS SOCIALES: CASO DE APLICACIÓN ANÁLISIS DE DATOS ESPACIALES A LA INMIGRACIÓN Y SEGREGACIÓN ESPACIAL.

Luis Antonio Hurtado

Universidad del Valle

Instituto de Educación y Pedagogía

Licenciatura en Educación Básica con Énfasis en Matemática

2019

CONTRIBUCIONES DE LAS MATEMÁTICAS A LAS CIENCIAS SOCIALES: CASO DE APLICACIÓN ANÁLISIS DE DATOS ESPACIALES A LA INMIGRACIÓN Y SEGREGACIÓN ESPACIAL.

Luis Antonio Hurtado

Código: 201257167

Trabajo de grado para optar por el título de Licenciado en Educación Básica Con Énfasis en Matemáticas

Director

Mg. Steev Romero Agredo

Universidad del Valle

Instituto de Educación y Pedagogía

Licenciatura en Educación Básica con Énfasis en Matemática

2019

Contenido

Tabla de ilustraciones

Índice de tablas.

Resumen

En este trabajo se exhibe la importancia de la contribución de algunos objetos matemáticos en la modelación que utilizan otras ciencias del conocimiento, particularmente en las ciencias sociales para explicar, extrapolar y pronosticar ciertos fenómenos de interés.

A partir de este análisis se mostrará una serie de métodos enfocados en el análisis de datos espaciales, un tema propio de la economía regional, que permitirá evidenciar algunos conceptos matemáticos como los índices de correlación, álgebra matricial y regresiones lineales.

Así, el análisis de datos espacial permitirá llevar un estudio sobre la inmigración de personas en la ciudad de Cali, analizadas a partir de modelos estadísticos como el índice de Moran y modelos de regresión espacial. Este estudio se llevará a cabo a partir de una serie de datos que deberán ser sistematizados a partir de un software llamado GeoDa, donde se observará como resultados de otras ciencias tienen el sustento teórico en la matemática.

Palabras clave: Análisis espacial, índice de Morán, GeoDa, Modelos de Regresión Espacial, segregación espacial.

Abstract

In this work the importance of the contribution of some mathematical objects in the modeling used by other knowledge sciences is shown, particularly in the social sciences to explain, extrapolate and forecast certain phenomena of interest.

From this analysis we will show a series of methods focused on the spatial analysis of data, a subject specific to spatial econometrics, which will allow us to demonstrate some mathematical concepts such as correlation indexes, matrix algebra and linear regressions.

Thus, spatial econometrics will allow us to carry out a study on the immigration of Antioquia people to the city of Cali, analyzed from statistical models such as the Moran index and regression models. This study will be carried out from a series of geographic data that should be systematized from a software called GeoDa, where it will be observed as results of other sciences have the theoretical support in mathematics

Key Words: Spatial analysis, Moran index, GeoDa, Spatial Regression models, Segregation spatial.

Introducción.

La matemática se considera una ciencia importante por tener como finalidad la creación de estructuras formales que permitan modelar y representar fenómenos evidenciados en la realidad y así obtener datos significativos que pueden ser utilizados para construir, optimizar, pronosticar, entre otras cosas más. En este caso, se pretende analizar como la matemática se utiliza como puente para consolidar teorías, confirmar hipótesis y establecer conclusiones que bajo estimaciones deberían ser útiles para describir una realidad. Para esta tarea, nos enfocaremos en un área de las ciencias sociales, la economía regional, ya que existe el imaginario de que pueden estar poco vinculadas *per sé* y se llegó a resultados que muestran que no es así.

Este proyecto se fundamenta principalmente en dos marcos teóricos, uno de ellos utiliza resultados específicos del álgebra de matrices y modelación mediante regresiones, el otro marco, el de la economía regional, considera variables sociales en función del espacio apelando al análisis de datos espaciales. Así, el estudio en conjunto, permite visibilizar como algunos conceptos y objetos de la matemática y estadística, facilitan el análisis de los efectos que se pueden presentar las relaciones de las variables demográficas, sociales y económicas en el espacio.

Teniendo en cuenta lo anterior, el planteamiento de este trabajo de grado presentará tres etapas: La inicial muestra algunos aspectos matemáticos y estadísticos importantes, en especial lo que nos brinda las teorías del algebra lineal y los modelos de regresión. Una segunda etapa, presenta los hechos estilizados de la economía regional, que se lograron gracias a la interdisciplinariedad dada por las ciencias exactas y, por último, la tercera etapa muestra un ejercicio de simulación y modelación, que evidencia el trabajo conjunto de ambas áreas, es decir la relaciona intrínseca del marco teórico de las ciencias exactas y una ciencia social.

Capítulo 1. Principales conceptos matemáticos utilizados en la economía regional para el análisis de datos espaciales.

1.1. Introducción

Este capítulo realiza una descripción de algunos conceptos matemáticos importantes en la explicación de diferentes fenómenos que se presentan en la economía regional. Se toman elementos fundamentales del algebra lineal como son las matrices, lo que conlleva a ahondar en sumatorias, importantes para lograr desarrollar múltiples operaciones de múltiples datos. También se estudian elementos de la estadística importantes para el desarrollo del capítulo 3, como lo son los índices de correlación simples y multivariados, como los métodos de regresión lineal para generar modelos estimados.

1.2. Sumatorias

La sumatoria es un concepto que se aplica comúnmente en varias áreas de la matemática y la estadística, y es fundamental en el análisis sobre datos espaciales, pues permitirá que se expresen de una manera más ordena y generaliza todos los datos que se prevén para un estudio.

Según Bagur (2009), este concepto tiene sus inicios con Johann Carl Friedrich Gauss, quien tuvo aportaciones importantes en la teoría de los números, análisis matemático, cálculo diferencial, entre otras áreas del conocimiento. La aproximación más conocida para este concepto se evidencia a través de la famosa tarea impuesta a Gauss en la escuela, que consistió en obtener el resultado de la suma de todos los números del 1 al 100, donde sin mayor complicación respondió 5050.

Fácilmente se concluye que 101 se debía multiplicar por 50, y el algoritmo se explica teniendo en cuenta el siguiente esquema:

$$\begin{array}{c} 1 \quad 2 \quad 3 \quad 4 \ldots\ldots.\ 47 \quad 48 \quad 49 \quad 50 \\ 100 \quad 99 \quad 98 \quad 97 \ldots\ldots.54 \quad 53 \quad 52 \quad 51 \\ \hline 101+101+101+101+\cdots+101+101+101+101=(101)\times(50)=5050 \end{array}$$

Obteniendo de esta manera la base para formar o determinar una fórmula más concreta para cualquier n que se desee sumar:

$$S_n = \frac{(a_1 + a_2)n}{2}$$

Con las condiciones siguientes:

- a_1 es el primer número de la serie
- a_2 es el último número de la serie
- n es el número de términos a sumar

Así entonces fue surgiendo el concepto de sumatoria, que al día de hoy se utiliza en su gran mayoría en los estudios e investigaciones matemáticos que recurren a la suma de un conjunto n de números, los cuales permiten manipular de una mejor manera grandes sumandos.

Por lo cual Sandoval (2011) determina que las sumatorias son de gran ayuda a la hora de poder expresar un conjunto numérico de n elementos $\{a_1, a_2, \dots, a_n\}$ permitiendo representar sumas de muchos sumandos, n o en el caso de infinitos sumandos, expresadas con la letra griega sigma $\sum$ (cuya letra corresponde a nuestra S de suma), de esta manera la notación sigma es la siguiente:

$$\sum_{k=m}^{n} X_k = X_m + X_{m+1} + X_{m2} + \cdots + X_n$$

Así la notación se interpreta como sumatorio sobre k, desde m hasta n, de X sub- k. En donde la variable k es el índice de suma llamado límite inferior, la cual recorrerá los valores enteros hasta alcanzar el límite superior, n cumpliéndose que $m < n$.

1.2.1. Propiedades de la sumatoria

Las sumatorias son una técnica muy trabajada en la estadística pues a partir de ella, se pueden describir eventos o fenómenos de la vida diaria, de los cuales se presentan a través de modelos estadísticos que involucran variables discretas como en este caso las sumatorias. Así entonces el método de las sumatorias como su nombre lo indica se centra en la suma de términos numéricos finitos, de la cual se deben conocer todos sus datos.

De esta manera se presentan las propiedades de las sumatorias por su utilidad en campo de la estadística, donde permitirán que este método arroje resultados importantes al analizar cambios o variaciones en el estudio de una población.

- Propiedad conmutativa

$$\sum_{k=1}^{n}(X_k + Y_k) = X_1 + Y_1 + X_2 + Y_2 + \cdots + X_n + Y_n$$

$$= Y_1 + X_1 + Y_2 + X_2 + \cdots + Y_n + X_n = \sum_{k=1}^{n}(Y_k + X_k)$$

- Propiedad asociativa

$$\sum_{k=1}^{n}(X_k + Y_k) + \sum_{k=1}^{n} z_k = X_1 + Y_1 + X_2 + Y_2 + \cdots + X_n + Y_n + z_1 + z_2 + \cdots + z_n$$

$$= X_1 + X_2 + \cdots + X_n + Y_1 + Z_1 + Y_2 + Z_2 + \cdots + Y_n + Z_n$$

$$= \sum_{k=1}^{n} X_k + \sum_{k=1}^{n}(Y_k + Z_k) = \sum_{k=1}^{n}(X_k + Y_k + Z_k)$$

- Propiedad distributiva

$$a \cdot \sum_{k=1}^{n} x_k = a \cdot (x_1 + x_2 + \cdots + x_n)$$

$$= ax_1 + ax_2 + \cdots + ax_n = \sum_{k=1}^{n}(ax_k), \qquad a \in \mathbb{R}$$

1.2.2. Doble sumatoria

Las sumatorias según Churata (2016) se definen por visualizar las variables y que tipo de valores dependen de estas, es decir si existen valores dependientes, en este caso dobles, se tendrán dos sub- índices para la variable, en donde se representaría un elemento en particular de la manera a_{ij}. Lo anterior se puede ver mejor mediante arreglos rectangulares de datos, ordenados en filas y columnas, lo cual se explicará más detalladamente en el apartado §1.3

Así, teniendo un arreglo con n filas y m columnas, $A = \left(a_{ij}\right) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$

Con $i, j = 1, \dots, n$, y se quisiera sumar todos los elementos de A, se deberá primero sumar los elementos de cada fila y luego sumar los n valores resultantes, es decir que la suma de todos los elementos está dada por: $\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$. Pero si se desarrolla primero el sumatorio en j y después en sumatorio de i se obtiene lo siguiente:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} = \sum_{i=1}^{n} (a_{i1} + a_{i2} + \cdots + a_{in})$$

$$= (a_{11} + a_{i2} + \cdots + a_{1n}) + (a_{21} + a_{22} + \cdots + a_{2n}) + \cdots + (a_{n1} + a_{n2} + a_{nn})$$

Entonces si los sumatorios de i y j toman el mismo valor se deberá escribir los dos sub- índices en una sola sumatoria, indicando que tanto i como j van desde 1 hasta n, en la cual se podría escribir la doble sumatorio de una forma más compacta:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} = \sum_{i,j=1}^{n} a_{ij}$$

1.2.3. Propiedades de la doble sumatoria

Para tener un mejor manejo y entendimiento de las sumatorias dobles, se muestran sus propiedades de los sumatorios dobles:

- Propiedad conmutativa:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} + y_{ij}) = \sum_{i=1}^{n} \sum_{j=1}^{n} (y_{ij} + x_{ij})$$
$$= \sum_{j=1}^{n} \sum_{i=1}^{n} (y_{ij} + x_{ij}) = \sum_{j=1}^{n} \sum_{i=1}^{n} (x_{ij} + y_{ij})$$

- Propiedad asociativa:

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_{ij} + \sum_{j=1}^{n} y_{ij} \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} + y_{ij})$$

- Propiedad distributiva:

$$a \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} = a \cdot \sum_{i=1}^{n}(x_{i1} + x_{i2} + \cdots + x_{in})$$

$$= a \cdot [(x_{11} + x_{12} + \cdots + x_{1n}) + \cdots + (x_{n1} + x_{n2} + \cdots x_{nn})]$$

$$= (ax_{11} + ax_{12} + \cdots + ax_{1n}) + \cdots + (ax_{n1} + ax_{n2} + \cdots + ax_{nn})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} ax_{ij,} \qquad a \in \mathbb{R}$$

1.3. Matrices

El origen de las matrices Rosales (2009), se remite a los siglos II y III a.c. Donde el uso de este concepto se relaciona con el estudio de sistemas de ecuaciones lineales, y que a partir de los años 200 y 100 a.c. se crearon escritos como el libro chino “Nueve capítulos sobre el Arte de las matemáticas” el cual utiliza el método de matrices para la solución de un sistema de ecuaciones simultáneas.

Además, a lo largo de la historia el término de matriz ha sido trabajado asiduamente y se ha permitido evolucionar por los aportes de grandes matemáticos como Leibniz, quien desarrollo los determinantes en 1693 para solucionar de una manera más adecuada los sistemas de ecuaciones lineales. También Cayley, trajo sus contribuciones al introducir en 1858 la notación matricial concluyendo que un sistema de m ecuaciones lineales posee n incógnitas.

Es por ello que el concepto de matriz en su constante desarrollo ha sido un aporte relevante para los matemáticos y para la sociedad, que se desempeña en la resolución de sistemas de ecuaciones y que de esta manera, se ha logrado entender de manera intuitiva según Martínez & Sanabria (2008), como un arreglo rectangular compuesta de filas y columnas, donde se ubican números llamados componentes o elementos, y desde su estructura y propiedades pueden llevarse a cabo ciertas operaciones, que resultan útiles para el tratamiento d datos y modelaciones, un concepto primordial en el campo del algebra lineal.

Con lo anterior, una matriz puede ser expresada como $A = a_{ij}$. Cada elemento de la matriz lleva dos subíndices El primero de ellos "i", indica la fila en la que se encuentra el elemento, y el segundo, "j", la columna, es decir:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \text{ Filas de la matriz A}$$

Columnas de la matriz A

Así entonces la matriz puede tener un orden o tamaño que viene dado por el número de filas y columnas, de la cual se puede generalizar como una matriz de $m\ x\ n$, si posee m filas y n columnas.

Este concepto matemático es de gran importancia en el análisis de la economía regional, puntualmente en la econometría espacial, pues se considera una herramienta imprescindible en la aplicación y el cálculo de los datos espaciales georreferenciados. Así las matrices se deben considerar como parte fundamental de la estructura del análisis de datos espaciales, pues modelan y definirán una matriz de contigüidad que logrará el desarrollo de la correlación espacial, a cuál se profundizará más adelante.

1.3.2. Operaciones entre matrices

Las operaciones entre matrices juegan un papel importante a la hora de dar una solución a un determinado modelo, en este caso se analizaran operaciones de una matriz por un escalar, la multiplicación entre matrices y el determinante de una matriz el cual permite reducirla a un número real.

Estos procedimientos que se realizan entre las matrices resultan esenciales a la hora de estudiar en este caso una población, pues servirán de acotamiento en los datos y así poder conocer el comportamiento que puedan tener dichos datos en un análisis estadístico poblacional.

1.3.3. Multiplicación de una matriz por un escalar

Según Grossman & Flores (2012), al utilizar matrices es muy común de números reales como escalares se habla de números como escalar, para ello recalca el origen de este término en Hamilton

(1844), el cual basa su definición en el artículo *Philosophical Magazine* utilizando el concepto de *cuaternio*[1] incluía una parte real, donde puede tomar todos los valores contenidos en la escala de la progresión de números desde el infinito negativo hasta el infinito positivo.

Así una multiplicación de una matriz por un escalar está dada de la siguiente manera, si $A = (a_{ij})$ sería una matriz de $m\ x\ n$ y se incluiría un escalar α, de la siguiente manera:

$$\alpha A = (\alpha A_{ij}) = \begin{pmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{pmatrix}$$

Donde $\alpha A = (\alpha A_{ij})$ sera la matriz obtenida al multiplicar cada componente de A por el escalar α.

1.3.4. Multiplicación de matrices

La multiplicación de matrices, tiene una restricción que se expone de manera muy sencilla en Gonzales & Meneses (2009), pues dadas dos matrices A y B para calcular el producto AB, debe cumplirse que el número de las columnas de la matriz A tengan el mismo número de fila de la matriz B. obteniendo como resultado de la multiplicación una nueva matriz llamada C que tendrá el mismo número de filas de A como de Columnas de B.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{a_{i1}} & \boldsymbol{a_{i2}} & \cdots & \boldsymbol{a_{in}} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & \cdots & b_{1j} & \cdots & b_{1p} \\ b_{21} & \cdots & b_{2j} & \cdots & b_{2p} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_{n1} & \cdots & b_{nj} & \cdots & b_{np} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{i1} & \cdots & c_{ij} & \cdots & c_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{m1} & \cdots & c_{mj} & \cdots & c_{mp} \end{pmatrix}$$

Así entonces se definen las matrices, las cuales $A \in M_{mxn}$ y $B \in M_{nxp}$, donde el producto AB dará como resultado una matriz $C \in M_{mxp}$, donde el elemento C_{ij} será el producto de la fila i de A por la columna j de B:

$$C_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

[1] Los cuaternios son una extensión de los números reales generada por las unidades imaginarias i, j, k.

1.3.5. Determinante de una matriz

El determinante de una matriz, según Arce, Castillo & Gonzales (2003) es de gran importancia por su relación con otros conceptos matemáticos, en donde su utilidad permitirá reducir una matriz a un número real que puede interpretarse de varias formas: Si el número es cero el conjunto de columnas o filas de la matriz son vectores linealmente dependientes, es decir que la matriz no es invertible, si por el contrario es distinto de cero los vectores columna o fila serán linealmente independientes.

Cuando se tiene una matriz de orden n se debe generalizar la definición de los determinantes, donde se visualizan y se aplican algunas condiciones:

Esta se define por $D: M(n, \mathbb{R}) \to \mathbb{R}$ donde

- $n = 1, D(a) = a$
- Para $n > 1$ y para cualquier fila i de la matriz A, se tiene que el $D(A)$ sera la suma ponderada de n determinantes de submatrices de orden $n - 1$, es decir

$$D(A) = (-1)^{i+1} a_{i1} D(A_{i1}) + \cdots + (-1)^{i+n} a_{in} D(A_{in}) = \sum_{k=1}^{n} (-1)^{i+k} a_{ik} D(A_{kj})$$

1.3.7. Matriz inversa

Teniendo en cuenta la característica que se han presentado sobre las matrices, se centrara este aparato en un tipo de matriz como la inversa la cual es de gran aporte al estudio que se quiere realizar, para ello se estudiara un método que se aplica a sistemas de n ecuaciones lineales con n incógnitas.

Así entonces Gonzales & Meneses (2009) definen una A matriz cuadra con la condición de que existe una matriz C tal que $CA = I$, entonces C sera llamada inversa de A, de lo cual se dira que A es invertible.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix} \ y \ C = \begin{bmatrix} 7 & -2 \\ -3 & 1 \end{bmatrix}$$

$CA = \begin{bmatrix} 7 & -2 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$

De esta manera la matriz C será la inversa de A, pero se debe tener en cuenta varias características de este tipo de matriz:

- La inversa de A, si existe solo es única
- Si A y B son invertibles entonces $(AB)^{-1} = B^{-1}A^{-1}$
- Si A es invertible, A^t también es invertible.

1.3.8. Rango de matrices

Para realizar la obtención de un rango de matrices, según Aranda (2013) se debe recurrir a la aplicación de los determinantes, en el cual se deben calcular los determinantes de menor orden k de una matriz A para así poder obtener el rango al tener el mayor orden de entre los menores de A no nulos.

Entonces para calcular el menor orden k de una matriz $A \in M_{mxn}(B)$ se debe considerar el determinante de una submatriz cuadrada de A de orden k. Ya con los determinantes se puede considerar el rango de una matriz $A \in M_{mxn}(B)$ al mayor orden de entre los determinantes menores de A no nulos.

Un ejemplo que ilustra el rango de una matriz seria el siguiente, se tiene una matriz A y B las cuales son equivalentes:

$$A = \begin{pmatrix} 0 & 2 & 2 & 1 \\ -1 & 1 & 2 & 1 \\ -2 & 2 & 3 & 1 \\ 1 & 0 & 2 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 2 & 2 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = B$$

La matriz B es escalonada reducida, pero solo bastaría con la condición de que fuera escalonada, así pues la matriz posee dos filas no nulas de la cual el $Rng(B) = 2$

1.4. Correlación.

Según Vinuesa (2016), la correlación es una medida la cual presenta una relación lineal entre dos variables cuantitativas continuas (x, y). Para detectar si las variables están correlacionadas basta con determinar si existe covarianza, es decir si varían conjuntamente.

De esta manera la correlación es catalogada como una medida de asociación o de covariación lineal entre dos variables y a través del índice de correlación (r), que está acotado entre los valores $[-1,1]$, permitie diagnosticar que tipo de correlación se encuentra presente, sea positiva o negativa, es decir si ambas variables varían en un mismo sentido serán positivas, de lo contrario serán negativas cuando las variables varíen en sentidos opuestos. Ahora si un valor de $r = 0$ indicara que no existe relación lineal entre las variables.

De igual manera, se presentan rangos que nos ayudan a entender la correlación de la siguiente manera:

- Correlación baja: $|0.1| < r \leq |0.3|$
- Correlación media: $|0.3| < r \leq |0.5|$
- Correlación alta: $r > |0.5|$

Ahora, desde una perspectiva rigurosa, se debe tener en cuenta que la correlación se define mediante dos conceptos estadístico como lo son la varianza (s^2) de las variables x, y, y la covarianza ($cov(x, y)$), la cual será una medida de la variación conjunta de las dos variables.

De este modo la varianza (s^2) permitirá representar el promedio de la desviación correspondiente a los datos y su respectiva media:

$$Variancia(s^2) = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{\sum(x_i - \bar{x})(xi - \bar{x})}{N-1}$$

Ahora la covarianza $cov(x, y)$ de las dos variables será una medida promedio. Es la desviación promedio del producto cruzado entre ellas:

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Con lo anterior se definen y teniendo claridad sobre la naturaleza de las variables, la correlación pueden calcularse de diferentes formas, Moral (2013) indica que existen medidas de asociación importantes al momento de evaluar el grado de correlación que posean las variables y no solo quedarse en el comprobar si existe o no una relación entre dichas variables. Para ello se debe calcular coeficientes de correlación y evaluar la conexión de las dos variables independientes entre sí, para ello el grado de dependencia puede medirse a través de varios métodos, siendo uno de los más utilizados el de Pearson, Spearman y Kendall. El uso de cada uno de estos coeficientes varía según la estructura de los datos, pues serán tratados de forma diferente si son paramétricos y no paramétricos.

1.4.1. Coeficiente de correlación de Pearson (Datos Paramétricos)

La estadística paramétrica según Rojas (2003) emerge de una de las ramas como la estadística inferencial, el cual se caracteriza por unas condiciones en cuanto a su distribución. Es decir que se deben conocer uno o varios datos para que los parámetros y procedimientos estadísticos tenga validez y así de algún modo propicien resultados oportunos en el estudio de alguna población.

Un ejemplo muy general es que se conozca el peso de las personas y que esta siga una distribución normal, Sin embargo, que se desconozcan parámetros como la media y la desviación, dando a pasó a que se estimen estos parámetros.

De esta manera conocemos una técnica estadística paramétrica como la correlación la cual se calcula a través de la varianza y la covarianza entre ambas variables, dando lugar a que sea calculado siempre y cuando ambas variables se distribuyan normalmente:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

La correlación o grado de asociación de dos variables se mide a través del coeficiente de correlación de Pearson, el cual se denota con la letra r y su valor está comprendido entre $[1, -1]$. Según sea el valor del coeficiente de correlación r se tiene que:

- Si r es positivo la relación lineal entre las variables es directa
- Si r es negativo la relación lineal entre las variables es inversa
- Si $r = 0$ no existe relación lineal entre las variables
- Si $r = 1$ existe una relación de dependencia total directa entre las variables
- Si $r = -1$ existe una relación de dependencia total inversa entre las variables

1.4.2. Coeficiente de correlación de Spearman y de Kendall. (Datos no paramétricos).

La estadística no paramétrica según Badii, Guillen, Araiza, Cerna, Valenzuela & Landeros (2012) se utiliza como alternativa para obtener resultados de una distribución libre, en este caso las pruebas no paramétricas son aquellas que no recurren a los supuestos de una distribución sobre los parámetros de la población. Así los datos más utilizados en esta prueba son los nominales y ordinales.

Estas pruebas paramétricas resultan importantes en el estudio de rangos o diferencias entre observaciones, pues se pueden a través de los datos asignarse por características como lo socio económico, textura, sabor o género.

El coeficiente de correlación de Spearman según Cabrera (2009), es una medida de asociación lineal en la cual se utilizan los rangos y se ordenan los hechos para cada variable, en el caso de existir dicha asociación lineal, se observaría que el rango de la variable x sería el mismo valor del rango de la variable y, es decir que el coeficiente se calcula a través de la diferencias entre los rangos de ambas variables, pretendiendo que esta diferencia sea 0.

De igual manera si las diferencias de las variables en estudio se alejan la correlación dejaría de ser perfecta, y para que estas diferencias positivas no se involucren con las negativas se utilizaría un estadístico como el siguiente:

$$r_s = 1 - \frac{6 \sum {d^2}_i}{n(n^2 - 1)}$$

Mientras el coeficiente de correlación de Kendall parte de no ser paramétrico, pues no cumple con el supuesto de una distribución normal de las variables a comparar, este tipo de coeficiente se utiliza cuando se tiene una gran variedad de datos dentro de un mismo rango, en el cual el signo del coeficiente indicara que dirección tomara la relación y el valor absoluto indicara la magnitud de esta misma.

$$r = \frac{(s_a - s_b)}{\frac{n(n-1)}{2}}$$

Partiendo de las anteriores métricas, como una forma de entender la covariación entre un conjunto de datos, se valida mediante un solo número, pues nos muestra el carácter de la relación, positiva o negativa y si es fuerte o débil, sin embargo, también se puede visualizar mediante otras representaciones que funcionan de manera intuitiva para realizar análisis y algunos pronósticos. Por eso es importante mostrar la más importante de ellas.

1.4.3. Diagrama de dispersión

Los diagramas de dispersión se muestran en Altamirano & Espinoza (2009) de gran utilidad cuando se tienen estudios de muchas observaciones, en la cuales se da la relación existente entre 2 variables (x, y), así la correlación será positiva cuando el valor de x aumenté proporcionalmente al valor de y, pero por el contrario si aumenta el valor de x disminuirá en igual proporción el valor de y dando una correlación negativa.

En este sentido Moore (2002) ilustra el tipo de diagrama que se pueden presentar en el estudio de las variables x e y.

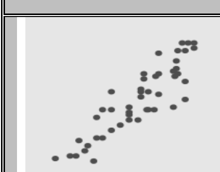
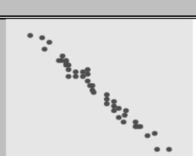
Diagrama	Relación
	No hay relación entre las variables
	Baja correlación positiva, el valor de x aumenta ligeramente, mientras aumenta el valor de y
	Débil correlación negativa, el valor de x dismunuye mientras el valor de y aunmenta.
	Alta correlación positiva, el valor de y se incrementa mientras el valor de x aumenta.
	Fuerte correlación negativa, el valor de x disminuye mientras aumenta el valor de y

Tabla 1. Gráficos de correlación según su dependencia. Fuente: Análisis de relaciones. Moore (2002)

1.5. Análisis de regresión.

El análisis de regresión es una técnica estadística utilizada para estudiar la relación entre variables, este método es uno de los más utilizados y aplicados por su adaptabilidad a una amplia variedad de situaciones.

En este sentido Berenson, Levine & Krehbiel (2006) detallan que características posee un análisis de regresión, del cual se encuentran inmersas la variable dependiente y la variable independiente, estas dos variables juegan un papel importante al desarrollar un modelo de regresión donde se quiera predecir algunos valores, pues la dependiente es la variable que se desea predecir y la independiente se utilizara para hacer dichas predicciones. Sin embargo este análisis también permite identificar otros factores como la relación matemática que existe entre las variables y como estas varían.

Según lo anterior, el análisis de regresión puede ser llevado a cualquier contexto, en este caso a las ciencias sociales, pues dentro de su naturaleza esta predecir algunos tipos de fenómenos relacionados con la sociedad.

1.5.1. Regresión lineal

El modelo de regresión lineal es el más básico de todos y parte por identificar cual es la variable dependiente y cuál es la variable independiente. Este modelo establece que y como variable independiente, es una función de solo una variable independiente $x,$ por tal razón se representa de esta manera: $y = f(x)$. Estuardo (2012), aclara que un análisis de regresión lineal es utilizado para fines de predicción, en el cual se busca hallar una ecuación a relaciones que se den entre dos o más variables, Así el primer paso para obtener una ecuación que relacione las variables de estudio es clasificar y ordenar los datos que cumplan las consideraciones de las variables correspondientes.

A partir de un diagrama de dispersión se puede determinar una ecuación en la cual se interpreten dichos datos, para ello se involucra un ajuste curvas que permite estimar una de las variables (dependiente) conociendo la otra variable (independiente), dejando así visualizar el proceso de regresión.

Así entonces Walpole, Myers, Myers & Ye (2007) estudian la ecuación de una regresión lineal, enfocándose en la predicción que posea dicha línea. Para ello se utilizará la siguiente ecuación.

Una condición de esta regresión es que el valor predicho de Y es igual a la intersección en Y más la pendiente multiplicada por el valor de X

$$\hat{y}_i = b_0 + b_1 X_i \ \ donde$$

- $\hat{y}_i$ = valor de Y para la observación i
- X_i= valor de X para la observación i
- b_0= intersección de la muestra y
- b_1= Pendiente de la muestra

Este tipo de ecuación lineal tiene dos coeficientes de regresión como lo son b_0 y b_1 para poder encontrar estos coeficientes se debe remitir a un método como el de los mínimos cuadrados. Este método minimizara la suma del cuadrado en cuanto a las diferencias entre los valores de y_i y los valores de $\hat{y}_i$

Así al realizar un análisis de regresión se debe tener en cuenta la línea recta que más se ajuste a los datos, para así lograr que la diferencia entre los valores de y y los valores estimados de $\hat{y}$ sean lo más pequeñas posible.

Es decir que la recta de mínimos cuadrados que más se aproxima al conjunto de pares ordenados como $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ tiene por ecuación la recta $y = b_0 + b_1 x$ en el cual las constantes a y b se determinan al resolver las siguientes ecuaciones (normales):

$\sum y = b_0 n + b_1 \sum x$ y $\sum xy = b_0 \sum x + b_1 \sum x^2$

Otra forma de determinar las constantes a y b que se deducen del sistema de ecuaciones anteriores es la siguiente:

$$b_0 = \frac{n \sum xy \ \sum x \sum y}{n \sum x^2 \ (\sum x)^2} \quad , \quad b_0 = \hat{y} \quad , \quad b_1 \widehat{x}$$

En donde $\hat{x}$ y $\hat{y}$ son los promedios estimados de los correspondientes a x e y.

En el caso de que la variable x sea dependiente y la variable y sea la independiente se tendría la ecuación de la recta como $x = b_0 + b_1 y$ y el sistema de ecuación seria el siguiente:

$$\sum x = b_0 n + b_1 \sum y \ , \ \ \sum xy = b_0 \sum y + b_1 \sum y^2$$

Obteniendo

$$b_1 = \frac{n \sum xy \ \sum x \sum y}{n \sum y^2 \ (\sum y)^2} \qquad , \qquad b_0 = \hat{x} \quad , \ b_1 \hat{y}$$

De la cual la recta de mínimos cuadrados se diferencia de la ecuación obtenida anteriormente.

1.5.2. Análisis de residuo (error).

Este tipo de análisis sirve para probar si el modelo lineal es el adecuado en relación con los datos en estudio, para ello se tiene como ecuación $y = b_0 + b_1 x + e$ en la cual la variable x es llamada independiente o de predicción y la otra variable y dependiente o de respuesta:

- b_0, e s el valor de la ordenada cuando se intercepta con el eje y
- b_1, es el coeficiente de regresión poblacional (pendiente de la recta)
- e, es el error

En el caso de e se entiende como esa diferencian entre el valor observado y y el valor estimado $\hat{y}$ obteniendo la siguiente ecuación:

$$e = y - \hat{y}$$

En donde permitirá establecer conclusiones, si la función de regresión es lineal o al contrario si no lo es, además de comprobar si el modelo de regresión se ajusta a todos los datos o si por el contrario existen una o varias observaciones atípicas, teniendo en cuenta que si el número de datos es mayor a (30) no se van a considerar dichas observaciones.

1.5.3. Medidas de variación en una regresión lineal.

Las medidas de variación en una regresión lineal serán la suma total de los cuadrados, es decir que este método de mínimos cuadrados permitirá expresar la variación total de las y. Para ello se necesitan calcular tres medidas de variación. La primera es la suma total de cuadrados (SST, por sus siglas en inglés) la segunda la suma de los cuadrados de la regresión (SSR) y por último la suma de los cuadrados del error (SSE).

Estas medidas de variación permitirán tener una igualdad interesante como que, la suma total de los cuadrados va hacer = a la suma de los cuadrados de la regresión + el error de la suma de los cuadrados.

- **Suma Total de Cuadrados (SST)**

La suma total de cuadrados es igual a la suma del cuadrado de las diferencias entre y y $\bar{y}$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **Suma de Cuadrados de la Regresión (SSR)**

La suma de cuadrados es igual a la suma del cuadrado de las diferencias entre el valor predicho de y y $\bar{y}$.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- **Error de la suma de cuadrados (SSE)**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\hat{y}_i = b_0 + b_1 X_i$

1.5.4. Coeficiente de determinación (R^2).

El coeficiente según Steel & Torrie (1996) denominado R^2 es un estadístico que se ajusta a un modelo del cual se pretende explicar o predecir futuros resultados, además de proporcionar la varianza total de la variable obtenida por la regresión. Este coeficiente R cuadrado tiene una particularidad cuando se refiere a una regresión lineal pues será el cuadrado del coeficiente de correlación de Pearson solo en una regresión lineal simple.

Las características de acuerdo a Laguna (2014) del coeficiente R cuadrado, es que primero es una cantidad adimensional que solo puede tomar valores en $[0,1]$, luego se debe analizar cuando un ajuste es bueno en R^2, si este toma un valor cercano a 1 se podrá decir que existe una asociación entre ambas variables, si por el contrario el ajuste no sea tan conveniente R^2 será cercano 0 en donde no habrá asociación entre x y y.

Así de acuerdo a las medidas de variación anteriormente mencionadas como SSR, SSE y SST se utilizarán la suma de los cuadrados de la regresión sobre la suma total de los cuadrados. La cual proporcionara la variación en y que se explica por la variable x del modelo de regresión, es decir que a esta nueva razón se le llamara coeficiente de determinación r^2 dada de la siguiente manera:

$$r^2 = \frac{suma\ de\ cuadrados\ de\ regresion}{suma\ total\ de\ cuadrados}$$

1.5.5. Regresión lineal múltiple.

La regresión múltiple se caracteriza por el estudio de la relación entre variables independientes y otra variable dependiente, representa por un modelo general como $y = f(x, w, z)$. Así las variables representadas mediante una función lineal, no serán ahora una recta como en el caso de la regresión lineal simple si no un plano o un hiperplano dependiendo del número de variables independientes.

Uriel (2013), establece que un análisis de regresión lineal simple no es el más apropiado a la hora de modelar fenómenos, ya que no puede explicar factores que intervienen en una variable económica. Es así como la regresión múltiple pretende modelar la relación existente entre variables con el propositito de poder pronosticar una de ellas, la variable dependiente a partir del conocimiento de las otras variables independientes, para ello sería de la siguiente manera $y = b_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n + e$ donde los parámetros $b_1 + b_2 + b_3 + \cdots + b_n$ son constantes y desconocidos.

Las variables independientes o explicativas son $x_1,\ x_2, \dots, x_n$ que requieren de un componente como e que sería el error para dar respuesta a la variable dependiente.

Para comprender el cálculo de estos términos, las formas para hacerlo requieren de matrices, lo cual se complejiza al aumentar el número de variables explicativas. Es por esto que el sistema debe formarse de una manera de arreglos rectangulares, no obstante la lógica expresada en §1.5.4 es la misma, encontrar los parámetros y el término de error.

Este sistema de ecuaciones puede expresar de una forma más compacta utilizando la notación matricial:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{21} & x_{31} & \dots & x_{n1} \\ 1 & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{2n} & x_{3n} & \dots & x_{nk} \end{pmatrix} \ \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix} \ \boldsymbol{e} = \begin{pmatrix} \boldsymbol{u_1} \\ \boldsymbol{u_2} \\ \dots \\ \boldsymbol{u_n} \end{pmatrix}$$

La matriz X es la matriz conformada por las variables explicativas o regresores, y esta toma el valor de “1” para todas las observaciones, así el modelo de regresión múltiple quería expresado en forma matricial de la siguiente manera:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{21} & x_{31} & \dots & x_{n1} \\ 1 & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{2n} & x_{3n} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}$$

Así el modelo regresión lineal múltiple puede ser expresado por $y = Xb + e,$ donde y es un vector $n\ x\ 1$, X es una matriz $n\ x\ k$, b es un vector $k\ x\ 1$ y e es un vector $n\ x\ 1$.

Con la estructura mostrada, se establece el estudio de la regresión a partir de los parámetros $b_1 + b_2 + b_3 + \dots + b_n$, provenientes de una muestra dada, logrando un modelo estimado mutivariable de la forma $\hat{y}_1 = \hat{b}_1 + \hat{b}_2 x_{2i} + \hat{b}_3 x_{3i} + \dots + \hat{b}_n x_{ni}$. Esto permite calcular el valor ajustado de $(\hat{y}_i)$ correspondiente a y_i. En donde $\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_n$ son los estimadores de $b_1 + b_2 + b_3 + \dots + b_n$.

En este caso el residuo será la diferencia entre y_i e $(\hat{y}_i)$, el cual será la diferencia entre el valor muestral y su correspondiente valor ajustado:

$$\hat{u}_1 = y_1 - \hat{y}_1 = y_1 - \hat{b}_1 - \hat{b}_2 x_{2i} - \hat{b}_3 x_{3i} - \cdots - \hat{b}_n x_{ni}$$

El sistema de ecuaciones se denotará expresando la notación matricial

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{pmatrix} \quad \hat{b} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \cdots \\ \hat{b}_n \end{pmatrix} \; \hat{u} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \cdots \\ \hat{u}_n \end{pmatrix}$$

Entonces el modelo ajustado para todas las observaciones de la muestra estará dado por $\hat{y} = X\hat{b}$ y el vector de los residuos será igual a la diferencia entre los valores de los vectores observados y ajustados por $\hat{u} = y - \hat{y} = y - Xb$, concluyendo que lo analizado para un modelo básico de regresión lineal es también análogo a este caso.

Capítulo 2. Estructura del análisis de datos espaciales.

2.1. Introducción

Este capítulo tiene como finalidad mostrar lo más relevante del marco teórico del análisis de datos espaciales, desarrollos como la matriz de contigüidad, el índice de morán, modelos de regresión con rezago espacial que permiten comprender a fondo las interacciones de algunas variables socio económicas teniendo en cuenta las métricas del *espacio*, asumiendo este último como una variable explicativa más en un sistema a modelar. Es claro que el capítulo anterior fue una antesala para comprender como la matemática y la estadística permean los resultados y hechos estilizados utilizados por el análisis de datos espaciales y por tal razón, este apartado busca con la divulgación de los trabajos de Luc Anselin principalmente, mostrar la estrecha relación entre estas dos áreas del conocimiento y los resultados que en su momento innovaron la discusión en las ciencias sociales.

2.2. Principios básicos en el tratamiento de los datos espaciales.

El análisis de datos espaciales, tiene un énfasis marcado en el estudio de los componentes del espacio, explicando formalmente sus elementos constitutivos y la manera como éstos se comportan bajo ciertas condiciones. Para ello se vale de un conjunto de herramientas provistas por la *econometría espacial*, que se sirve elementos de la estadística, algebra y del cálculo para evaluar la presencia de relaciones, contrastes y efectos en la modelación de fenómenos económicos y sociales. Paelinck y Klaassen (1979), en su trabajo procuran describir estas técnicas fundamentando las diferentes posibilidades a partir de cinco principios básicos: La *Interdependencia* que hace referencia a que todo modelo espacial debe incorporar relaciones mutuas entre las observaciones de las variables económicas, sociales, demográficas, la *Asimetría* que postula las relaciones espaciales, son en principio caóticas y muy raramente métricas, la *alotopía* que busca “a priori” la causa de un mismo fenómeno espacial en otro lugar; la ***No linealidad*** que argumenta que las posibles relaciones no pueden ser tan simple para explicarse por auto regresiones lineales y por último *la inclusión de variables topológicas*, piedra angular del análisis de datos espaciales pues defiende la inclusión de variable que antes no se consideraban necesarias y so las coordenadas, distancias, superficies, densidades, entre otras.

2.3. Dependencia espacial

La dependencia espacial es un concepto necesario para comprender la estructura del análisis de datos espaciales que se presenta en la econometría espacial, es por esto que Anselin (1998) y Goodchild (1987) dedican el estudio de las relaciones que se pueden presentar en el espacio, en virtud de que una variable puede tomar un valor en una unidad geográfica y esta a su vez puede estar condicionada por el valor que tome está en regiones próximas, y para ello se desarrollan una serie de representación que se valen de matrices para explicar y poder visualizar la relación que se puede presentar entre cada una de las regiones espaciales puestas en estudio.

Es por esto que al aplicar lo anterior, puede presentarse el caso de algunos efectos provocados por el mismo espacio que provoquen variaciones que se puedan ser significativos a la hora de intentar modelar una situación que involucren el concepto de posición, espacio y distancia, y los más conocidos son la heterogeneidad espacial y la auto correlación espacial. La primera se encarga de analizar la variación de las relaciones que se obtienen en el espacio, y se distinguen dos clases de heterogeneidad: *la inestabilidad estructural* presentada por la falta de estabilidad en el espacio, en el cual el comportamiento de la variable bajo estudio y los parámetros varían según la localización, generando variables diferenciadas en todo el conjunto de datos y *la heterocedasticidad* que se manifiesta por la omisión de variables u otras formas de errores de especificación que conllevan a la aparición de errores de medida.

No obstante, existen varios métodos que pueden dar solución a este tipo de efectos que podrían obstaculizar el análisis que se está llevando a cabo, por lo que debe darse una cierta a prioridad a lo que se conoce como *auto correlación espacial*.

2.3.1. Autocorrelación espacial

La dependencia o autocorrelación espacial en Moreno & Vaya (2000) se explica de una manera bastante sencilla, pues cuando se toma un evento en una región cualquiera y esta repercute o se relaciona con lo que ocurre en otra región del espacio, se está mostrando la existencia de patrones que influyen por una consecuencia de la localización geográfica y la naturaleza de la variable. Es decir que una variable puede comportarse y cuantificarse en específico en una zona, pero de alguna manera estará afectada por el valor que tome esa variable en otras regiones distantes o vecinas y no solamente por las condiciones internas. De una manera mucho más básica, este concepto puede verse cuando una zona que tiene un promedio alto de salarios y está rodeado a su vez de zonas con el mismo comportamiento, logrando una convergencia o agrupación de zonas con correlación positiva en cuanto a los salarios. Generalmente esto puede darse por la una escasa correspondencia entre la extensión espacial del fenómeno que se encuentra en estudio y las unidades espaciales de observación, ya que las relaciones de este tipo pueden ser multidireccionales y para su mejor entendimiento puede evidenciarse de manera más clara a través de una matriz de contigüidad y pesos espaciales.

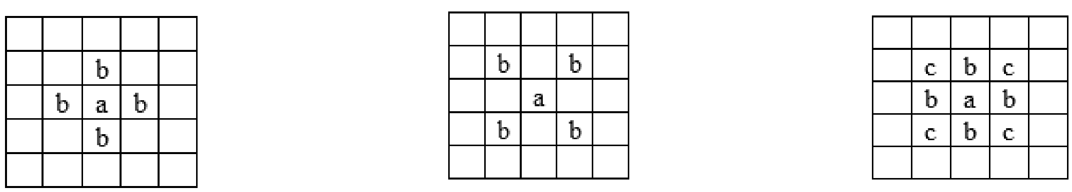
2.3.2. Matriz de contigüidad

La matriz de contigüidad es uno de los pilares que sostiene el análisis de datos espaciales, pues en su artículo Spatial Econometrics (1988) Luc Anselin introduce que existen localizaciones que son influenciadas por otras posiciones geográficas, estas últimas las etiqueta como vecinos del espacio con el fin de determinar que unidades en el sistema espacial tienen influencia sobre una unidad en particular bajo la misma consideración, expresando así la existencia de contigüidad entre regiones y por consiguiente correlación espacial.

Teniendo en cuenta el conjunto de vecinos que pueden ser hallados a través de las unidades espaciales, es decir que tan relacionados están por cercanía o lejanía, se propone un arreglo matricial simétrico notado como W, en la cual cada una de las filas y de las columnas representa una región en el espacio, en la cual se manifiesta la relación que tienen cada una de las regiones con las demás regiones del espacio, tal como se vería en un mapa. De esta manera la matriz de contigüidad puede ser construida a través de la notación binaria donde 1 representa la presencia de contigüidad espacial entre dos unidades y 0 la ausencia de contigüidad espacial entre dos unidades. Teniendo presente que pueden presentarse infinidad de configuraciones de las regiones en el espacio y existen varias formas para definir esta presencia o ausencia de contigüidad espacial, las más representativas son:

Criterio de vecindad	Número de vecinos	Definición
Criterio lineal	2	Serán vecinas de i las regiones que comparten el lado izquierdo o derecho de i
Criterio torre	4	Serán vecinas de i las regiones que comparten algún lado con i
Criterio alfil	4	Serán vecinas de i las regiones que comparten algún vértice con i
Criterio reina	8	Serán vecinas de i las regiones que comparten algún lado o vértice con i

Tabla 2. Técnicas econométricas para el tratamiento de datos espaciales. Fuente: Moreno y Vaya (2000).



Contigüidad Torre **Contigüidad del Alfil** **Contigüidad Reina**

Figura 1. Configuración de matrices de contigüidad Fuente: Acevedo y Velásquez (2008).

De esta manera pueden resultar más formas que permitan procesar la matriz contigüidad, se debe tener en cuenta que dicha matriz se debe ajustar al problema de estudio y puede resultar que la matriz de contigüidad binaria no provee una buena especificación, por lo cual es necesario tener en mente otras posibles matrices basadas en la similaridad o disimilaridad existente entre datos y características inherentes a las variables en estudio, para esto se debe analizar la matriz de pesos espaciales.

2.3.3. Matriz de pesos espaciales

La matriz de Pesos espaciales o también llamada W tiene una estrecha relación con el concepto de autocorrelación espacial, pues se basan en la distancia o alejamiento entre regiones, es por ello que autores como Cliff & Ord (1981) y Anselin (1980) definen una matriz W de la siguiente manera:

$$w_{ij} = d_{ij}^{-a} \beta_{ij}^{b}$$

Donde W dispone de d_{ij} como la distancia entre las unidades i y j , luego β_{ij} será la distancia entre las fronteras comunes entre i y j teniendo en cuenta el perímetro de i, de esta manera a y b serán los parámetros a estimar.

Ya conociendo la definición de W se da paso a que se implemente en los estudios que requieran conocer la distancia entre regiones, estandarizando dicha matriz W de tal forma que dividen cada elemento w_{ij} por la suma total de la fila a la que pertenece, dejando que la suma de cada fila sea igual a la unidad.

2.4. Análisis exploratorio de los datos espaciales y la visualización de datos espaciales.

El análisis exploratorio tiene gran incidencia en el estudio de patrones y asociaciones de datos, pues en una gran variedad de investigaciones espaciales, se recurre a esta técnica por su manera de contrastar y generalizar datos espaciales. Es por ello que autores como Tukey (1977), Good (1983) y Cleveland (1993) impulsaron a que el análisis exploratorio de datos tuviera una incidencia en la estadística aplicada y que fuera un elemento indispensable a la hora de analizar y visualizar un estudio sobre datos espaciales.

De esta manera Anselin (1996) reconoce la importancia de los datos espaciales y clasifica unas técnicas para dicho estudio, una de ellas son los indicadores globales y locales de asociación espacial y la otra los estadísticos basados en la vecindad y la distancia.

Dichos indicadores globales resultan ser la forma más tradicional al efecto de dependencia espacial, pues se resume a un indicador único, en este caso el índice de Moran es un claro ejemplo, y para los indicadores locales (LISA) es determinar si una región se encuentra rodeada por otros valores altos o bajos, dejando ver la ausencia de autocorrelación espacial global.

Teniendo en cuenta dichas técnicas del análisis exploratorio de datos, Moreno & Vaya (2000) enfatizan la importancia de involucrar la visualización en las distribuciones espaciales, pues se tiene la capacidad de interactuar y construir una percepción de los datos que se pretenden analizar. Resaltando que en el estudio y en el análisis de datos estadísticos se involucran herramientas que, por medio de mapeos o gráficos dinámicos, permiten tener la información requerida a partir de la localización y de aspectos relevantes lo cuales rodean dichos valores, es posible hablar de cómo se pueden analizar índice mediante ciertas técnicas establecidas. Un ejemplo es como la identificación de valores atípicos espaciales se da a través de cuartiles, ya que se puedan percibir los valores atípicos inferiores y posteriores, que son visibles a través de las observaciones que están fuera del límite de un diagrama de cajas, otro sería el ya mencionado grafico de correlación que muestra las zonas.

Por esta razón un punto de partida es el mapeo cuartil que permitirá obtener los valores atípicos y además contando con la relación de los diagramas de cajas, darán como resultado que tipo de valores en las observaciones se tienen para así asociarlas con las variables que correlacionen, es decir este tipo de análisis permite seleccionar las variables que se deben presentar en un modelo donde no se tienen fundamentos a priori.

De esta manera se implementa un sistema de información geográfica capaz de combinar la localización, visualización, y exploración de datos geográficas, permitiendo así dar la información requerida y eficaz de las variables de interés, esto se realizaría en el marco del *análisis exploratorio de datos espaciales*, pues permite describir las distribuciones de una manera dinámica y simultánea a través de un mapeo ya sea un histograma o un diagrama de cajas en el cual se pueden visualizar y trabajar en subgrupos los datos espaciales para poder ser comparados en una distribución.

Es por eso que se proponen unas técnicas o métodos que se pueden utilizar para analizar una distribución espacial ya sea desde un punto de vista global o local y para ello se toma un esquema de Yrigoyen (2003).

Distribución espacial	Univariante	• Diagrama /Mapa de caja.
	Multivariante	• Diagrama dispersión/ Caja.
Asociación espacial	Global	• Mapa de contigüidades espaciales. • Grafico del retardo espacial. • Diagrama mapa de dispersión de Moran.
	Local	• Puntos atípicos en el diagrama de dispersión de Moran. • Mapa LISA. • Diagrama de caja LISA.
	Multivariante	• Diagrama de dispersión Multivariante de Moran.
Heterogeneidad espacial	• Mapa del histograma de frecuencia. • Diagrama de dispersión.	

Tabla 3. Métodos Gráficos del Análisis Exploratorio de Datos Espaciales. Fuente: Yrigoyen (2003).

2.4.1. El contraste de Autocorrelación Espacial Global y el índice de Morán.

La necesidad de encontrar un punto de vista global del fenómeno de auto correlación espacial, se motiva por la presencia de estructuras espaciales, es decir, características que se modelan a partir de la distribución de una variable sobre un espacio geográfico dado por interacciones de otras zonas contiguas a él. En términos más precisos, se trata de corroborar la hipótesis de que una variable que se encuentre distribuida de forma totalmente aleatoria en un sistema espacial o, verificar si existe algún tipo de asociación significativa de valores similares o distintos entre regiones vecinas.

Para tal fin, se han propuesto varias alternativas estadísticas de dependencia espacial, pero definitivamente el más utilizado y en el que se centra este apartado, es el índice y test de Moran.

Este índice es utilizado como una prueba de diagnóstico, siendo presentado por Moran (1948) y consolidándose como una de las pruebas más utilizadas para conocer si existe o no auto correlación entre las unidades espaciales, además permite realizar un análisis espacial en cuanto a los valores que se obtienen utilizando un diagrama de dispersión, pues si los valores se precipitan estos se agruparan de manera espacial y serán positivos, por el contrario si dichos valores se dispersan estos valores serán negativos, y además se podrán obtener cuales localizaciones posean valores muy altos respecto a regiones vecinas, para así poder dar paso analizar variables de interés.

Cabe aclarar que el índice de Moran permitirá estudiar la presencia o ausencia de la dependencia espacial univariante, es decir si una y solo una variable se encuentra distribuida de forma aleatoria en el espacio o si por el contrario existen valores similares o disimiles entre regiones vecinas. Para consolidar la estructura del test, se utiliza el edificio matemático teniendo como resultado la siguiente expresión:

$$I = \frac{N \sum_{ij}^{N} w_{ij} \, (X_i - \bar{X})(X_j - \bar{X})}{S_0 \sum_{i=1}^{N} (X_i - \bar{X}\,)^2} \; i \neq j$$

Donde:

- w_{ij} es la posición correspondiente de cada elemento (i, j) de la matriz de contigüidad
- S_0 es la suma de los pesos espaciales $S_0 = \sum_{ij}^{N} w_{ij}$
- $\bar{X}$ es el valor esperado de la variable X
- N es el número de regiones.

Si se requiere utilizar la matriz de contigüidades estandarizadas, el I de Moran cambiaria de la siguiente manera:

$$I = \frac{N \sum_{ij}^{N} w_{ij} \, (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X}\,)^2} \; i \neq j$$

En el cual la suma de los elementos de cada fila dará como resultado la unidad, es decir que $\sum_{ij}^{N} w_{ij}$ será N.

Como parte del resultado de este índice, también se construye un diagrama de dispersión de Moran en el que se representa con una línea de regresión, cuya pendiente, en este caso, será el valor del test *I* de Moran que, por este motivo, puede ser utilizado como indicador del grado de ajuste, así como de la presencia de valores atípicos en la nube de puntos. Esta ayuda visual funciona como el índice de correlación visto en el apartado § 1.4 y se relaciona de manera muy fuerte con la posible covariación de la variable dependiente e independiente.

Por último, tampoco debe olvidarse la importancia que tiene la matriz de pesos espaciales en el proceso de contraste del fenómeno de autocorrelación espacial global, pues está bien comprobado que los resultados obtenidos de *I* pueden variar, a veces de forma sensible, en función de la matriz *W* especificada.

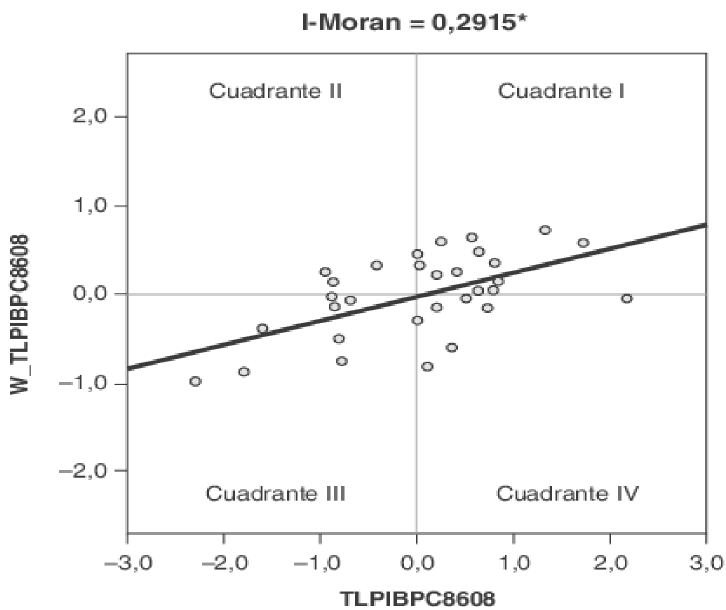


Figura 2. Diagrama de dispersión de Moran para tasa de crecimiento. Fuente: Asuad & Quintana (2010)

2.4.2. Contrastes de Autocorrelación Espacial Local y los Indicadores Locales de Asociación Espacial (LISA)

En la autocorrelación global, los resultados están centrados en el análisis de dependencia general propia de todas las unidades de un espacio geográfico y no son capaces de detectar otros aspectos o algunas estructuras locales de asociación entre las variables como *clusters* (resultados donde se asocian o convergen puntos con mismas características, altos con altos, bajos con bajos), *outliers* (resultados donde difieren las variables, altos con bajos) e inestabilidades locales que pueden no estar presentes en una estructura global de dependencia. Por esto, Getis & Ord (1992), ahondan en este problema de la dependencia espacial local desde la posibilidad de que, en un espacio dado, no se detecte la presencia de autocorrelación espacial global en la distribución de una variable, aunque existan pequeños clusters espaciales en los que dicha variable experimenta una concentración importante o que habiéndose detectado dependencia a nivel global en una variable, no todas las regiones del espacio considerado contribuyan con igual peso en el indicador global, es decir, que coexistan unas zonas en las que la variable se distribuya de forma aleatoria junto a otras con una importante contribución a la dependencia existente.

Para dar respuesta a estos planteamientos, Anselin (1995), definen unos contrastes de asociación local espacial que indican hasta qué punto una región se encuentra rodeada por otras con valores altos o bajos de una variable determinada y parten de la hipótesis nula de ausencia de autocorrelación espacial global. Es comúnmente conocido por sus siglas en inglés, LISA *("Local Indicators of Spatial Association"),* capaces de detectar la contribución de cada región a un indicador de dependencia espacial global. Este tipo de indicadores permiten el reconocimiento de inestabilidad espacial, es decir, la presencia de valores atípicos que también pueden ser visualizados mediante diagramas de dispersión y su forma matemática está muy relacionada al índice global *I* de Moran visto en el apartado anterior.

$$I = \frac{xi - \bar{x}}{s^2} \sum_{j-1} wij(xj - \bar{x})$$

Con este resultado se obtiene una serie de indicadores locales que puede tomar cualquier valor, por ejemplo, un valor positivo y estadísticamente significativo indica la presencia de un clúster (una zona con características similares a sus zonas vecinas). A la inversa, un signo negativo sugiere que un distrito tiene valores diferentes a sus vecinos, es decir, se trata de un outliers. La interpretación del estadístico local *I* de Moran como un indicador de inestabilidad local se desprende fácilmente de la relación entre estadísticos locales y globales. En concreto, la media de *I* local será igual a la del estadístico global *I* por un factor de proporcionalidad. Las máximas contribuciones de los valores de *I* local al estadístico global *I* pueden ser identificadas a través de criterios sencillos, ya sea por un diagrama de dispersión del *I* de Moran (Scatterplot de Moran) o mapas de las zonas gracias a los paquetes de computación avanzados.

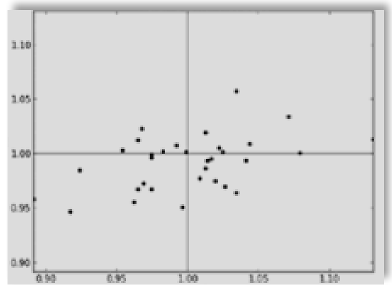


Figura 3. Ausencia de dependencia espacial. Fuente: elaboración propia

2.5. Análisis confirmatorio.

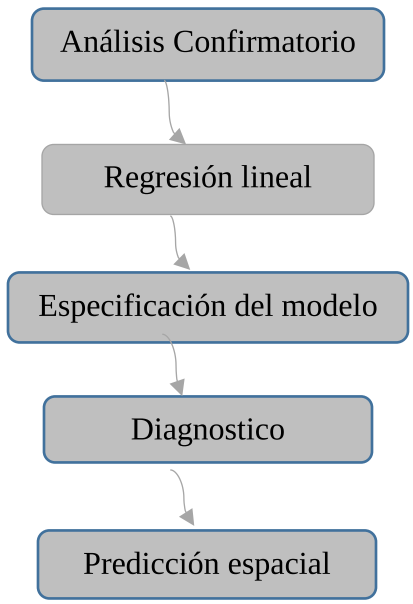
Este tipo de análisis *Olivares* (2002), lo caracteriza desde una perspectiva netamente matemática y estadística, proponiendo el estudio y análisis de datos que son puestos en juego a partir de un modelo defino a priori, y que se enfoca en comprobar o rechazar una hipótesis, para ello se utilizan o se emplean indicadores estadísticos como la media, la varianza, los coeficientes de correlación, coeficientes de regresión y por supuesto las pruebas de hipótesis. Llegando así, hacer un modelo de análisis de datos muy efectivo para ser empleado en las ciencias sociales.

Ahora bien, ya un poco más enfocado en los datos espaciales, *Moreno & Vaya* (2000) pretende que este análisis se maneje desde una perspectiva de modelización, en donde involucra distintos métodos de estimación, contrastes de especificación y procedimientos de validación. Para así poder llevar acabo modelos multivariantes que posean observaciones de corte transversal y que de igual manera estén georreferenciados.

Por lo tanto, cuando se incluye estos tipos de modelos multivariantes se debe tener en cuenta cuales son las características principales que debe poseer como la selección de las variables y como estas van hacer relacionadas a partir de una manera funcional. Todo esto con el fin de poder estructurar un análisis adecuado, pero sin embargo cuando no se sabe qué tipo de modelo utilizar por no tener una fuente teórica a priori que oriente dicho modelo se puede recurrir como punto de partida un análisis exploratorio de datos.

Sin involucrar ningún tipo de efecto espacial, ya que los resultados en este caso los residuos servirán como referencia a la hora de diagnosticar una dependencia espacial. De todas maneras, si se diagnostica una autocorrelación residual el paso a seguir será utilizar un término de error que se incorpore en la misma estructura de la dependencia espacial, es decir que la estimación y los diagnósticos de validación de un modelo juegan un papel importante en la búsqueda de seleccionar el modelo más adecuado.

Es por ello que se presentara un esquema de las etapas básicas que se deben seguir a la hora de realizar un análisis de datos espacial confirmatorio:



Flujograma de procedimiento. Fuente: Moreno y Vaya (2000)

2.5.1. Modelos de regresión con rezago espacial

Este tipo de modelos con rezago espacial poseen una particularidad y es ya haber estudiado la dependencia espacial dando a paso a que se analice a través de modelos de regresión espacial en el cual se puedan representar y añadir la dependencia espacial.

Según Pineda (2006) la dependencia especial que se presenta en los modelos de regresión tiene una consecuencia de autocorrelación entre las variables ya sea dependientes e independientes y además por la aparición de un esquema espacial en las perturbaciones. Estas variantes que se presenta en la dependencia espacial se pueden introducir en los modelos de regresión espacial a través de matrices de pesos como:

- Variables dependientes espacialmente rezagadas: Wy
- Variables explicativas espacialmente rezagadas: Wx
- Términos de error espacialmente rezagados: Wu

La elección del tipo de modelo espacial que se desea utilizar para contrarrestar los efectos detectados viene seguido de las variantes más usuales de dependencia espacial como residual y sustantiva.

Estos tipos de dependencia espacial tienen una característica importante considerada en los procesos espaciales como las formas autorregresivas las cuales permitirán captar el efecto multidireccional de las variables espaciales. Para ello se plantea un modelo espacial de primer orden como:

$$Y = pWy + u$$

- y y u son vectores de variables y el termino de error
- W la matriz de pesos espaciales
- pWy es la estructura autorregresiva en la variable dependiente que intenta explicar el modelo (rezago espacial).

Así los modelos espaciales más usuales tienen como característica, una estructura lineal autorregresiva en donde se identifican dos tipos de modelos de regresión como el de datos de corte trasversal y datos que combina tiempo y espacio.

Capítulo 3. Aplicación de una situación de análisis de datos espaciales y modelación matemática: El caso de la segregación espacial de inmigrantes en la ciudad de Cali.

3.1 Introducción

En esta última parte se pretende estudiar la interacción de los conceptos del análisis de datos espaciales y el marco teórico provisto por la matemática, mediante un ejercicio de aplicación que se enfoca en la inmigración y la segregación, un tema subyacente en el área de las ciencias sociales y que impacta en la economía regional y urbana. Este fenómeno social permitió analizar bajo tratamientos matemáticos, algunos patrones y comportamientos que hacen más fácil su entendimiento. Teniendo en cuenta que existen paquetes informáticos que procesan información geográfica con algoritmos matemáticos, se utilizó el sistema de información geográfico (SIG) llamado GeoDa y así lograr una mejor visualización y resultados más robustos.

3.2. Metodología para realizar la aplicación del análisis de datos espaciales en la segregación de inmigrantes.

Teniendo presente que parte de la literatura revisada involucra un fuerte componente matemático para explicar de manera más estricta lo correspondiente a ciertos eventos sociales, demográficos y económicos, esta sección se enfocará en el estudio de la inmigración y la segregación espacial desde una perspectiva matemática, y de alguna forma se mostrará como dos conceptos de origen social, se apoyan de las estructuras matemáticas para generar análisis con un soporte sólido y preciso. Para esto es necesario iniciar con una aproximación a la definición inmigración y segregación, y como se ha intentado medir con trabajos descriptivos y cualitativo hasta formas complejas como modelos de regresión.

Para involucrar de manera armoniosa ambos marcos teóricos, debemos contar con una matriz de datos los cuales nos dé una antesala de cómo se comporta la migración y la segregación, y eso se logró gracias a la información pública de las bases de datos que provee el Departamento administrativo nacional de estadísticas (DANE), permitiendo que esta caracterización cuantitativa sirviera para ser tratada matemáticamente.

Ya teniendo claro lo ejes que ayudarán a este proceso, se remite a los visto en la sección § 2.4 correspondiente al análisis exploratorio de datos espaciales, donde utilizaremos las herramientas matemáticas y propias del análisis de datos espaciales para realizar un acercamiento a la relación entre inmigración y segregación de forma matemática, pues estableceremos unas primeras conclusiones mediante la matriz de contigüidad, el test I de Morán y diagramas de correlación y dependencia espacial. Por último, aplicaremos el análisis confirmatorio visto en la sección § 2.5 para complementar las discusiones de nuestro caso aplicado con un modelo de regresión lineal , que aterrizado en un contexto geográfico evoluciona a un modelo de regresión con rezagos espaciales, instrumentos estadísticos que ayuda a explicar la inherencia del espacio en las explicaciones de este fenómeno. Cabe mencionar, que todos los modelos se realizaron mediante un software reconocido en este campo, llamado Geo Da, el cual implementa herramientas geografías y matemáticas para realizar los respectivos cálculos.

3.3. Inmigración y segregación.

Debido a que la inmigración es un fenómeno muy común en nuestro contexto y es referenciado en varios estudios de orden social como en Rodríguez (2011) y Dane (2003), se logra evidenciar algunos atisbos de ciertos patrones que prefiguran comportamientos estándares y posiblemente repetitivos que con algunos conceptos matemáticos se podrían modelar conveniente. Es claro, que la inmigración está estrechamente relacionada con el espacio geográfico en virtud de donde inicia y donde termina este proceso, también algunos estudios como Romero & Vargas (2016) y Vivas (2013), nos hablan de que los inmigrantes no llegan de manera arbitraria a los lugares donde se asientan, sino que existe un conjunto de elementos que lo causan, lo cual nos lleva a la investigación de la *segregación espacial*. Para esclarecer mejor este concepto, que es de un carácter muy sociológico, una definición apropiada sería entenderla como "la aglomeración geográfica que tiene como factor principal las familias que comparten una misma condición social o clase social que puede estar distribuida y concentrada en algunas áreas socialmente homogéneas" (Sabatini, 1999), aunque de manera más aterrizada al objetivo de este trabajo, la segregación espacial se utilizaría para explicar y entender como grupos o individuos de características muy parecidas tienden a agruparse en el espacio bajo una serie de condiciones, es decir, a localizarse más próximas entre sí, que respecto a otros grupos, permeando un comportamiento con una posible relación espacial y por consiguiente involucrando la teoría de análisis de datos espaciales. Resumiendo, se buscará matemáticamente la forma de medir y de una manera inicial, asentar un pronóstico que explique cómo los migrantes llegan a un lugar determinado.

Para robustecer más la aplicación que se va a mostrar en este capítulo y puntualizar la intervención de los conceptos matemáticos en este tema, se revisaron los artículos de Massey & Denton (1988) y Duncan (1995), donde el común denominador es tratar de medir mediante formas matemáticas la segregación socio espacial, obteniendo que cualquier análisis tenga el soporte riguroso mediante indicadores cuantitativos, resultados ofrecidos por la matemática y estadística, al servicio de varias áreas de conocimiento. Los indicadores que generalmente se usan para analizar la segregación espacial y social son el índice de Disimilitud (D) y el índice de segregación (SI), representados matemáticamente como:

$$D = \frac{1}{2}\sum_{i}^{n}\left|\frac{N_{1i}}{N_1} - \frac{N_{2i}}{N_2}\right| \qquad 0 \leq D \leq 1 \qquad \textbf{y} \qquad IS = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{t_i - x_i}{T - X}\right| \qquad 0 \leq SI \leq 1$$

Estos parten de la existencia de dos grupos poblacionales, uno minoritario y otro mayoritario y son indicadores sintéticos o globales en la medida que aporta índices agregados, además incluyen aspectos que subyacen a la segregación espacial como provocados por la uniformidad o igualdad, la exposición, la concentración, la centralización y la aglomeración, los cuales pueden inducir efectos en las dimensiones de mediciones. Estos indicadores son fundamentales para realizar la regresión, pues prefiguran la estimación de la variable independiente de la modelación que se revisará más adelante.

3.4. Naturaleza de los datos.

Los datos con los cuales se realizará la situación de aplicación son obtenidos de los archivos originados del procesamiento del censo de la población realizado en Colombia en el año 2005 por el DANE las cuales son de acceso público, enfocándose principalmente en las dinámicas del municipio de Cali en el Valle del Cauca, lo anterior teniendo como referencia el trabajo realizado por Romero & Vargas (2016).

La base de datos está configurada de tal forma que se distinguieran dos grupos, el de interés que son *los inmigrantes recientes*, que corresponde a la población que en el año 2000 vivía en una ciudad diferente a Cali y el grupo de población que para el año 2000 residía en la ciudad de Cali, considerando principalmente los nacidos en el municipio y los migrantes que llevaban un tiempo considerable y los cuales se notarán como *no inmigrantes recientes*. Esta información fue obtenida del censo mediante de la pregunta *¿en dónde vivía…hace cinco años?* del módulo de personas del formulario censal. Por lo anterior, el análisis esta alrededor de las personas que han llegado a la ciudad a partir de ese año y como se distribuyen en el espacio geográfico.

Teniendo en cuenta que los datos generaron una matriz bastante grande, se optó por seguir las recomendaciones para configurar el territorio de una forma fácil de trabajar.

> *Dado que el país cuenta con más de 1.100 municipios, se construyó una regionalización a partir de la agrupación de municipios, que dio un total de 11 regiones. Los criterios considerados para esta delimitación implicaron el cruce de diferentes aspectos, que consideraron la identificación de la cuenca migratoria tradicional de la ciudad, dando cuenta de los municipios que han estado históricamente vinculados a ella y su área de influencia; el análisis de la cantidad de inmigrantes recientes provenientes de cada uno de los municipios del país, lo que llevó por ejemplo a considerar a Bogotá como una sola región dado el peso de los inmigrantes recientes aportados, igualmente se hizo con Antioquia. Y, por último, la consideración de unas regiones geohistóricas (Pacífica, Atlántica, etc.). (Romero, Vargas, 2016, p.10)*

El área que servirá para la aplicación esta de limitada a la zona urbano de la ciudad Cali y toda la información fue procesada a nivel de sector censal de barrio, lo cual arrojo 344 polígonos (datos), siendo homologado a través del uso de herramientas propias de los sistemas de información geográfica (SIG).

3.4.1. Algunas variables de interés para el estudio de la inmigración y segregación.

Teniendo en cuanta que el análisis de datos espaciales en este trabajo, tiene como gran finalidad medir la interacción de la segregación en los inmigrantes en la ciudad de Cali, es imperante valerse de ciertos atributos sociales, que al transponerse en un lenguaje matemático son de gran ayuda para obtener resultados más precisos. Es el caso de la variable endógena o dependiente que se supone explica la concentración de los inmigrantes, es decir la segregación espacial, pues no está explicita en los matrices de datos, pero se logró medir utilizando un índice de concentración desarrollado en Romero & Vargas (2016), denominado $Q_{M_{ij}}$ donde ayuda a representar la concentración de algunos grupos poblacionales dependiendo de su origen migratorio y así poder observar como estos logran ubicarse en algunas zonas de Cali. El índice se estructura así:

$$Q_{M_{ij}} = \frac{\left[M_{ij} / \sum_{i=1}^{n} M_{ij}\right]}{\left[\sum_{j=1}^{m} M_{ij} \,/\, \sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}\right]}$$

Donde

- $Q_{M_{ij}}$ índice de concentración regional.
- M_{ij} inmigrantes recientes de la región i en el barrio j.
- $\sum_{i=1}^{n} M_{ij}$ inmigrantes de todas las regiones en el barrio j.
- $\sum_{j=1}^{m} M_{ij}$ inmigrantes recientes de la región i en los demás barrios.
- $\sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}$ inmigrantes recientes de la ciudad.

Por lo visto anteriormente, se logra establecer una línea estrecha entre un comportamiento social y un hecho estilizado de la matemática, pues se logra entender cómo se ubican o distribuyen los inmigrantes, según su origen en cada barrio de la ciudad utilizando una métrica rigurosa, algo que se observó someramente en los índices de tradicionales de segregación como el índice de disimilitud (D) o el índice de segregación (IS), pero aquí más precisos al tener una unidad de análisis como son los barrios.

Para fundamentar el carácter social de la segregación, sin olvidar la idea de matematizar este argumento, se incluyó el impacto que provoca de manera indirecta la calidad de las interacciones posibles con los vecinos más próximos y se logra por la métrica que propone Vivas (2007), donde aplicando el concepto de índice nuevamente logra obtener una expresión matemática que se denotará como $\frac{K_h}{K_l} i$ que recoge lo anterior mencionado y se estructura como sigue:

$$\frac{K_h}{K_l} i = \left[\sum_{i=1}^{n} \frac{\left(\frac{H_{edu}}{L_{edu}}\right)_i - \left(\frac{H_{edu}}{L_{edu}}\right)_{min}}{\left(\frac{H_{edu}}{L_{edu}}\right)_{max} - \left(\frac{H_{edu}}{L_{edu}}\right)_{min}} \right] * 100$$

Donde *H_edu/L_edu* es construido a partir del nivel educativo y básicamente obedece a un cociente, entre educación alta y educación baja obtenida desde la base de datos por el nivel estudio de cada hogar, en cada barrio *i*.

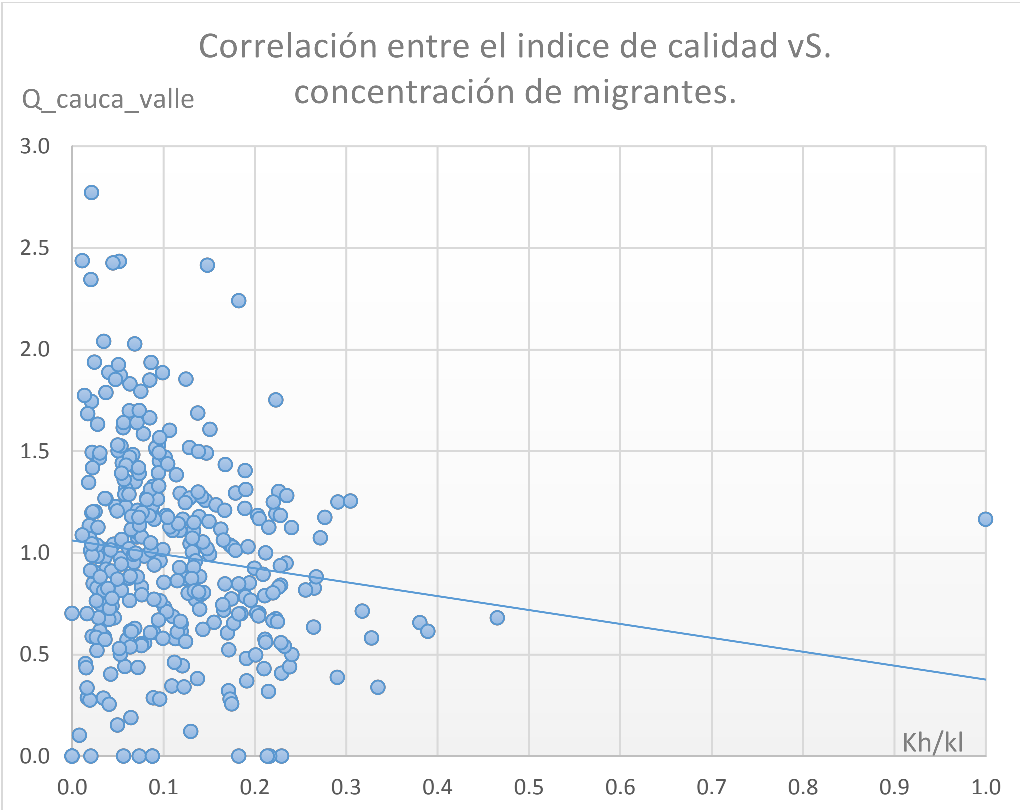


Figura 4. Correlación entre índice de calidad kh/kl y concentración de los inmigrantes región norte Cauca y centro del Valle. Fuente: Elaboración propia.

Cabe anotar que las demás variables de la matriz de dato son obtenidas de manera cuantitativa y son la denominas demográficas y socio económicas que suelen estar presente en base de datos de información.

3.5. Análisis exploratorio de datos espaciales.

Haciendo análisis previos sobre la base de datos, se logra establecer que existe cierta dependencia entre zonas de concentración (segregación) y lo inmigrantes de las regiones Pacifico, Bogotá, Antioquia y norte del Cauca centro del Valle, esto utilizando principalmente índices de correlación y el índice de Morán. Por efectos de interés de este trabajo se toma para analizar la región Norte del Cauca y centro Valle.

Teniendo la región de análisis seleccionada, ahora es una tarea necesaria contrastar el marco de la dependencia espacial visto en la sección §2.3, y así establecer si existen efectos espaciales que puedan argumentarnos que la distribución de los inmigrantes de esta región no es aleatoria y por el contrario perite un patrón de distribución y segregación que valida la aplicación de los análisis globales y locales propuestos. A partir de estos, se pudo asegurar si la segregación espacial descrita por la variable concentración (Q_M), de manera sistemática determinado por la auto correlación espacial.

La herramienta seleccionada por su facilidad de interpretación y facilidad al calcular para el análisis global es el índice de Moran, ya que al estar relacionado con la matriz de contigüidad puede suponer una dependencia espacial que intentamos verificar. Se debe recordar que matemáticamente este índice mide a través de matrices la relación fuerte o débil de concentración en el espacio de las variables, matemáticamente se fundamenta en una correlación sencilla, pero aquí introduce una matriz para determinar el impacto de las cercanías de los datos.
El índice de auto correlación espacial global I de Moran estimado para la región Norte del Cauca y centro Valle es de 0,16 lo cual sugiere un comportamiento sistemático, aunque no tan marcado puesto que es menor a 0.3, pero ayuda a considerar esta hipótesis como posible. (Ver gráfico).

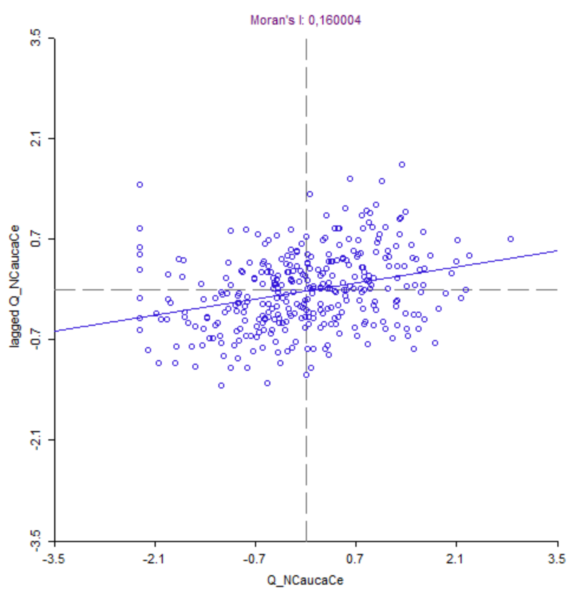


Figura 5.Scatter Plots de Morán univariante Región Norte del Cauca –centro del Valle. Fuente: Estimaciones propias en GeoDa.

El eje x del gráfico representa el índice de concentración de la región analizada y el eje y el rezago espacial de primer orden de índice. Cada punto representa un barrio.

Ya ejecutado el primer paso, el que nos permite concebir una pista de que la segregación o concentración no se da de manera arbitraria a nivel general, entramos en el área del análisis local, pues como se dijo anteriormente se tiene 344 polígonos referentes a los barrios y el LISA que es "un estadístico (resultado de una distribución) que cumple con dos requerimientos: a) El Lisa para cada observación da una indicación de la medida de la significancia de la agrupación espacial de valores similares alrededor de una observación y b) La suma de los Lisa para todas las observaciones es proporcional al índice global de asociación espacial." (Anselin, Local indicators of spatial association, 1995), nos permite conocer cómo y dónde se agrupan. Esto se logra de una manera explícita y exitosa por la matriz de pesos espaciales consecuencia de la matriz de contigüidad.

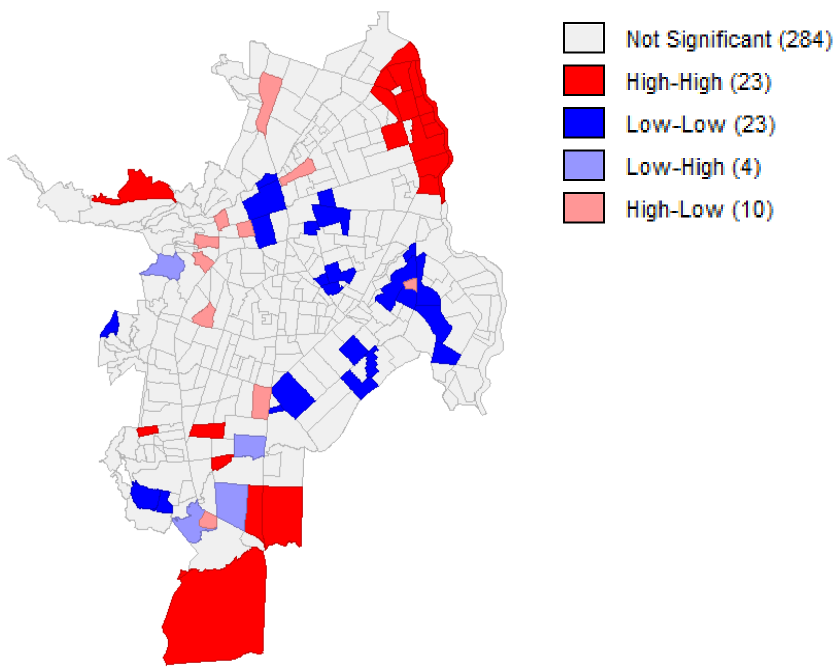


Figura 6. LISA para el índice de concentración región norte del Cauca y centro del Valle. Fuente: Estimaciones propias en GeoDa.

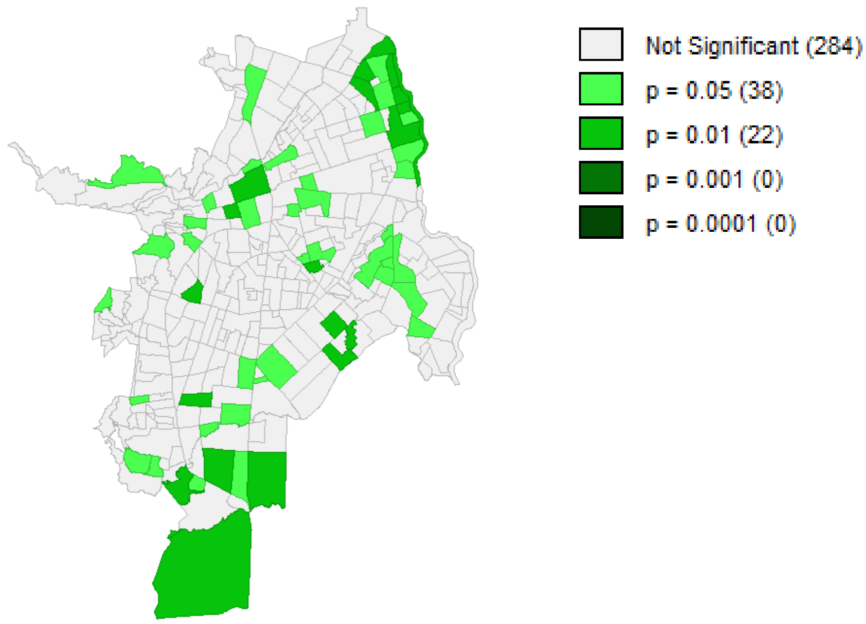


Figura 7. . Significancia del LISA para el índice de concentración región Norte del cauca y centro del Valle Fuente: Estimaciones propias en GeoDa.

Los resultados obtenidos por los análisis globales y locales muestran claramente la existencia de barrios con concentración representativa de los inmigrantes recientes del norte del Cauca y centro del Valle, mostrándonos que hay una convergencia de este grupo están rodeados por barrios con alta concentración de esta misma población, identificándose así dos clústeres. Uno en la zona oriental y nororiental. Contrastando los resultados del LISA con el índice Kh/Kl, queda en evidencia que los inmigrantes recientes de la región norte del Cauca y centro del Valle se localizan en barrios de la ciudad en donde el nivel de las interacciones posibles se da entre personas con bajo capital humano, un patrón antes visto en la literatura pues se dice que en esa zona se asientan.

> *"poblaciones que comparten rasgos comunes, entre los que se destacan la presencia de asentamientos informales, dificultades de acceso a los mercados formales de trabajo, bajos niveles de calidad de vida, altas tasas de subutilización de la mano de obra y restricciones en la provisión de bienes públicos locales de buena calidad, como es el caso de la educación y la salud". (Vivas, P, 2013, P.146)*

Este apartado mostró como teniendo establecidos niveles de proximidad mediante matrices, índices de correlación entre variables y niveles de significancia se puede concluir de manera sólida sobre patrones de concentración.

3.6. Análisis confirmatorio

En esta sección se acude a la estimación de modelos de regresión que permiten estudiar las variables que contribuyen, explican y cuantifican la segregación espacial en los inmigrantes de la zona norte del Cauca – centro del Valle. El ejercicio anterior nos bosqueja un panorama que nos permite decir que matemáticamente los asentamientos obedecen a un patrón de distribución, y las ciencias sociales deberán explicar en su contexto el porqué. Ahora, lo que las regresiones permiten es que, teniendo los datos, podamos decir cuánto es la influencia de cada variable en este comportamiento, se podrá atribuir cuantitativamente que impacto tiene el nivel educativo, o el estrato, o el estado civil en establecerse o no en determinad zona. Teniendo en cuenta algunas metodologías de modelación como las de David F. Hendry & Cristopher A. Sims, comentadas en Pagan (1987), se obtienen aproximaciones que permiten esclarecer el comportamiento de la segregación por parte de los inmigrantes recientes.

Se sigue una estrategia convencional de acuerdo a lo sugerido por Anselin (1988) consistente en estimar un modelo de mínimos cuadrados ordinarios (MCO), donde se supone que los datos se comportan bajo las características de una distribución normal.

Adicionalmente se realizan pruebas para verificar si efectivamente existe dependencia espacial sustantiva. Los test aplicados son I-Moran (vistos anteriormente) y multiplicadores de LaGrange LM_ρ y LM_λ, ya sea por un rezago espacial (lag spatial) o una estructura autoregresiva en los errores. Para seleccionar el modelo más adecuado, entre un modelo de rezagos espaciales (LM_{lag}) o un modelo de dependencia espacial residual (LM_{error}), se establece el más significativo comparando los valores que arrojan los test. Por lo que si la significancia de $LM_\rho > LM_\lambda$ se estima un modelo LM_{lag}, de lo contrario, si $LM_\lambda > LM_\rho$ se estima un modelo LM_{error}. Dado el caso de que ambos sean significativos, se debe analizar los errores robustos siguiendo el mismo criterio de comparación.

De la matriz de datos compuesta por variables individuales, de hogares, sociales y del barrio, resultaron ser de gran interés el índice de capital humano (Kh/Kl), la población total del barrio, el nivel de educación secundaria y que no se declaren de ninguna etnia. Es evidente que Kh/Kl mostró una correlación con el índice de concentración, las demás variables se concibieron de un abanico amplio de posibles predictores, que terminaron por una parametrización del modelo, estructurándolo de la siguiente forma, tanto para la región:

$$Q_{cauca-valle_i} = \beta_o + \beta_1 * Kh/Kl_i + \beta_2 * PobBarrio_i + \beta_3 * NoEtnia_i + \beta_4 * Secun_i + U_i$$

Los resultados parciales obtenidos en el software muestran que las regresiones de MCO no cumplen con la condición de distribución idéntica e independiente en los errores, condicionando la necesidad de un modelo diferente, en este caso de rezago espacial, dado que los valores del I-Moran y de los multiplicadores LM_λ, LM_ρ son significativos. (Ver tabla 4). Así, por la sección §2.5.1 se puede utilizar el modelo dado a continuación:

$$Q_{cauca-valle_i} = \rho * W * Q_{cauca-valle_i} + \beta_o + \beta_1 * Kh/Kl_i + \beta_2 * PobBarrio_i + \beta_3 * NoEtnia_i + \beta_4 * Secun_i + U_i$$

La variación notable está en el término $\rho * W * Q_{región_i}$, donde W es la matriz cuadrada de contactos espaciales de contigüidad, que captura la dependencia espacial de un barrio con sus vecinos. $W * Q_{región_i}$ es el rezago espacial del índice de concentración de los inmigrantes recientes, que facilita la estimación e identificación de las interrelaciones existentes en cada barrio; y ρ es un parámetro autorregresivo que modela el nivel de intensidad de las interdependencias, interacciones y retroalimentaciones a través de las observaciones muéstrales. Realizando el tratamiento adecuado de matrices y para una mejor visualización de los efectos de propagación, el modelo se puede reescribir

$$Q_{cauca-valle_i} = (I_n - \rho W)^{-1} * + \beta_o + \beta_1 * Kh/Kl_i + \beta_2 * PobBarrio_i + \beta_3 * NoEtnia_i + \beta_5 * Secun_i + (I_n - \rho W)^{-1} * U_i$$

Q_Cauca-Valle	***MCO***	***LAG***
Constante	-32,401***	-43,47***
kh/kl	-0,592**	-0,4386*
Pob_Barrio	0,008***	0,008***
No_etnia	0,451***	0,4696**
Edu_secundaria	-0,36*	-0,364*
$\boldsymbol{\rho}$	-	0,2049
R2 ajustado	0,6296	-
L-L	1732,23	-1724,1
N	344	344
Moran´s I	0,2558***	-
$\boldsymbol{LM_\rho}$ $\boldsymbol{LM_\lambda}$	*** ***	-
Test robustos	$prob(LM_\rho) < prob(LM_\lambda)$	-
Akaike	3476,47	3462,2

Tabla 4.Modelos de regresión Índice de concentración región norte Cauca y centro del Valle. Variable endógena: Índice local de concentración de inmigrantes recientes según región de procedencia. L-L: Máxima Verosimilitud, N: Cantidad de unidades censales, LM_ρ y LM_λ: Multiplicadores de LaGrange para modelos de rezago y error. Significancia $(*)$ para $p < 0,05$; $(**)$ para $p < 0,1$ y $(***)$ para $p < 0,001$. Fuente: Estimaciones propias en GeoDa a partir de la base construida

Los resultados de la regresión para la región confirman lo visto en el análisis exploratorio para el índice de capital humano. Tiene el signo negativo, confirmando la relación directa evidenciada previamente. Si se da el aumento de una unidad en el índice de capital humano entonces se disminuye la concentración de inmigrantes recientes de la región norte Cauca centro Valle en 0,4386 puntos.

Con referencia la población de cada barrio, se encuentra que, aunque es significativo solo logra impactar en 0,008 puntos el incremento de la concentración de inmigrantes de la región.

El resultado para el predictor no se reconoce en un grupo étnico incide de manera positiva, es decir que se aumenta la concentración de inmigrantes de la región en 0,451 puntos.

Ahora un hecho que puede generar un análisis bastante oportuno es que, al tener una educación secundaria, la persona que desea llegar a la ciudad tiende a disminuir la segregación espacial en 0,36 puntos.

Es importante señalar que estos valores varían en el espacio por la incidencia de la matriz de multiplicadores espaciales $(I_n - \rho W)^{-1}$, en la proporción del coeficiente de auto correlación espacial **ρ,** que para el modelo matemático arrojado aumenta el nivel de concentración de inmigrantes recientes en un valor de 0,2049. Estos resultados nos permiten rechazar la hipótesis nula, verificando que la configuración espacial de la concentración de los inmigrantes recientes de la región Norte del cauca centro del Valle **no se produce de manera trivial** en la ciudad de Cali, quedando en evidencia la interdependencia y retroalimentación entre los barrios, ratificando unos efectos de vecindario.

Conclusiones

Uno de los principales resultados de este estudio es que a partir de las diversas herramientas que nos proporciona el marco teórico matemático implementado en la economía regional y urbana, especialmente en el análisis de datos espaciales, fue posible afirmar que los inmigrantes que llegan a la ciudad de Cali de diversas regiones, siguen tendencias de concentración en zonas específicas.

El hecho de tener todo el tratamiento y la estructura lógica de ciertos procesos matemáticos, es posible matematizar diferentes situaciones que generan resultados utilizados en otras ciencias. Por ejemplo, los resultados de los índices locales mostraron ciertos *clústeres* en detenidas zonas de la ciudad, que si se analizan desde una perspectiva socioeconómica, puede determinarse que coinciden con zonas de estratos medios tendientes a bajos y pueden ser destinos de fácil acceso para las personas que migran con recursos limitados.

Todo el tratamiento matemático y geográfico que se le pudo dar a este ejercicio de aplicación, nos lleva a pensar que no solo existen variables comunes que pueden explicar cierto evento, sino que el espacio, partiendo de la definición básica de espacio euclídeo visto en topología, configura un rol importante en la explicación y comprensión de los fenómenos sociales, económicos, y en este caso, lo que corresponde a las migraciones y la segregación. Queda comprobado, pues la variable que más aporta en la explicación del análisis realizado es **ρ,** el coeficiente de auto correlación espacial calculado a través de una regresión teniendo en cuenta la matriz de contigüidad.
También se logra mostrar que la interdisciplinariedad del marco teórico de la matemática logra grandes aportes en otras áreas del conocimiento y que es un grave error considerar que la matemática no tiene funcionalidad alguna fuera de las áreas afines a ella como la aplicación en ciencias exactas, naturales y las licenciaturas. Este trabajo ha sido bastante enfático en concebir por separado conceptos matemáticos y de otras ciencias, pero vinculados en un objetivo común, logrando resultados objetivos y bien fundamentados. Se pudo ver claramente la contribución de la matemática en virtud de buscar la solución a una problemática social.

Por último, se quiere establecer la premisa sobre generar espacios en el aula que permitan conocer este tipo de ejercicios de aplicación de la matemática en las diversas áreas del conocimiento, especialmente en las licenciaturas que forman a los futuros docentes y tendrán que responder preguntas como ¿Para qué sirve eso? ¿Eso donde se aplica? o peor aún, en una labor que no sea enseñar matemática, ¿de qué me servirá en mi trabajo? Es importante que se manifieste que las matemáticas no son un “estandarte” de cálculos sin relación alguna con la cotidianidad, y esa misión recaería plenamente en los nuevos docentes, que tienen el gran reto de fundamentar al estudiante en la parte teórica pero también en la parte aplicativa. Ejercicios como el anterior, permitirán concebir que la matemática es una ciencia que pretende modelar de la manera más sencilla cualquier evento de la existencia misma.

Bibliografía

Anselin, L. (1980). Estimation methods for spatial autoregressive structures. Ithaca NY: Comell University.

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Netherlands: Kluwer Academic Publishers.

Anselin, L. (1998). Exploratory Spatial data analysis in a geocomputational environment. West Virginia University: Research Paper.

Anselin, L. (1996). The Moran Scatterplot as an ESDA tool to assess local instability in spatial association. Spatial analytical perspectives on GIS. London, Taylor and Francis.

Anselin, L. (1995). Local Indicators of spatial association- LISA. Geographical Analysis, 27, 93-115.

Arce, C, Castillo W & Gonzales J. (2003). Algebra lineal, Costa rica: Carlos Arce.

Aranda, E. (2013). Algebra lineal con aplicaciones y phyton, España: Ernesto Aranda.

Altamirano, P & Espinosa, E. (2009). *Guía para la presentación de gráficos estadísticos*. Perú: INEI.

Asuad, S & Quintana, R. (2010). Crecimiento económico, convergencia y concentración económica especial en las entidades federativas. Investigaciones regionales, 18, 83-106. Recuperado de: http://www.redalyc.org/pdf/289/28916988004.pdf.

Acevedo, I & Velásquez, E. (2008). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. *Ecos de economía*, (27), 9-34. Recuperado de: http://publicaciones.eafit.edu.co/index.php/ecos-economia/article/view/705/627.

Bagur, A. (2009). El príncipe de las matemáticas. *Matemáticas para todos* 10(88), 1-4. Recuperado de http://matematicas.reduaz.mx/home/publicaciones/Boletin_Mate_88_Mzo_09.pdf.

Badii, Guillen, Araiza, Cerna, Valenzuela & Landeros (2012). Métodos no paramétricos de uso común. Universidad Spenta. Recuperado de: http://www.spentamexico.org/v7-n1/7(1)132-155.pdf.

Cliff, A. & Ord, J. (1981).Spatial Process. Models and Applications. London: Pion.

Cabrera, E. (2009). El coeficiente de correlación de los rangos de Spearman caracterización. *Habanera de ciencias médicas,* 7(2), 1-19. Recuperado de http://scielo.sld.cu/pdf/rhcm/v8n2/rhcm17209.pdf.

Chasco Yrigoyen, Coro. (2003). Métodos Gráficos del Análisis Exploratorio de Datos Espaciales. Madrid. Recuperado de: https://www.researchgate.net/publication/256174755_Metodos_graficos_del_analisis_exploratorio_de_datos_espaciales.

Cleveland, W. (1993). Visualizing data, Summit, NJ, Hobart Press.

Dane, 2003. Evidencia reciente del comportamiento de la migración interna en Colombia a partir de la Encuesta Continua de Hogares. Bogotá. Recuperado de: https://www.dane.gov.co/files/banco_datos/Migracion/migracion_interna_Clbia.pdf.

Doodchild, M (1987). Spatial analytical perspective on geographical Information systems, *international journal of geographical information systems.* (1), 327-334.

Duncan, O (1995). Methodological Analysis of Segregation Indexes. *American Sociological Review*, (41), 210- 217.

Estuardo, A. (2012). Estadística y probabilidades, Chile: Morales

Grossman, S & Flores, J. (2012). Algebra lineal, México: Mc Graw Hill.

Gonzales & Meneses (2008). Algebra Lineal. *Universidad de Sevilla, 1-149.* Recuperado de: http://matematicas.unex.es/~navarro/algebralineal/meneses.pdf.

Getis, A & Ord, k. (1992). The analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis.* (24), 1-18. Recuperado de: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1538-4632.1992.tb00261.x.

Good, I. (1983). The philosophy of exploratory data analysis. *Philosophy of Science,* (50), 283-295.

Hamilton (1844). On quaternions, or on a new System of imaginaries in algebra. *Philosophical Magazine.* 1-92. Recuperado de: https://www.emis.de/classics/Hamilton/OnQuat.pdf.

Hendry, D & Sims, C. (1987). Tres metodologías econométricas: una evaluación crítica. (1), 1-2.

Levine. M. Krehbiel. C & Berenson. (2006). Estadística para administración. México: Pearson.

Laguna, C. (2014). Correlación y regresión lineal, PP. 1-18. Recuperado de: http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T04.pdf.

Martínez, H & Sanabria, A. (2008). Algebra lineal, 1-214. Recuperado de https://viterick.files.wordpress.com/2010/10/algebra-lineal_martinez.pdf

Moral, I. (2013). Medidas de asociación, 1-10. Recuperado de http://www.revistaseden.org/files/13-cap%2013.pdf.

Moreno, S & Vaya, E*, (2002). Econometría espacial: nuevas técnicas para el análisis regional. Una aplicación a las regiones europeas,* Investigaciones Regionales, *1-24.*

Moore (2002). Análisis de relación. Estadística Aplicada Básica. Recuperado de: http://www.econ.upf.edu/~satorra/dades/moore_02.pdf.

Moran, P. (1948). The interpretation of statistical maps. Journal of the Royal Statistical Society B, 10, 243-251.

Massey, D & Denton, N. (1998). The Dimensions of Residential Segregation. Oxford University Pres. (67), 281-315.

Olivares, P. (2002). Análisis Exploratorio y Análisis Confirmatorio de Datos. *Universidad del Zulia*. 11, 1-11. Recuperado de: http://www.redalyc.org/articulo.oa?id=12211106.

Pineda, J. (2006). Econometría espacial y ciencia regional. *Investigación Económica.* (258), 129-160. Recuperado de: http://www.redalyc.org/pdf/601/60125804.pdf.

Paelinck, J. & Klaassen, L. (1979). *Spatial Econometrics*, Saxon House, Farnborough, 1-211.

Romero, S. & Vargas, L. (2016). *Segregación socioespacial de los inmigrantes recientes en la ciudad de Cali. Un abordaje desde la econometría espacial evaluando dos regiones de origen: Pacífico y Bogotá.* Pontificia Universidad Javeriana. Colombia.

Rodríguez, M. (2011). *La matemática y su relación con las ciencias como recurso pedagógico*, Didáctica de las matemáticas, 1-15.

Rojas (2003). Técnicas estadísticas paramétricas y no paramétricas equivalentes: Resultados comparativos por simulación (Tesis de pregrado). Recuperado de:

http://www.iuma.ulpgc.es/~nunez/mastertecnologiastelecomunicacion/RecursosGenerales/TesisEstadisticaParametricayNoParametrica.pdf.

Rosales, A. (2009). Evolución histórica del concepto de matriz. *Educación e internet* 9(2), 1-20. Recuperado de https://tecdigital.tec.ac.cr/revistamatematica/ContribucionesV9_n1_2008/Evolucion_Historica_del_concepto_de_matriz.pdf.

Sandoval, I. (2011). Cuadernillo de apuntes calculo integral. *Gobierno del estado de México, 1-64.* Recuperado de http://www.tesoem.edu.mx/alumnos/cuadernillos/2011.006.pdf.

Steel, R & Torrier, J. (1960). *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*, pp. 187 – 287.

Sabatini, f. (1999). La segregación social del espacio en las ciudades de américa latina. *Banco interamericano de Desarrollo.* 1-46. Recuperado de: https://www.researchgate.net/publication/254421887_La_segregacion_social_del_espacio_en_las_ciudades_de_America_Latina.

Tukey, J. (1977). Exploratory data analysis. Reading, Addison – Wesley.

Uriel, E. (2013). Regresión lineal simple. *Universidad de valencia*, 1- 38. Recuperado de https://www.uv.es/=uriel/2%20El%20modelo%20de%20regresion%20lineal%20simple%20estimacion%20y%20propiedades.pdf

Vinuesa, P. (2016). Correlación: Teoría y práctica, 1-26. Recuperado de http://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.pdf.

Churata, N. (2016). *Notación de sumatoria*. Perú: Senati.

Vivas, H. (2007). Educación, background familiar y calidad de vida de los entornos locales en Colombia. Barcelona: Tesis Doctoral, Departamento de economía aplicada, universidad autónoma de Barcelona.

Vivas, P.H. (2013) persistencia de la segregación residencial y composición del capital humano por barrios en la ciudad de Cali. Ensayos sobre POLITICA ECONOMICA, (31), 122-155.

Walpole, R. Myers, R. Myers, S & Ye. (2007). Probabilidad y Estadística para Ingeniería y ciencias. Pearson.