UNIVERSITAT DE BARCELONA

Tres ensayos sobre la movilidad laboral. Aspectos metodológicos y evidencia empírica

Omar García León

Doctorat en Empresa
Facultat d'Economia i Empresa

Tres ensayos sobre la movilidad laboral. Aspectos metodológicos y evidencia empírica.

Omar García León

Tesis Doctoral dirigida por
Dr. Raúl Ramos Lobo

Doctorat en Empresa

Título de la Tesis:

Tres ensayos sobre la movilidad laboral. Aspectos metodológicos y evidencia empírica

Doctorando:

Omar García León

Director de Tesis:

Dr. Raúl Ramos Lobo

Septiembre 2015

ÍNDICE

TABLAS

FIGURAS

1. Introducción.

En el año de 2014, la densidad de población de España estimada por la Organización de las Naciones Unidas para la Agricultura y la Alimentación y el Banco Mundial fue de 93 habitantes/Km^2. Para este mismo año se estima la densidad de la población en la provincia de Barcelona en 714.3 habitantes/Km^2, mientras que para la comunidad de Madrid aumenta a 809 habitantes/Km^2.

En el caso de México, la densidad de población en 2014, según estimación de los mismos organismos, fue de 64 habitantes /Km^2, con una aumento drástico para el Distrito Federal cuya densidad estimada es de 5862 habitantes/Km^2, según el Instituto Nacional de Geografía y Estadística.

La razón del por qué la densidad de población en la provincia de Barcelona y en la comunidad de Madrid es mayor que la media nacional y el por qué la densidad de población es tan grande en la ciudad de México en comparación con la densidad del país, es simple y se debe a que la población se distribuye de manera desigual.

Las personas se agrupan por diversas causas o razones, sociológicas, culturales, geográficas, etc., pero primordialmente se aglomeran por causas económicas. La distribución de la población y la actividad económica a través del espacio es muy desigual, con una aglomeración de la actividad económica en puntos importantes.

La gente se agrupa y se mueve en función primordialmente de las fuentes de trabajo existentes y es así como van creciendo las ciudades y los núcleos urbanos. Diariamente las personas se mueven del hogar hacia el trabajo y viceversa, van de compras, acuden a la escuela, etc.; además hay flujos de dinero, bienes, productos básicos, materias primas y es en estos datos de flujos donde esencialmente se encuentra las fuentes de datos secundarios relativas a las personas, sus organizaciones, comportamientos y actividades en el espacio.

El concepto de regiones funcionales proporciona un marco para la investigación de la estructura del espacio geográfico. Una región funcional se caracteriza por la aglomeración de las actividades económicas en un centro regional que concentra los movimientos de las personas del interior o de los alrededores.

Una región funcional se describe generalmente en términos de la estructura de las interacciones espaciales, lo que ayuda a delinear regiones, naturalmente, mediante la

maximización de la diferencia entre dentro-región y las interacciones entre-región (Noronha y Goodchild 1992). Por lo tanto, el patrón de los flujos geográficos es un indicador eficaz para la comprensión de la estructura de las actividades económicas de las personas en el espacio (Taaffe, Gauthier y O'Kelly 1996).

Los flujos de desplazamiento diario al trabajo describen por lo general la estructura espacial de la actividad urbana: ya que los flujos de población representan las actividades económicas de las regiones, muchos sistemas de planificación urbana han utilizado datos de flujo para el diseño de las regiones funcionales (ver Masser y Brown 1975; Masser y Scheurwater 1980; Rosing y ReVelle 1986; Noronha y Goodchild 1992; Alvanides, Openshaw, y Duke-Williams 2000)

Varios términos se han utilizado para describir la regionalización de los flujos: regiones nodales, teoría de los lugares centrales, mercados laborales, cuencas, zonas comerciales, zonas de venta, viajes a las áreas de trabajo y regiones funcionales.

Una amplia gama de técnicas se han utilizado para regionalizar datos de flujo: La delimitación de las regiones funcionales por lo general implica la agregación de unidades territoriales pequeñas en menos grupos, que se basan generalmente en la contigüidad espacial y el mantenimiento de propiedades similares entre las unidades de área en cada grupo (Brown y Holmes 1971).

Más formalmente, el problema se puede definir como la identificación de un conjunto de p grupos, que se agrupan en n unidades de área a la vez que se optimiza una función objetivo predefinida con un determinado conjunto de criterios o restricciones. Las funciones objetivo pueden ser formuladas para minimizar la disimilitud de las unidades de área o para maximizar su similitud.

En esta Tesis el primer ensayo tiene como objetivo determinar mercados laborales por sector en España, a partir de la matriz de viajes origen-destino, que permitirá formar grupos o regiones homogéneas en base a la cantidad de personas que comparten destinos en su viaje al trabajo y lugares de residencia.

Estas regiones se integraran por grupos de municipios, considerando como criterio de similaridad entre municipios, los destinos (trabajo) y orígenes (residencia) que las personas tienen en su vida diaria, es decir dos municipios se considerarán similares si las personas que viven en esos dos municipios tienen destinos similares en su viaje al trabajo. Para la agrupación de los municipios se plantea utilizar un enfoque de estadística multivariable.

Ahora bien, los mercados laborales tienen heterogeneidad espacial, varían en su estructura, en su contexto social y en su historia, en formas que no se capturan fácilmente mediante las variables explicativas de una regresión global estándar.

Cuando en los modelos de regresión se asume que el poder explicativo es igual para todo el conjunto de observaciones, la idea de un comportamiento uniforme y constante del ajuste a través del espacio geográfico resulta por lo menos sospechosa.

En el modelo de mínimos cuadrados ordinarios, OLS, cada observación es independiente. Pero lo cierto es que los datos espaciales no cumplen la hipótesis de independencia, debido a que normalmente están autocorrelacionados, por lo que relación entre las variables del modelo no será la misma en toda el área de estudio (Clark, 2007). En este sentido, Lloyd and Shuttleworth (2005) destacan la necesidad de incluir en los modelos especificaciones más apropiadas, que consideren la naturaleza intrínseca de los datos espaciales, que normalmente están autocorrelacionados.

El enfoque de regresión global ignora uno de los principios fundamentales de la ciencia regional; lo relacionado con la localización espacial. Los científicos regionales esperan no sólo que las variables explicativas se diferencien a través del espacio, sino también que las respuestas marginales a los cambios en las variables explicativas pueden variar a través del espacio.

Una técnica de modelado relativamente reciente para el análisis de datos espaciales es la regresión geográficamente ponderada (GWR) que permite variaciones locales en los coeficientes a estimar.

Un factor importante en el crecimiento del empleo es el capital humano cuyas características no serán las mismas en todo el espacio de estudio.

El segundo ensayo en esta Tesis consistirá en determinar el crecimiento del empleo en función de diferentes variables relacionadas con el nivel de educación de la población con el enfoque de la regresión ponderada geográficamente lo que permitirá determinar la variación de la relación entre el crecimiento del empleo y las variables explicativas a través de todo el país.

En la calle de Donceles desde Isabel la Católica hasta República de Argentina en el centro de la ciudad de México, se localizan tiendas de cámaras fotográficas, video, grabadoras, luces y todo lo relacionado con la labor de fotógrafos, productores y periodistas. Es un lugar adecuado para este tipo de tienda, pero en las 668 manzanas, 9.7 km^2, que ocupa el centro seguramente hay

otros sitios que podrían ser convenientes. ¿Por qué entonces los dueños de las tiendas de artículos para la comunicación seleccionan Donceles? ¿No les importa lo perjudicial que podría ser tener tanta competencia cerca? Los clientes de esos artículos se dirigen a Donceles porque esperan encontrar un buen número de tiendas donde buscar lo que satisface sus necesidades y las tiendas se localizan ahí porque los dueños saben que tendrán acceso a un buen número de clientes potenciales.

El ejemplo que la calle de Donceles muestra a un nivel micro ilustra el concepto de aglomeración. La aglomeración ocurre a muchos niveles, desde distritos locales de venta que atienden a áreas residenciales dentro de una ciudad hasta regiones económicas especializadas como Silicon Valley que proveen al mercado mundial. La distribución de la población y la actividad a través del espacio es muy desigual.

Marshall (1890), sugirió un modelo de distrito industrial que se incluye en el concepto de aglomeración, donde la naturaleza y la calidad del mercado de trabajo local es interno al distrito y altamente flexible. Las personas se mueven de una empresa a otra, y los propietarios, así como los trabajadores viven en la misma comunidad, donde se benefician del hecho de que "el secreto de la industria está en el aire", es decir, hay una atmósfera industrial, como él lo define. Los trabajadores parecen estar comprometidos con el distrito en lugar que con la empresa, y, además, la emigración de la mano de obra se supone mínima. El distrito es visto como una comunidad relativamente estable que permite la evolución de la fuerte identidad cultural local y la experiencia industrial compartida.

Los distritos industriales son un fenómeno muy importante en Italia, país que ha encontrado en este modelo organizativo un rasgo peculiar de su economía y una fuente importante de desarrollo socio-económico. En Italia hay más de 200 distritos industriales, principalmente en el ramo textil, en la moda y en la industria de muebles. Este modelo creció a mediados de la década de 1970, cuando una serie de industrias y ciudades fueron un éxito económico. Los casos más notables fueron la industria textil en Carpi y Prato, la industria del mueble en Brianza y Cascina, y la industria del calzado en Vigevano, e incluso en Puglia; por primera vez la industria italiana de las máquinas-herramientas exportaba a toda Europa, las máquinas de envasado de Bolonia exportaban a Japón.

Varias regiones se han identificado como distritos industriales, debido a sus patrones de crecimiento, a su competitividad, su aglomeración y ciertas similitudes con el modelo de distrito industrial proporcionada por Marshall, o su variante de estilo italiano (Piore y Sabel, 1984). Los ejemplos norteamericanos más conocidos son las regiones de Hollywood, Silicon

Valley y el condado de Orange (Hall y Markusen, 1985), aunque muchos otros se han estudiado (Porter, 1998). En Reino Unido, los investigadores han identificado el área entre Londres y Bristol; en Francia, Grenoble, Montpellier y Sophia-Antipolis; en Suecia el distrito Gnosjo; en Alemania, Baden-Württemberg; algunas zonas de España y Dinamarca y otros fuera de Europa, como Ishikawa y otros en Japón (Friedman, 1988), la India, Brasil y México (Schmitz, 1995; Rabellotti, 1997).

En México se puede distinguir distritos industriales naturales que se originan como resultado de la evolución histórica de una o más industrias como el de cuero-calzado en la ciudad de León, Guanajuato y forzados como la industria electrónica en Guadalajara, cuyo desarrollo se debe a las políticas de fomento y apoyo a la instalación de empresas.

Un caso digno de mencionar lo constituye las empresas maquiladoras, plantas que importan materias primas, componentes y maquinaria para procesarlos o ensamblarlos en México y reexportarlos, principalmente a Estados Unidos y pagan impuestos sólo sobre el valor agregado. Estas empresas se aglomeraron preferentemente en el norte del país formando distritos industriales exitosos, pero a partir del año 2000 atravesaron por la principal crisis de su historia; se perdieron más de 300,000 empleos a nivel nacional, 60% de ellos en municipios fronterizos, y aproximadamente 890 fábricas maquiladoras fueron cerradas. Se calcula que la mitad de ellas, aproximadamente, fueron relocalizadas a los países asiáticos. Actualmente la industria maquiladora sigue operando, pero los distritos industriales se han reducido en algunas zonas y en otras han desaparecido.

El tercer ensayo de esta Tesis plantea la identificación de los distritos industriales en México agrupando primero municipios en función a los flujos de viajes al trabajo para después aplicar la metodología propuesta por Boix y Galleto (2005) para decidir cuáles de estos clusters tienen características de distritos industriales. Es importante señalar que en México no se tiene información censal sobre los flujos de viajes por lo que se estimarán mediante el modelo de radiación propuesto por Simini et al (2012).

La presente Tesis Doctoral cuyo eje central es investigar la movilidad laboral está estructurada de la forma siguiente: en la primera parte se revisan los modelos estadísticos que se utilizarán en las metodologías de los tres ensayos, a continuación se presentan los mismos para finalizar con las conclusiones y las líneas de investigación que se pueden derivar del trabajo desarrollado.

La estructura de cada ensayo contempla una introducción, revisión de la literatura, datos, metodología, evidencia empírica y conclusiones.

2. Modelos Estadísticos

2.1. Análisis de Conglomerados

Los métodos de análisis de grupos clasifican grupos de casos o elementos de acuerdo a criterios cualitativos o cuantitativos (distancias o similaridades). Entre los métodos estadísticos que tratan de analizar la pertenencia de casos a diversos grupos está el Análisis de Conglomerados o Cluster. En éste análisis no se tienen grupos predefinidos; éstos se definen mediante el cálculo de distancias o similaridades, a partir de los valores de algunas variables que se consideran adecuados para ello.

El objetivo de un análisis cluster es obtener grupos de objetos de forma, que por un lado, los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí y, por el otro, los objetos pertenecientes a grupos diferentes tengan un comportamiento distinto con respecto a las variables analizadas.

2.1.1. Medidas de similaridad y disimilaridad.

Un coeficiente de similaridad indica la fuerza de la relación entre dos puntos de datos (Everitt, 1993). Dos puntos de datos se parecen más entre sí, cuanto mayor sea su coeficiente de similaridad. Sea $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ y $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)$ dos puntos de datos d-dimensionales. Entonces, el coeficiente de similaridad entre $\boldsymbol{x}$ y $\boldsymbol{y}$ será alguna función de los valores de sus atributos, es decir,

$$s(\boldsymbol{x}, \boldsymbol{y}) = s(x_1, x_2, \ldots, x_d,\ y_1, y_2, \ldots, y_d\ ).$$

La similaridad generalmente es simétrica, es decir, $s\ (\boldsymbol{x}, \boldsymbol{y}) = s\ (\boldsymbol{y}, \boldsymbol{x})$. Constantine y Gower (1978), discutieron medidas de similaridad asimétricas. Una métrica es una función de distancia f definida en un conjunto E que satisface las cuatro propiedades siguientes (Anderberg, 1973; Zhang y Srihari, 2003):

1. No negatividad: $f\ (\boldsymbol{x}, \boldsymbol{y}) \geq 0$;
2. Reflexividad: $f\ (\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$;
3. Commutatividad: $f\ (\boldsymbol{x}, \boldsymbol{y}) = f\ (\boldsymbol{y}, \boldsymbol{x})$;
4. Desigualdad triangular: $f\ (\boldsymbol{x}, \boldsymbol{y}) \leq f\ (\boldsymbol{x}, \boldsymbol{z}) + f\ (\boldsymbol{y}, \boldsymbol{z})$,
   Donde $\boldsymbol{x}, \boldsymbol{y}$, y $\boldsymbol{z}$ son puntos de datos arbitrarios.

Una función de disimilaridad es una métrica definida en un conjunto. Por una función de similaridad, nos referimos a una función s $(\cdot,\cdot)$ medida en un conjunto de datos que satisface las siguientes propiedades (Kaufman y Rousseeuw, 1990):

1. $0 \leq s(\boldsymbol{x}, \boldsymbol{y}) \leq 1$,
2. $s(\boldsymbol{x}, \boldsymbol{x}) = 1$,
3. $s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$,

Donde $\boldsymbol{x}$, $\boldsymbol{y}$, y $\boldsymbol{z}$ son puntos de datos arbitrarios.

La elección de las distancias o de las medidas de similaridad es importante en las aplicaciones, y la mejor elección se logra a través de una combinación de experiencia, habilidad, conocimiento y suerte. Aquí se enumeran algunas distancias usadas comúnmente.

2.1.1.1. Matriz de covarianzas.

La covarianza es un concepto bien conocido en estadística. Sea D un conjunto con n objetos, los cuales se describen por medio de sus d atributos $v_1, v_2, \ldots, v_d$. Los atributos $v_1, v_2, \ldots, v_d$, se llaman también variables. La covarianza entre dos variables v_r, v_s se define como:

$$c_{rs} = \frac{1}{n}\sum_{i=1}^{n}(x_{ir} - \bar{x}_r)(x_{is} - \bar{x}_s)$$

Donde x_{ij} es el j-ésimo componente del punto de datos $\boldsymbol{x}_i$ y $\bar{x}_j$ es la media de todos los puntos de datos de la j-ésima variable, es decir:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad j = 1,2, \ldots., d$$

La matriz de covarianza es una matriz $d \times d$ en la cual el elemento (r, s) es la covarianza entre las variables v_r y v_s.

$$\sum = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1d} \\ c_{21} & c_{22} & \cdots & c_{2d} \\ \vdots & \vdots & & \vdots \\ c_{d1} & c_{d2} & & c_{dd} \end{pmatrix}$$

La matriz de covarianza se puede escribir como:

$$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$$

Donde $\boldsymbol{X}^T$ es la matriz transpuesta de **X** y **X** es una matriz $n \times d$ con el (i,j)-ésimo elemento, $x_{ij} - \bar{x}_j$, es decir

$$\boldsymbol{X} = \left(x_{ij} - \overline{x}_j\right)_{n\times d} = \begin{pmatrix} \boldsymbol{x}_1 - \overline{x}_1\,\boldsymbol{e}_d \\ \boldsymbol{x}_2 - \overline{x}_2\,\boldsymbol{e}_d \\ \vdots \\ \boldsymbol{x}_d - \overline{x}_d\,\boldsymbol{e}_d \end{pmatrix}$$

Donde $\boldsymbol{e}_d$ es el vector identidad d-dimensional, es decir, $\boldsymbol{e}_d = (1, 1, \ldots., 1)$.

2.1.1.2. Similaridad coseno

La medida de similaridad coseno fue propuesta por Salton y McHill (1983) y también por Xiao y Dunham(2001). La similaridad coseno es una medida existente entre dos vectores en un espacio que posee un producto interno con el que se evalúa el valor del coseno del ángulo comprendido entre ellos. Esta función trigonométrica proporciona un valor igual a 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar.

Para cualquier ángulo existente entre los vectores, el coseno arrojará un valor inferior a uno. Si los vectores fuesen ortogonales el coseno se anularía, y si apuntasen en sentido contrario su valor sería -1. De esta forma, el valor de esta medida se encuentra entre -1 y 1, es decir en el intervalo cerrado [-1,1].

La similaridad coseno para dos vectores de puntos de datos está dada por:

$$cos(t_i, t_j) = \frac{\langle t_i, t_j \rangle}{\|t_i\| \cdot \|t_j\|}$$

Donde ⟨·,·⟩ representa el producto interno y ‖·‖ se refiere a la norma del vector.

2.1.1.3. Distancia Euclídea

La distancia euclídea es probablemente la distancia más comúnmente utilizada para datos numéricos. Para dos puntos de datos $\boldsymbol{x}$ y $\boldsymbol{y}$ en el espacio d-dimensional, la distancia euclídea entre ellos, se define como:

$$d_{euc}(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{j=1}^{d} (x_j - y_j)^2\right]^{\frac{1}{2}} = [(\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^T]^{\frac{1}{2}}$$

La distancia euclídea al cuadrado viene dada por:

$$d_{euc\ cuad}(\boldsymbol{x}, \boldsymbol{y}) = d_{euc}(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j=1}^{d} (x_j - y_j)^2 = (\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^T$$

Se debe observar que de hecho la distancia euclídea al cuadrado no es una distancia.

2.1.1.4. Distancia de Manhattan

La distancia de Manhattan también llamada “distancia city block”, define la distancia entre dos puntos como la suma de las diferencias (absolutas) de sus coordenadas. Es decir, para dos puntos de datos $\boldsymbol{x}$ y $\boldsymbol{y}$ en el espacio d-dimensional, la distancia de Manhattan entre ellos es:

$$d_{man}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{d} |x_j - y_j|$$

Si los puntos de datos $\boldsymbol{x}$ ó $\boldsymbol{y}$ tienen valores perdidos, entonces la distancia de Manhattan se define como (Wishart, 2002):

$$d_{manw}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{d} \frac{w_j|x_j - y_j|}{\sum_{k=1}^{d} w_j}$$

donde $w_j = 1,$ si tanto $\boldsymbol{x}$ como $\boldsymbol{y}$ tienen observaciones del j-ésimo atributo y $w_j = 0$ en caso contrario.

La distancia segmental de Manhattan es una variante de la distancia de Manhattan. En la distancia segmental de Manhattan, sólo una parte de toda la dimensión se utiliza para calcular la distancia. Se define como (Aggarwal et al., 1999):

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \sum_{j\varepsilon P} \frac{|x_j - y_j|}{|P|}$$

donde P es un conjunto no vacío de $\{1, 2, \ldots, d\}$.

2.1.1.5. Distancia Máxima

La distancia máxima también llamada distancia "sup", se define como el máximo valor de las distancias de los atributos; es decir, para dos puntos de datos $\boldsymbol{x}$ y $\boldsymbol{y}$ en el espacio d-dimensional, la distancia máxima entre ellos es:

$$d_{max}(\boldsymbol{x},\boldsymbol{y}) = \max_{1\leq k\leq d} |x_j - y_j|$$

2.1.1.6. Distancia de Minkowski

La distancia euclidiana, la distancia de Manhattan, y la distancia máxima son tres casos particulares de la distancia definida por Minkowski

$$d_{min}(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{j=1}^{d} |x_j - y_j|^r\right)^{\frac{1}{r}}, \quad r \geq 1$$

r es el orden de la distancia Minkowski. Sí r toma los valores de $2, 1, \infty$, se obtienen respectivamente la distancia euclídea, la distancia de Manhattan, y la distancia máxima. Si el conjunto de datos tiene conglomerados compactos o separados, la distancia de Minkowski funciona bien (Mao y Jain, 1996); de lo contrario el atributo de mayor escala tiende a dominar a los demás; para evitar esto, se debe estandarizar o utilizar esquemas de ponderación (Jain et al., 1999).

2.1.1.7. Distancia de Mahalanobis

La Distancia de Mahalanobis (Jain y Dubes, 1988; Mao y Jain, 1996), evita la distorsión de la distancia causada por las combinaciones lineales de los atributos. Se define por:

$$d_{mah}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{y})^T}$$

Donde $\boldsymbol{\Sigma}$ es la matriz de covarianzas del conjunto de datos. Por lo tanto, esta distancia aplica un esquema ponderado. Otra propiedad importante de la distancia de Mahalanobis es ser invariante bajo todas las transformaciones no singulares. Por ejemplo, sea C cualquier matriz no singular $d \times d$ aplicada a los datos originales $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$

$$\boldsymbol{y}_i = C\boldsymbol{x}_i, \quad i = 1, 2, \ldots, n.$$

La nueva matriz de covarianza se convierte en:

$$\frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y} = \frac{1}{n}(\boldsymbol{X}C^T)^T(\boldsymbol{X}C^T)$$

La distancia de Mahalanobis entre $\boldsymbol{y_i}$ y $\boldsymbol{y_j}$es:

$$d_{mah}(\boldsymbol{x}, \boldsymbol{y})$$

$$= \sqrt{(\boldsymbol{y}_i - \boldsymbol{y}_j)\left(\frac{1}{n}Y^T\boldsymbol{Y}\right)^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j)^T}$$

$$= \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)C^T \left(\frac{1}{n}(\boldsymbol{X}^T C^{\boldsymbol{T}})(\boldsymbol{X}C^T)\right)^{-1} C(\boldsymbol{x}_i - \boldsymbol{x}_j)^T}$$

$$= \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)C^T \left(\frac{1}{n}(\boldsymbol{X}^T \boldsymbol{X})\right)^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)^T}$$

$$= d_{mah}(\boldsymbol{x_i}, \boldsymbol{x_j})$$

Lo cual muestra que la distancia de Mahalanobis es invariante bajo transformaciones no singulares.

Morrison (1967), propuso una distancia generalizada de Mahalanobis mediante la ponderación de las variables. Sea $\lambda_j (j = 1, 2, \ldots, d)$ los pesos asignados a la j-ésima variable y la matriz diagonal $d \times d$ contiene los pesos d, es decir:

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}$$

Entonces la distancia generalizada de Mahalanobis es:

$$d_{gmah}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})\Lambda \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})^T}$$

La distancia de Mahalanobis tiene algunas desventajas como requerir mucho esfuerzo computacional, ya que la matriz de covarianzas se calcula con todos los puntos de datos.

2.1.1.8. Distancia Media

Legendre y Legendre (1983), señalaron el siguiente inconveniente de la distancia euclídea: Dos puntos de datos sin valores de atributo en común pueden tener una distancia menor que otro par de puntos de datos que contienen los mismos valores del atributo. Para estos casos propusieron el uso de la distancia media.

La distancia media se modifica a partir de la distancia euclídea. Dados dos puntos de datos $\boldsymbol{x}$ y $\boldsymbol{y}$ en un espacio d-dimensional, la distancia media se define por:

$$d_{prom}(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{1}{d} \sum_{J=1}^{d} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2 \right)^{\frac{1}{2}}$$

En la Tabla 1, se muestran otras medidas de disimilitud para variables numéricas.

2.1.2. Estandarización de los datos

En muchas aplicaciones de análisis de conglomerados, los datos crudos o de medidas reales, no se utilizan directamente a menos que se cuente con un modelo probabilístico para la generación de un patrón (Jain y Dubes, 1988). La preparación de datos para el análisis de conglomerados requiere algún tipo de transformación, como es la estandarización o normalización.

Es necesario estandarizar las variables en casos en que la medida de disimilaridad, tal como la distancia euclidiana, sea sensible a las diferencias en las magnitudes o escalas de la variables de entrada (Milligan y Cooper, 1988). Después de la estandarización, todo el conocimiento de la localización y la escala de los datos originales se puede perder; la escala de los datos originales se pierde ya que la estandarización de los datos los hace adimensionales. El enfoque de la estandarización o normalización de las variables es esencialmente de dos tipos: estandarización global y estandarización dentro de la agrupación.

La estandarización global estandariza las variables a través de todos los elementos del conjunto de datos. La normalización o estandarización dentro del cluster se refiere a la estandarización que se realiza a cada variable dentro de los grupos. Algunas formas de estandarización se pueden utilizar tanto en la estandarización global como en la estandarización dentro de los grupos, pero otras formas sólo se pueden utilizar en la estandarización global.

Tabla 1. Otras medidas de disimilitud para variables numéricas

Medida	$d(\boldsymbol{x}, \boldsymbol{y})$	Referencia
Diferencia media carácter	$\frac{1}{d}\sum_{j=1}^{d}\lvert x_i - y_j\rvert$	Czekanowski (1909)
Índice de asociación	$\frac{1}{2}\sum_{j=1}^{d}\left\lvert \frac{x_i}{\sum_{l=1}^{d} x_l} - \frac{y_j}{\sum_{l=1}^{d} y_l}\right\rvert$	Whittaker (1952)
Métrica de Canberra	$\sum_{j=1}^{d}\frac{\lvert x_i - y_j\rvert}{(x_i + y_j)}$	Legendre y Legendre (1983)
Coeficiente Czekanowski	$1 - \frac{2\sum_{j=1}^{d} min\{x_i, y_j\}}{\sum_{j=1}^{d}(x_i + y_j)}$	Johnson y Wichem (1998)
Coeficiente de divergencia	$\left(\frac{1}{d}\sum_{j=1}^{d}\left(\frac{x_i - y_j}{x_i + y_j}\right)^2\right)^{\frac{1}{2}}$	Legendre y Legendre (1983)

Es imposible estandarizar directamente las variables dentro de las agrupaciones en el análisis de conglomerados, ya que los grupos no se conocen antes de la normalización. Para superar esta dificultad, Overall y Klett (1972), propusieron un enfoque iterativo, que primero obtiene agrupaciones basadas en estimaciones globales y luego utiliza estos grupos para ayudar a determinar las varianzas dentro del grupo, para estandarizar en el análisis del segundo cluster.

Milligan y Cooper (1988), presentaron un examen a fondo de la estandarización o normalización de las variables cuando se utiliza la distancia euclídea como medida de disimilaridad.

Antes de revisar varios métodos para la normalización de datos, se debe remarcar que la elección adecuada de un método de estandarización depende del conjunto de datos original y del particular campo de estudio.

Para estandarizar los datos crudos, se resta una medida de localizacióny se divide por una medida de escala para cada variable. Esto es,

$$x_{ij} = \frac{x_{ij}^{*} - L_j}{M_j}$$

Se obtienen diferentes métodos de normalización eligiendo diferentes valores de L_j y M_j en la ecuación anterior. Algunos métodos de normalización bien conocidos son la media, la mediana, la desviación estándar, el rango, la estimación de Huber, la estimación biponderada de Tukey, la estimación de onda de Andrew. En la Tabla 2, se muestran algunas formas de estandarización, donde $\bar{x}_j^*$, R_j^* y σ_j^* son la media, el rango y la desviación estándar de la variable j-ésima, es decir:

$$\bar{x}_{ij}^* = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^*$$

$$R_j^* = \max_{1\le i\le n} x_{ij}^* - \min_{1\le i\le n} x_{ij}^*$$

$$\sigma_j^* = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}^* - \bar{x}_j^*\right)^2\right]^{\frac{1}{2}}$$

Transformar a puntuaciones z, es una forma de estandarización utilizada para transformación normal. Dado un conjunto de datos crudos D^*, la fórmula para obtener las puntuaciones z es:

$$x_{ij} = \boldsymbol{Z}_1\left(x_{ij}^*\right) = \frac{x_{ij}^* - \bar{x}_j^*}{\sigma_j^*}$$

La variable transformada tiene una media igual a cero y una varianza igual a uno. Con este tipo de estandarización se pierde la localización y la información de la escala de la variable original.

Una restricción importante de la estandarización a puntuaciones z, es que sólo debe aplicarse en la normalización o estandarización global y no dentro del cluster (Milligan y Cooper, 1988). Por ejemplo, en el caso en el que existen dos cluster bien separados, si una de las muestras se encuentra en cada uno de los dos centroides, la normalización dentro del grupo estandarizaría las muestras situadas en los centroides como vectores cero; cualquier algoritmo de agrupamiento uniría los dos vectores cero, lo que significa que las dos muestras originales serían agrupados en un cluster, lo que sería una agrupación muy engañosa.

Tabla 2. Algunos métodos de estandarización

Nombre	L_j	M_j
Puntuaciones z	$\bar{x}_j^*$	σ_j^*
USTD	0	σ_j^*
Máximo	0	$\max_{1\le i\le n} x_{ij}^*$
Media	$\bar{x}_j^*$	1
Mediana	$x^*_{\frac{n+1}{2}j}$ sí n es non $\frac{1}{2}\left(x^*_{\frac{n}{2}j} + x^*_{\frac{n+2}{2}j}\right)$ sí n es par	1
Suma	0	$\sum_{i=1}^{n} x_{ij}^*$
Rango	$\min_{1\le i\le n} x_{ij}^*$	R_j^*

La estandarización de la desviación estándar ponderada sin corregir, USTD, es similar a la estandarización a puntuaciones z y se define como:

$$x_{ij} = \boldsymbol{Z}_2\left(x_{ij}^*\right) = \frac{x_{ij}^*}{\sigma_j^*}$$

La variable transformada $\boldsymbol{Z}_2$ tendrá una varianza igual a 1. Dado que los resultados no se han centrado restando la media, la información de localización de las puntuaciones no se pierde. Por lo tanto, la estandarización $\boldsymbol{Z}_2$ no presentará el problema de la pérdida de información acerca de la centroides del cluster.

El tercer método de estandarización presentado por Milligan y Cooper (1988), utiliza la puntuación máxima de la variable:

$$x_{ij} = \boldsymbol{Z}_3\left(x_{ij}^*\right) = \frac{x_{ij}^*}{\max_{1\le i\le n} x_{ij}^*}$$

Una variable X transformada por $\boldsymbol{Z}_3$ tendrá una media igual a $\frac{\bar{X}}{max(X)}$ y una desviación estándar igual a $\frac{\sigma_X}{max(X)}$, donde $\bar{X}$ y σ_X son la media y la desviación estándar de la variable original. $\boldsymbol{Z}_3$ es susceptible a la presencia de valores atípicos (Milligan y Cooper, 1988). Si hay un solo valor muy grande, $\boldsymbol{Z}_3$, estandarizará los valores restantes hasta casi 0. $\boldsymbol{Z}_3$ parece ser significativa sólo cuando la variable esté medida en una escala de razón (Milligan y Cooper, 1988).

Milligan y Cooper (1988), propusieron dos métodos de estandarización utilizando el rango de la variable:

$$x_{ij} = \boldsymbol{Z}_4\left(x_{ij}^*\right) = \frac{x_{ij}^*}{R_j^*}$$

$$x_{ij} = \boldsymbol{Z}_5\left(x_{ij}^*\right) = \frac{x_{ij}^* - \min\limits_{1\leq i\leq n} x_{ij}^*}{R_j^*}$$

donde R_j^* es el rango del j-ésimo atributo.

Una variable $\boldsymbol{X}$ transformada mediante $\boldsymbol{Z}_4$ y $\boldsymbol{Z}_5$ tendrá una media igual a $\frac{\bar{X}}{max(X)-min(X)}$, $\frac{\bar{X}-min(X)}{max(X)-min(X)}$ respectivamente y tendrán la misma desviación estandar igual a $\frac{\sigma_X}{max(X)-min(X)}$.

Las transformaciones $\boldsymbol{Z}_4$ y $\boldsymbol{Z}_5$ son sensibles a la presencia de valores extremos*, outliers*.

Milligan y Cooper (1988), también presentaron una estandarización basada en la suma de las observaciones:

$$x_{ij} = \boldsymbol{Z}_6\left(x_{ij}^*\right) = \frac{x_{ij}^*}{\sum_{i=1}^{n} x_{ij}^*}$$

La transformación estandariza la suma de los valores transformados a la unidad y la media transformada será igual a $\frac{1}{n}$, por lo que la media será constante a través de todas las variables.

Un enfoque muy diferente de la estandarización consiste en realizar un ranking o clasificación de los resultados (Milligan y Cooper, 1988):

$$x_{ij} = \boldsymbol{Z}_7\left(x_{ij}^*\right) = Ranking\left(x_{ij}^*\right)$$

Una variable transformada mediante $\boldsymbol{Z_7}$, tendrá una media igual a $\frac{n+1}{2}$ y una desviación estándar igual a $(n+1)\left(\frac{2n+1}{6}-\frac{n+1}{4}\right)$. Esta transformación reduce el impacto de los valores extremos. Conover and Iman (1981), sugirieron cuatro tipos de transformación de *ranking*. La primera transformación consiste en ordenar los valores del más pequeño al más grande y entonces asignar un *ranking* igual a 1 al valor más pequeño, un *ranking* igual a 2 al siguiente y así sucesivamente; en caso de empate se asigna el promedio de los valores de *ranking* que les corresponda.

2.1.3. Conglomerados Jerárquicos

Los algoritmos fuertes de agrupamiento se subdividen en algoritmos jerárquicos y algoritmos particionales. Un algoritmo particional divide un conjunto en una sola partición de datos, mientras que un algoritmo jerárquico divide un conjunto de datos en una secuencia de particiones anidadas. Los algoritmos jerárquicos se subdividen en algoritmos jerárquicos aglomerativos y algoritmos jerárquicos divisivos, como se muestra en la Figura 1. La agrupación jerárquica aglomerativa comienza con clusters o conglomerados de un solo objeto. Luego se repite la unión de la pareja más cercana de los conglomerados de acuerdo con algunos criterios de similaridad hasta que todos los datos están en un solo cluster. La agrupación jerárquica aglomerativa tiene algunas desventajas como es que los puntos de datos que se han agrupado de forma incorrecta en una etapa temprana no pueden reasignarse y que las diferentes medidas utilizadas para medir la similaridad entre clusters pueden conducir a resultados diferentes. Si tratamos la agrupación jerárquica aglomerativa como método de agrupamiento de abajo hacia arriba, la agrupación jerárquica divisiva puede ser vista como un método de agrupación de arriba hacia abajo. La agrupación divisiva jerárquica comienza con todos los objetos en un solo cluster y se repite la división de conglomerados grandes en grupos más pequeños. La agrupación jerárquica divisiva tiene los mismos inconvenientes que la agrupación jerárquica aglomerativa. En la Figura 1, se da un ejemplo de agrupación jerárquica aglomerativa y de agrupación jerárquica divisiva. Los algoritmos jerárquicos se pueden expresar mediante gráficas o algebra de matrices (Jain y Dubes, 1988). El dendograma, un tipo especial de estructura de árbol, es de uso frecuente para visualizar una agrupación jerárquica.

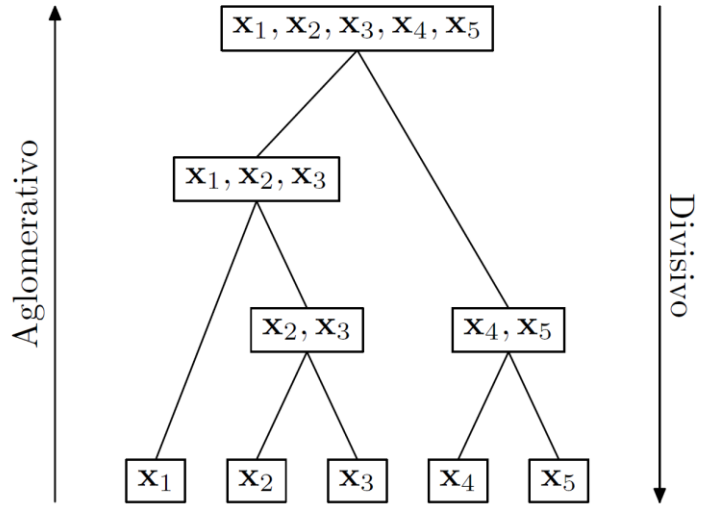

Figura 1. . Algoritmos Jerárquicos

2.1.3.1. Representación de conglomerados jerárquicos

Una agrupación jerárquica se puede representar mediante gráficas o una serie de símbolos abstractos. La gráfica de un agrupamiento jerárquico es mucho más fácil de interpretar. Los símbolos abstractos de una agrupación jerárquica se pueden utilizar internamente para mejorar el rendimiento del algoritmo. En esta sección, se revisan algunas representaciones comunes de agrupamientos jerárquicos.

n-árbol

Una agrupación jerárquica se representa generalmente por un diagrama de árbol. Un n-árbol es un diagrama de árbol simple jerárquicamente anidado que se puede utilizar para representar una agrupación jerárquica. Sea $D = \{x_1, x_2 \ldots, x_n\}$ un conjunto de objetos, un n-árbol en D se define como un conjunto T de subconjuntos de D sí satisfacen las siguientes condiciones (Bobisud y Bobisud, 1972; McMorris et al., 1983; Gordon, 1996):

1. $D \in T$;
2. El conjunto vacío $\emptyset \in T$;
3. $\{x_i\} \in T$ para todo $i = 1, 2, \ldots, N$;
4. Sí $A, B \in T$, entonces $A \cap B \in \{\emptyset, A, B\}$.

Un 5-árbol se ilustra en la Figura 2. Los nodos terminales u hojas representadas por una círculo abierto representa un solo punto de datos. Los nodos internos representados por un círculo relleno representan un conglomerado o cluster. A los n-árbol también se les conoce como árboles no rankeados (Murtagh, 1984b). Si un n-árbol tiene precisamente $n-1$ nodos internos, el árbol es un árbol binario o árbol dicotómico. Los diagramas de árbol, así como los n-árbol y los dendrogramas, discutidos más adelante, contienen muchas indeterminaciones. Por ejemplo, el orden de los nodos internos y el orden de las hojas se pueden intercambiar. También, los diagramas de árbol tienen muchas variaciones. Por ejemplo, al girar el árbol 90 grados se obtiene un árbol horizontal. Propiedades alternativas de árboles se presentan en (Hartigan,1967) y (Constantinescu, 1966).

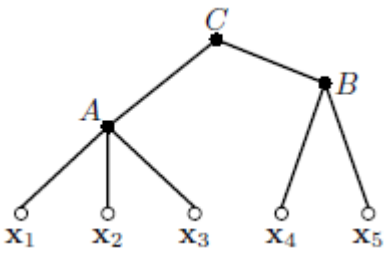

Figura 2. 5-árbol

Dendrograma

Un dendrograma es un n-árbol en el que cada nodo interno está asociado con una altura que satisface la condición:

$$h\,(A) \leq h\,(B) \Leftrightarrow A \subseteq B$$

para todos los subconjuntos de puntos de datos A y B, sí $A \cap B \neq \emptyset$, donde $h\,(A)\; y\; h\,(B)$ denotan las alturas de A y B, respectivamente.

A modo de ejemplo, la Figura 3, muestra un dendrograma con cinco puntos de datos. Las líneas punteadas indican las alturas de los nodos internos. Para cada par de puntos de datos (x_i, x_j), h_{ij} es la altura del nodo interno especificando el grupo más pequeño al que tanto x_i y

x_j pertenecen, por lo que un valor pequeño de h_{ij} indica una alta similaridad entre x_i y x_j. En el dendrograma de la Figura 3, por ejemplo, tenemos $h_{12} = 1$, $h_{23} = h_{13} = 3$ y $h_{14} = 4$.

Las alturas en el dendrograma satisfacen las siguientes condiciones de ultra métrica (Johnson,1967):

$$h_{ij} \leq max\{h_{ik}, h_{jk}\} \quad \forall i, j, k \in \{1, 2, \ldots, N\}.$$

De hecho, la condición ultra métrica también es una condición necesaria y suficiente para un dendrograma (Gordon, 1987).

Matemáticamente, un dendrograma se puede representar por una función de $c: [0, \infty) \rightarrow E(D)$ que satisface (Sibson, 1973)

1. $c(h) \subseteq c(h')$
2. Si $h \leq h', c(h)$ está eventualmente $D \times D$,
3. $c(h + \delta) = C(H)$ para algunas pequeñas $\delta > 0$,
   donde D se ajusta a los datos dados y $E(D)$ es el conjunto de relaciones de equivalencia en D.

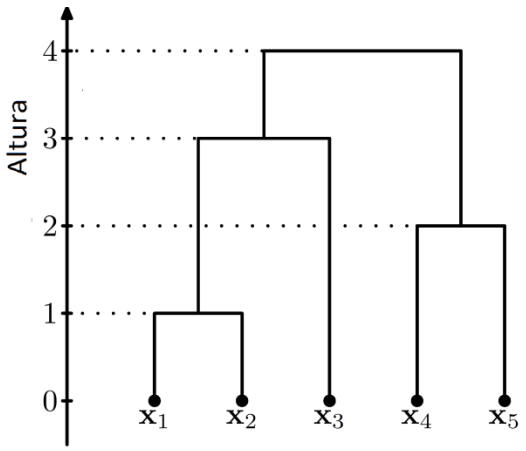

Figura 3. Dendrograma

Como ejemplo, la función *c* que se da a continuación contiene la información del dendrograma que se muestra en la Figura 3:

$$c(h) = \begin{cases} \{(i,i) : i = 1,2,3,4,5\} & si\ 0 \le h < 1 \\ \{(i,i) : i = 3,4,5\} \cup \{(i,\ j) : i,\ j = 1,2\} & si\ 1 \le h < 2 \\ \{(3,3)\} \cup \{(i,\ j) : i,\ j = 1,2\} \cup \{(i,\ j) : i,\ j = 4,5\} & si\ 2 \le h < 3 \\ \{\{(i,\ j) : i,\ j = 4,5\} \cup \{(i,\ j) : i,\ j = 1,2,3\} & si\ 3 \le h < 4 \\ \{(i,\ j) : i,\ j = 1,2,3,4,5\} & si\ 4 \le h \end{cases}$$

Otras caracterizaciones de un dendrograma se han presentado en (Johnson, 1967), (Jardine et al., 1967), y (Banfield, 1976). Van Rijsbergen (1970), sugirió un algoritmo para encontrar el dendrograma de enlace único de la matriz de disimilaridad de entrada. Algoritmos para el trazado de dendrogramas se discuten en (Rohlf, 1974), (Gower y Ross, 1969), y (Ross, 1969). Sokal y Rohlf, (1962), discutieron la comparación de los dendrogramas.

2.1.3.2. Métodos jerárquicos aglomerativos

De acuerdo con las diferentes medidas de distancia entre los grupos, los métodos jerárquicos aglomerativos pueden subdividirse en métodos de vinculación simple, vinculación completa, media de grupos, promedio ponderado, centroide, Ward y mediana.

Los métodos de vinculación simple, completa, media y promedio ponderado también se conocen como métodos gráficos, mientras que el método de Ward, el método centroide, y el método de la mediana hace referencia a métodos geométricos (Murtagh, 1983), ya que en los métodos gráficos un cluster puede ser representado por una subgráfica de puntos interconectados y en los métodos geométricos un cluster pueden ser representado por un punto central.

Murtagh (1983), proporcionó un inventario de los algoritmos de agrupamiento jerárquico, especialmente de los algoritmos de agrupamiento jerárquico aglomerativos. El desempeño de los algoritmos de agrupación jerárquicos se puede mejorar mediante la incorporación eficiente de la búsqueda del vecino más cercano en los algoritmos.

2.1.3.3. Fórmula de Lance y Williams

En los algoritmos de agrupamiento jerárquico aglomerativos, la fórmula de Lance-Williams se utiliza para calcular la disimilaridad entre un grupo y un grupo formado mediante la fusión de otras dos clusters. Lance y Williams (1967a) propusieron una fórmula de recurrencia que da la distancia entre un C_k cluster y un grupo C formado por la fusión de los grupos C_i y C_j, es decir, $C = C_i \cup C_j$. La fórmula está dada por:

$$
\begin{aligned}
&D(C_k, C_i \cup C_j) \\
&= \alpha_i D(C_k, C_i) + \alpha_j D(C_k, C_j) \\
&+\beta D(C_i, C_j) + \gamma \left| D(C_k, C_i) - D(C_k, C_j) \right|
\end{aligned}
$$

donde $D(\cdot,\cdot)$ es la distancia entre dos clusters

Para una elección adecuada de α_iy de los parámetros, α_j, β, y γ, se pueden obtener varias distancias interclusters utilizados por los algoritmos de agrupamiento jerárquico. En la Tabla 3, se muestran los valores comúnmente usados en la fórmula Lance-Williams, donde $\mathrm{n_i} = |\mathrm{C_i}|$ es el número de puntos de datos en $\mathrm{C_i}$ y $\sum_{\mathrm{ijk}} = \mathrm{n_i} + \mathrm{n_j} + \mathrm{n_k}$.

2.1.3.4. Método de Vinculación Simple

El método de vinculación simple es uno de los métodos más simples de agrupamiento jerárquico. Fue introducido por Florek et al. (1951) y luego de forma independiente por McQuitty (1957) y Sneath (1957). El método de vinculación simple es también conocido por otros nombres, como el método del vecino más cercano, el método del mínimo, y el método de conexión (Rohlf, 1982).

El método de vinculación simple es invariante bajo transformaciones monótonas como por ejemplo, la transformación logarítmica de los datos originales (Johnson, 1967).

Emplea la distancia del vecino más cercano para medir la disimilaridad entre dos grupos. Sean C_i, C_j, C_k tres grupos de puntos de datos. Entonces la distancia entre C_ky $C_i \cup C_j$ se puede obtener a partir de la fórmula Lance-Williams como sigue:

$$D(Ck, Ci \cup Cj) =$$
$$= 1/2\, D(C_k, C_i) + 1/2 D(C_k, C_j) - 1/2|D(C_k, C_i) - D(C_k, C_j)|$$
$$= min\{D(C_k, C_i), D(Ck, C_j)\},$$

donde $D(\cdot,\cdot)$ es la distancia entre dos cluster.

Tabla 3. Valores comúnmente usados para los parámetros en la fórmula Lance-Williams.

Algoritmo	$\propto_i$	$\propto_j$	β	γ
Vinculación individual	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Vinculación completa	$\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$
Método de Ward	$\frac{n_i + n_j}{\sum_{ijk}}$	$\frac{n_i + n_k}{\sum_{ijk}}$	$\frac{-n_k}{\sum_{ijk}}$	0
Promedio de grupo	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Promedio de grupo ponderado	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroide	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0

Rohlf (1982) ha clasificado los algoritmos de vinculación simple en cinco tipos diferentes:

1. Algoritmos de conexión,
2. Algoritmos basados en una transformación ultra métrica,
3. Algoritmos de estimación de densidad de probabilidad,
4. Algoritmos aglomerativos
5. Algoritmos basados en el árbol de expansión mínimo

Los algoritmos de conexión se basan en la teoría de grafos. En un algoritmo de conexión los puntos de datos se representan como vértices en una gráfica: Un par (i, j) de vértices se conectan con un borde sí y sólo si la distancia entre puntos de datos (i y j) $dij \leq \Delta$. Los

grupos de enlace único a nivel Δ corresponden a los subgrafos conectados de la gráfica. Los algoritmos de conección requieren una cantidad considerable de esfuerzo computacional.

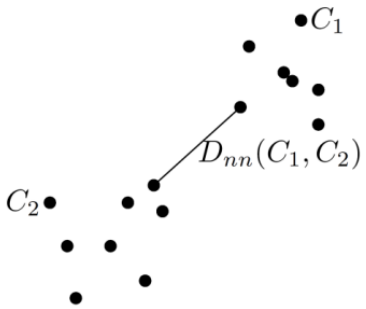

Figura 4. Vecino más cercano

2.1.3.5. Método de Vinculación Completa

A diferencia del método de enlace único, el método de vinculación completa utiliza la distancia del vecino más lejano para medir la disimilaridad entre dos grupos.

El método de vinculación completa es invariante bajo transformaciones monótonas (Johnson, 1967). Sean C_i, C_j y C_k tres grupos de puntos de datos. Entonces la distancia entre C_k y $C_i \cup C_j$ puede obtenerse de la fórmula Lance-Williams como sigue:

$$D(C_k, C_i \cup C_j)$$
$$= 1/2\, D(C_k, C_i) + 1/2 D(C_k, C_j) + 1/2|D(C_k, C_i) - D(C_k, C_j)|$$
$$= max\{D(C_k, C_i), D(C_k, C_j)\},$$

donde $D(\cdot,\cdot)$ es la distancia entre dos cluster.

La distancia definida en la ecuación anterior tiene la siguiente propiedad:

$$D(C, C') = \max_{\boldsymbol{x}\varepsilon C, \boldsymbol{y}\varepsilon C'} d(\boldsymbol{x}, \boldsymbol{y})$$

donde C y C' son dos, grupos no vacíos que no se superponen y $d(\cdot,\cdot)$ es la función de distancia con la cual se calcula la matriz de disimilaridad.

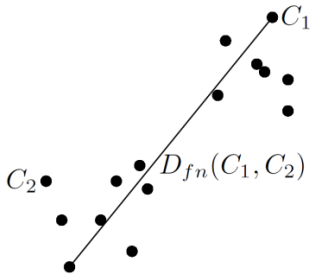

Figura 5. Vecino más lejano

2.1.3.6. Método de la Vinculación Promedio

El método de la vinculación promedio, también se conoce como UPGMA, método del par no ponderado del grupo utilizando promedios aritméticos (Jain y Dubes, 1988). En el método de la vinculación promedio del grupo, la distancia entre dos grupos se define como el promedio de las distancias entre todos los posibles pares de puntos de datos que pueden formarse tomando un miembro del grupo y otro miembro del otro. Sean C_i, C_j y C_k tres grupos de puntos de datos. Entonces la distancia entre C_k y $C_i \cup C_j$ puede obtenerse a partir de la fórmula Lance-Williams como sigue:

$$D(C_k, C_i \cup C_j)$$

$$= \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j)$$

donde $D(\cdot,\cdot)$ es la distancia entre dos clusters.

Sean C y C' dos grupos no vacíos y que no se superponen entonces:

$$D(C, C') = \frac{1}{|C||C'|} \sum_{\boldsymbol{x}\varepsilon C, \boldsymbol{y}\epsilon C'} d(\boldsymbol{x}, \boldsymbol{y})$$

donde $d(\cdot,\cdot)$ es la función de distancia con la cual se construye la matriz de disimilaridad.

Sean C_1, C_2, C_3 tres conglomerados no vacíos y mutuamente excluyentes, entonces:

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum (C_i, C_j), \qquad 1 \leq i < j \leq 3$$

donde $n_i = |C_i|$, $n_j = |C_j|$, y $\sum(C_i, C_j)$ es la distancia total entre los clustersC_i y C_j, dada por:

$$\sum (C_i, C_j) = \sum_{x \varepsilon C_i, y \varepsilon C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

De las ecuaciones anteriores tenemos:

$$D(C_1, C_2 \cup C_3)$$

$$= \frac{n_2}{n_2 + n_3} D(C_1, C_2) + \frac{n_3}{n_2 + n_3} D(C_1, C_3)$$

$$= \frac{n_2}{n_2 + n_3} \times \frac{1}{n_1 n_2} \sum (C_1, C_2) + \frac{n_3}{n_2 + n_3} \times \frac{1}{n_1 n_3} \sum (C_1, C_3)$$

$$\frac{1}{n_1 (n_2 + n_3)} \sum (C_1, C_2 \cup C_3)$$

ya que $(C_1, C_2) + (C_1, C_3) = (C_1, C_2 \cup C_3)$.

$$d_{AB} = (d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})/6$$

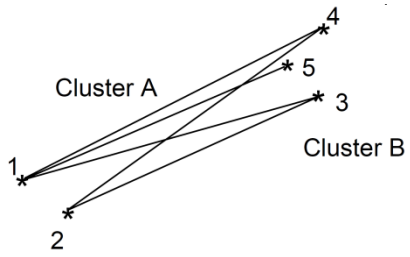

Figura 6. Media del grupo

El método del promedio puede producir vinculaciones intergrupos o intra grupos. La vinculación intra grupos mide la distancia entre dos conglomerados mediante la media aritmética de todas las posibles distancias entre pares, de forma que cada componente del par pertenece a un conglomerado distinto, mientras que la vinculación intra grupos considera la media aritmética de las distancias entre todos los pares, pertenezcan al mismo grupo o a grupos distintos.

2.1.3.7. Método del Centroide.

El método del centroide también se conoce como el "método del grupo de pares no ponderados usando centroides" (Jain y Dubes, 1988). Con el método del centroide, las distancias entre los clusters se pueden calcular mediante la fórmula de Lance-Williams:

$$D(C_k, C_i \cup C_j)$$

$$= \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) - \frac{|C_i||C_j|}{(|C_i| + |C_j|)^2} D(C_i, C_j) \quad (2.1)$$

donde C_k, C_i, C_j son tres grupos en el mismo nivel de agrupamiento.

Sea C y C' dos cluster cualquiera no traslapados, es decir, $C \cap C' = \emptyset$. Entonces de la ecuación 2.1, tenemos:

$$D(C, C')$$

$$= \frac{1}{|C||C'|} \sum d(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{2|C|^2} \sum_{\boldsymbol{x},\boldsymbol{y} \varepsilon C} d(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{2|C'|^2} \sum_{\boldsymbol{x},\boldsymbol{y} \varepsilon C'} d(\boldsymbol{x}, \boldsymbol{y}) \quad (2.2)$$

donde $d(\cdot,\cdot)$ es la función de distancia utilizada para calcular la matriz de disimilaridad.

Sea C_1, C_2 y C_3 tres clusters no vacíos y mutuamente excluyentes, se asume que:

$$D(C_i, C_j) = \frac{1}{n_i n_j}\sum(C_i, C_j) - \frac{1}{2{n_i}^2}\sum(C_i) - \frac{1}{2{n_j}^2}\sum(C_j) \qquad (2.3)$$

para $1 \leq i < j \leq 3$, donde $n_i = |C_i|$, $n_j = |C_j|$, $\sum(C_i, C_j)$ es la distancia total entre los cluster de C_i y C_j, es decir:

$$\sum(C_i, C_j) = \sum_{\boldsymbol{x}\varepsilon C_i, \boldsymbol{y}\varepsilon C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

$$\sum(C_i) = C_i = \sum_{\boldsymbol{x}, \boldsymbol{y}\epsilon C_i} d(\boldsymbol{x}, \boldsymbol{y})$$

y $\sum(C_j)$ se define de forma similar.

como:

$$D(C_1, C_2 \cup C_3)$$

$$= \frac{1}{n_1(n_2 + n_3)}\sum(C_1, C_2 \cup C_3) \qquad (2.4)$$
$$- \frac{1}{2{n_1}^2}\sum(C_1) - \frac{1}{2(n_2 + n_3)^2}\sum(C_2 \cup C_3)$$

de la ecuación 2.1 tenemos:

$$D(C_1, C_2 \cup C_3)$$

$$= \frac{n_2}{n_2 + n_3} D(C_1, C_2) + \frac{n_3}{n_2 + n_3} D(C_1, C_3) - \frac{n_2 n_3}{(n_2 + n_3)^2} D(C_2, C_3),$$

Sustituyendo la ecuación 2.3 en la ecuación anterior tenemos:

$$D(C_1, C_2 \cup C_3)$$

$$= \frac{n_2}{n_2+n_3}\left(\frac{1}{n_1 n_2}\sum(C_1, C_2) - \frac{1}{2{n_1}^2}\sum(C_1) - \frac{1}{2{n_2}^2}\sum C_2\right)$$
$$+ \frac{n_3}{n_2+n_3}\left(\frac{1}{n_1 n_3}\sum(C_1, C_3) - \frac{1}{2{n_1}^2}\sum(C_1) - \frac{1}{2{n_3}^2}\sum C_3\right)$$
$$- \frac{n_2 n_3}{(n_2+n_3)^2}\left(\frac{1}{n_2 n_3}\sum(C_2, C_3) - \frac{1}{2{n_2}^2}\sum(C_2) - \frac{1}{2{n_3}^2}\sum C_3\right)$$

$$= \frac{1}{n_1(n_2+n_3)}\sum(C_1, C_2 \cup C_3) - \frac{1}{2{n_1}^2}\sum(C_1)$$

$$- \frac{1}{2(n_2+n_3)^2}\left[\sum(C_2) + \sum(C_3) + 2\sum(C_2, C_3)\right]$$

$$= \frac{1}{n_1(n_2+n_3)}\sum(C_1, C_2 \cup C_3) - \frac{1}{2{n_1}^2}\sum(C_1) - \frac{1}{2(n_2+n_3)^2}\sum(C_2 \cup C_3).$$

En la ecuación anterior se utilizaron las siguientes igualdades:

$$\sum(C_1, C_2) + \sum(C_1, C_3) = \sum C_1\,, (C_2 \cup C_3)$$
$$\sum(C_2) + \sum(C_3) + 2\sum(C_2, C_3) = \sum(C_2 \cup C_3)$$

En particular, si se toma $d(\cdot,\cdot)$ en la ecuación 2.2, como la distancia euclidiana al cuadrado entonces la distancia $D(C, C')$ es exactamente la distancia euclidiana al cuadrado entre los centroides de C y C'.

Si $d(\cdot,\cdot)$ en la ecuación 2.2, es la distancia euclidiana al cuadrado, entonces:

$$D(C, C')$$
$$= \frac{1}{|C||C'|}\sum(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T - \frac{1}{2|C|^2}\sum_{\boldsymbol{x},\boldsymbol{y}\varepsilon C}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T$$
$$- \frac{1}{2|C'|^2}\sum_{\boldsymbol{x},\boldsymbol{y}\varepsilon C'}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T$$
$$= \frac{1}{|C|}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C}}\boldsymbol{x}\boldsymbol{x}^T - \frac{2}{|C||C'|}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C},\boldsymbol{y}\varepsilon \boldsymbol{C'}}\boldsymbol{x}\boldsymbol{y}^T + \frac{1}{|C'|}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C'}}\boldsymbol{x}\boldsymbol{x}^T$$
$$- \frac{1}{|C|}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C}}\boldsymbol{x}\boldsymbol{x}^T + \frac{1}{|C|^2}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C},\boldsymbol{y}\varepsilon \boldsymbol{C'}}\boldsymbol{x}\boldsymbol{y}^T + \frac{1}{|C'|}\sum_{\boldsymbol{x}\varepsilon \boldsymbol{C'}}\boldsymbol{y}\boldsymbol{y}^T$$

$$+\frac{1}{|C'|^2}\sum_{x,y\varepsilon C'} \boldsymbol{xy}^T$$

$$=\frac{1}{|C|^2}\sum_{x,y\varepsilon C} \boldsymbol{xy}^T+\frac{1}{|C'|^2}\sum_{x,y\varepsilon C'} \boldsymbol{xy}^T-\frac{21}{|C||C'|}\sum_{x\varepsilon C,y\varepsilon C'} \boldsymbol{xy}^T$$

$$=\left(\frac{1}{|C|}\sum_{x\varepsilon C} \boldsymbol{x}-\frac{1}{|C'|}\sum_{x\varepsilon C'} \boldsymbol{x}\right)\left(\frac{1}{|C|}\sum_{y\varepsilon C} \boldsymbol{y}-\frac{1}{|C'|}\sum_{y\varepsilon C'} \boldsymbol{y}\right)^T$$

ya que $(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T = \boldsymbol{xx}^T - 2\boldsymbol{xy}^T + \boldsymbol{yy}^T$; $\boldsymbol{xy}^T = \boldsymbol{yx}^T$

La ecuación 2.2, proporciona otra forma de calcular la distancia entre los conglomerados nuevos y los viejos.

2.1.3.8. Método de la Mediana

El método de la mediana, también conocido como "el método de los grupos de pares ponderados usando centroides" (Jain y Dubes, 1988) o el "método del centroide ponderado", lo propuso por primera vez Gower (1967), para resolver algunas desventajas del método del centroide. En el método del centroide, si los tamaños de los dos grupos a unirse son muy diferentes, entonces el centroide del nuevo grupo estará muy cerca al grupo más grande y puede permacer dentro de ese grupo (Everitt, 1993). En el método de la mediana, el centroide del nuevo grupo es independiente del tamaño de los grupos que se conglomeraron.

Una desventaja del método de la mediana es que no es adecuado para medidas como el coeficiente de correlación, ya que no es posible la interpretación en un sentido geométrico (Lance y Williams, 1967a).

En el método de la mediana, las nuevas distancias entre los grupos formados y los otros grupos se calculan así:

$$D\left(C_k, C_i \cup C_j\right) = \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D\left(C_k, C_j\right) - \frac{1}{4}D\left(C_i, C_j\right)$$

donde C_k, C_i y C_j son tres clusters en el mismo nivel de agrupamiento.

2.1.3.9. Método de Ward

Ward Jr. (1963) y Ward Jr. y Hook (1963), propusieron un procedimiento de agrupamiento jerárquico buscando formar particiones $P_n, P_{n-1}, \dots, P_1$, de manera que se minimice la pérdida de información asociada con cada unión. Por lo general la pérdida de información se cuantifica en términos de la suma de cuadrados del error, SCE, por lo que el método de Ward se conoce también como el método de la "mínima varianza".

Dado un grupo de puntos C, la SCE asociada con C está dada por:

$$SCE(C) = \sum_{x \epsilon C} \big(\boldsymbol{x} - \mu(C)\big)\big(\boldsymbol{x} - \mu(C)\big)^T$$

o

$$SCE(C) = \sum_{\boldsymbol{x} \epsilon C} \boldsymbol{x}\boldsymbol{x}^T - \frac{1}{|C|}\left(\sum_{\boldsymbol{x} \epsilon C} \boldsymbol{x}\right)\left(\sum_{\boldsymbol{x} \epsilon C} \boldsymbol{x}\right)^T$$

$$= \sum_{\boldsymbol{x} \epsilon C} \boldsymbol{x}\boldsymbol{x}^T - |C|\mu(C)\mu(C)^T \qquad (2.5)$$

donde $\mu(C)$ es la media de C, es decir:

$$\mu(C) = \frac{1}{|C|}\sum_{\boldsymbol{x} \epsilon C} \boldsymbol{x}$$

Suponga que hay k grupos $C_1, C_2, \dots, C_k$ en el mismo nivel de agrupamiento. Entonces la pérdida de información se representa mediante la suma de cuadrados del error dada por:

$$SCE = \sum_{i=1}^{K} SCE(C_i)$$

la cual es la SCE total dentro del grupo.

Para cada etapa del método de Ward, se considera la unión de cada posible par de grupos y los dos grupos que al unirse den como resultado el mínimo incremento en la pérdida de información son los que se fusionan.

Si la distancia euclidiana al cuadrado se usa para calcular la matriz de disimilaridad, entonces esta puede ser recalculada mediante la fórmula de Lance-Williams durante el proceso de conglomerado como sigue (Wishart, 1969):

$$D\big(C_k, C_i \cup C_j\big)$$

$$\frac{|C_k| + |C_i|}{\sum_{i,j,k}} D(C_k, C_i) + \frac{|C_k| + |C_j|}{\sum_{i,j,k}} D\big(C_k, C_j\big) - \frac{|C_k|}{\sum_{i,j,k}} D\big(C_i, C_j\big) \quad (2.6)$$

donde $\sum_{i,j,k} = |C_k| + |C_i| + |C_j|$

Para justificar esto se supone que C_i y C_j se seleccionan para unirse y el cluster que resulta es C_t, es decir $C_t = C_i \cup C_j$, entonces el incremento en la SCE es:

$$\Delta SCE_{ij} = SCE(C_t) - SCE(C_i) - SCE\big(C_j\big)$$

$$= \left(\sum_{\boldsymbol{x}\epsilon C_t} \boldsymbol{x}\boldsymbol{x}^T - |C_t|\mu_t\mu_t^T\right) - \left(\sum_{\boldsymbol{x}\epsilon C_i} \boldsymbol{x}\boldsymbol{x}^T - |C_i|\mu_i\mu_i^T\right) - \left(\sum_{\boldsymbol{x}\epsilon C_j} \boldsymbol{x}\boldsymbol{x}^T - |C_j|\mu_j\mu_j^T\right)$$

$$= |C_i|\mu_i\mu_i^T + |C_j|\mu_j\mu_j^T - |C_t|\mu_t\mu_t^T \quad (2.7)$$

donde μ_t, μ_i, μ_j son las medias de los clusters C_t, C_i y C_j respectivamente.

Como $|C_t|\mu_t = |C_i|\mu_i + |C_j|\mu_j$, elevando al cuadrado ambos lados de la ecuación:

$$|C_t|^2\mu_t{\mu_t}^T = |C_i|^2\mu_i{\mu_i}^T + |C_j|^2\mu_j{\mu_j}^T + 2|C_i||C_j|\mu_i{\mu_j}^T,$$

o

$$|C_t|^2\mu_t{\mu_t}^T = |C_i|^2\mu_i{\mu_i}^T + |C_j|^2\mu_j{\mu_j}^T + |C_i||C_j|\big(\mu_i{\mu_i}^T + \mu_j{\mu_j}^T\big) \quad (2.8)$$
$$-|C_i||C_j|\big(\mu_i - \mu_j\big)\big(\mu_i - \mu_j\big)^T$$

ya que

$$2\mu_i{\mu_j}^T = \mu_i{\mu_i}^T + \mu_j{\mu_j}^T - \big(\mu_i - \mu_j\big)\big(\mu_i - \mu_j\big)^T$$

Dividiendo ambos lados de la ecuación por $|C_t|$ y sustituyendo $|C_t|\mu_t\mu_t^T$ en la ecuación 2.7, se obtiene:

$$\Delta SCE_{ij} = \frac{|C_i||C_j|}{|C_i| + |C_j|}(\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{2.9}$$

Ahora si se considera el incremento en la SCE que resultaria de la unión potencial de los grupos C_k y C_t. De la ecuación 2.9,resulta:

$$\Delta SCE_{kt} = \frac{|C_k||C_t|}{|C_k| + |C_i|}(\mu_k - \mu_t)(\mu_k - \mu_t)^T \tag{2.10}$$

donde $\mu_k = \mu(C_k)$ es la media del grupo C_k.

ya que $\mu_t = \frac{1}{|C_t|}(|C_i|\mu_i + |C_j|\mu_j)$ y $|C_t| = |C_i| + |C_j|$ y de la ecuación 2.8:

$$\begin{aligned}
&(\mu_k - \mu_t)(\mu_k - \mu_t)^T \\
&= \frac{|C_i|}{|C_t|}(\mu_k - \mu_i)(\mu_k - \mu_t)^T + \frac{|C_j|}{|C_t|}(\mu_k - \mu_j)(\mu_k - \mu_j)^T \\
&\quad - \frac{|C_i||C_j|}{|C_t|^2}(\mu_i - \mu_j)(\mu_i - \mu_j)^T
\end{aligned}$$

Sustituyendo la ecuación anterior en la ecuación 2.10:

$$\begin{aligned}
\Delta SCE_{kt} = &\frac{|C_k||C_i|}{|C_k| + |C_t|}(\mu_k - \mu_i)(\mu_k - \mu_i)^T \\
&+ \frac{|C_k||C_j|}{|C_k| + |C_t|}(\mu_k - \mu_j)(\mu_k - \mu_j)^T \\
&- \frac{|C_k||C_i||C_j|}{|C_k| + |C_t|}(\mu_i - \mu_j)(\mu_i - \mu_j)^T
\end{aligned}$$

y utilizando la ecuación 2.9:

$$\begin{aligned}
&\Delta SCE_{kt} \\
&= \frac{|C_k| + |C_i|}{|C_k| + |C_t|}\Delta SCE_{ki} + \frac{|C_k| + |C_j|}{|C_k| + |C_t|}\Delta SCE_{kj} \\
&\quad - \frac{|C_k|}{|C_k| + |C_t|}\Delta SCE_{ij}
\end{aligned} \tag{2.11}$$

Lo anterior prueba la ecuación 2.6. Si se calcula la matriz de disimilaridad para un grupo de datos $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ usando la distancia euclidiana al cuadrado, entonces los valores (i, j) de la matriz de disimilaridad son:

$$d_{ij}^2 = d(, \boldsymbol{x_j}) = (\boldsymbol{x_i} - \boldsymbol{x_j})(\boldsymbol{x_i} - \boldsymbol{x_j})^T = \sum_{l=1}^{d} (x_{il} - x_{jl})^2$$

donde d es la dimensionalidad del conjunto de datos D

Si $C_i = \{\boldsymbol{x_i}\}$ y $C_j = \{\boldsymbol{x_j}\}$ en la ecuación 2.9, entonces el incremento en la SCE que resulta de la unión de $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$es :

$$\Delta SCE_{ij} = \frac{1}{2} d_{ij}^2$$

Ya que el objetivo del método de Ward es encontrar para cada etapa cuáles son los dos grupos que al unirse den el mínimo incremento en la SCE total dentro del grupo, los dos puntos con la mínima distancia cuadrada euclidiana se unirán en la primera etapa. Si se supone que entre $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$ hay la mínima distancia euclidiana al cuadrado, entonces $C_i = \{\boldsymbol{x_i}\}$ y $C_j = \{\boldsymbol{x_j}\}$ deberán unirse. Después de su unión, la distancia entre $C_i \cup C_j$ y los otros puntos deben calcularse. Sea $C_k = \{\boldsymbol{x_k}\}$, cualquier otro grupo entonces la SCE que resulta de la fusión potencial de C_k con $C_i \cup C_j$ se puede calcular a partir de la ecuación 2.11:

$$\Delta SCE_{k(ij)} = \frac{2}{3}\frac{d_{ki}^2}{2} + \frac{2}{3}\frac{d_{kj}^2}{2} - \frac{1}{3}\frac{d_{ij}^2}{2}$$

Si se recalcula la matriz la de disimilaridad usando la ecuación 2.6, entonces de la ecuación anterior:

$$\Delta SCE_{k(ij)} = \frac{1}{2} D\left(C_k, C_i \cup C_j\right)$$

Entonces si se recalcula la matriz de disimilaridad usando la ecuación 2.6, durante el proceso de agrupamiento, se deberán unir los grupos con la misma distancia.

Tabla 4. Métodos Jerárquicos Estándar Aglomerativos

Método	Nombre Alternativo	Por lo general usada con	Distancia entre clusteres definida como:	Observaciones
Vinculación individual Sneath (1957)	Vecino más cercano	Similaridad o distancia	Distancia mínima entre un par de objetos, uno en un cluster, uno en el otro.	Tiende a producir conglomerados desequilibrados y desordenados ("Encadenamiento'", sobre todo en grandes conjuntos de datos. No tiene en cuenta la estructura del cluster.
Vinculación completa Sorensen (1948)	Vecino más lejano	Similaridad o distancia	Distancia máxima entre un par de objetos, uno en un cluster, uno en el otro.	Tiende a encontrar conglomerados compactos con diámetros iguales (máxima distancia entre los objetos). No tiene en cuenta la estructura del cluster.
Vinculación Promedio (grupos). Sokal y Michener (1958)	UPGMA	Similaridad o distancia	Distancia promedio entre un par de objetos, uno en un cluster, uno en el otro.	Tiende a unir grupos con pequeñas varianzas. Intermedio entre la vinculación individual y completa. Toma en cuenta la estructura de grupos. Relativamente robusto.
Vinculación Centroide. Sokal y Michener (1958)	UPGMC	Distancia (requiere datos crudos)	Distancia Euclidiana al cuadrado entre vectores de medias (centroides)	Asume puntos que pueden representarse en el espacio euclidiano (para interpretación geométrica). El más numeroso de los dos grupos domina la fusión. Sujeto a restablecimiento.
Vinculación Promedio intragrupos	Intragrupos	Similaridad o distancia	Distancia promedio entre un par de objetos, uno en un cluster, uno en el otro.	Solución intermedia entre la vinculación simple y la vinculación completa. Relativamente robusto.
Vinculación Mediana Gower (1967)	WPGMC	Distancia (requiere datos crudos)	Distancia Euclidiana al cuadrado entre centroides ponderados	Asume puntos, pueden representarse en el espacio euclidiano para interpretación geométrica. El nuevo grupo es intermedio en posición entre los conglomerados unidos. Sujeto a restablecimiento.
Método de Ward Ward (1963)	Mínima suma de cuadrados	Distancia (requiere datos crudos)	Incremento en la suma de cuadrados dentro de grupos, después de la unión, sumadas sobre todas las variables	Asume puntos pueden representarse en el espacio euclidiano para interpretación geométrica. Tiende a encontrar cluster del mismo tamaño, agrupaciones esféricas. Sensible a valores atípicos.

Los métodos presentados se resumen en la Tabla 4, junto con algunas observaciones sobre algunas de sus características propias.

2.1.4. La matriz cofenética. Coeficiente de correlación cofenético.

Los métodos jerárquicos imponen una estructura sobre los datos y es necesario con frecuencia considerar si es aceptable o si se introducen distorsiones inaceptables en las relaciones originales. El método más usado para verificar este hecho, o sea, para observar si la relación entre el dendrograma y la matriz de proximidad es original, es el coeficiente de correlación cofénetica, el cual es simplemente la correlación entre los $n(n-1)/2$, elementos de la parte superior de la matriz de proximidades observada y los correspondientes en la llamada matriz cofenética, C, cuyos elementos, c_{ij} , se definen como aquellos que determinan la proximidad entre los elementos i y j cuando éstos se unen en el mismo cluster. Después de aplicar varios procedimientos de agrupamiento distintos, surge la pregunta acerca de cuál método se debe elegir como definitivo. La respuesta la da el coeficiente cofenético, ya que aquel método que tenga un coeficiente cofenético mayor será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio.

2.2. Modelo de regresión geográficamente ponderada (GWR)

Un modelo de regresión global viene dado por:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \tag{2.12}$$

En la calibración de este modelo, se estima un parámetro para la relación entre cada variable independiente y la variable dependiente y se asume que estas relaciones son constantes a través de la región de estudio.

La regresión geográficamente ponderada, GWR, es una técnica que extiende el marco de regresión tradicional permitiendo parámetros locales en lugar de globales, los cuales son estimados de manera que el modelo se reescribe como:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2.13}$$

donde (u_i, v_i) denota las coordenadas del punto i-ésimo en el espacio y $\beta_k(u_i, v_i)$ es una realización de la función continua $\beta_k(u, v)$ en el punto i. Es decir, se permite una superficie continua de valores de parámetros y las mediciones de esta superficie se toman en ciertos puntos para indicar la variabilidad espacial de la superficie. La ecuación (2.12) es un caso especial de la ecuación (2.13) en la que los parámetros se suponen espacialmente invariantes. La ecuación GWR en (2.13) reconoce que podrían existir variaciones espaciales en las relaciones y proporciona una manera en la que se pueden medir.

En el enfoque del método se asume que los coeficientes no son al azar, sino más bien que son funciones deterministas de algunas otras variables, en este caso, la ubicación en el espacio. Se debe tener en cuenta que aunque no es posible una estimación insesgada de los coeficientes locales, si es posible realizar estimaciones con sólo una pequeña cantidad de sesgo.

El proceso de calibración se puede conceptualizar como una solución de compromiso entre el sesgo y el error estándar. Suponiendo que los parámetros muestren algún grado de coherencia espacial, entonces los valores cercanos a lo que se está estimando deben tener magnitudes y signos relativamente similares. Por lo tanto, al estimar un parámetro en una ubicación dada i, se puede aproximar (2.13) en la región de i mediante (2.12), y realizar una regresión usando un subconjunto de los puntos del conjunto de datos que estén cerca de i. Por lo tanto, las $\beta_k(u_i, v_i)$ se estiman para i de la forma habitual y para la siguiente i, se usa un nuevo subconjunto de los puntos "cercanos"', y así sucesivamente. Estas estimaciones tendrán algún grado de sesgo, ya que los coeficientes de (2.13) exhibirán algún desvío a través del subconjunto de calibración local. Sin embargo, si la muestra local es lo suficientemente grande, se podrá realizar la calibración, aunque sea sesgada. Cuanto mayor sea el tamaño del subconjunto de calibración local, menor serán los errores estándar de las estimaciones de los coeficientes; pero al aumentar este subconjunto, también aumenta la probabilidad de que el coeficiente se desvíe y por tanto se introduzca sesgo; para reducir este efecto, se puede hacer un ajuste final. Suponiendo que los puntos más lejanos de i en el subconjunto de calibración son los que mayor probabilidad tienden a tener diferentes coeficientes, se utiliza una calibración ponderada, de modo que la mayor influencia en la calibración se atribuye a los puntos más cerca de i.

La calibración de la ecuación (2.13) presupone implícitamente que los datos observados cerca de la localización i tienen más influencia en la estimación de las $\beta_k(u_i, v_i)$ que la que tienen

los datos situados más lejos de i; en esencia, la ecuación mide las relaciones inherentes en el modelo alrededor de cada ubicación i. Por lo tanto los mínimos cuadrados ponderados proporcionan una base para la comprensión de cómo opera GWR.

En GWR una observación se pondera de acuerdo con su proximidad a la posición i de modo que la ponderación de una observación ya no es constante en la calibración porque varía con i. Los datos de observaciones cerca de i pesan más que los datos de las observaciones más lejanas, esto es:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{y} \qquad (2.14)$$

donde las negrillas indican una matriz, $\widehat{\boldsymbol{\beta}}$ representa una estimación de $\boldsymbol{\beta}$, y $\boldsymbol{W}(u_i, v_i)$ es una matriz de $n \times n$, cuyos elementos fuera de la diagonal son cero y los elementos en la diagonal denotan la ponderación geográfica de cada uno de los n datos observados para la regresión en el punto i.

Para ver esto más claramente, consideremos la ecuación de regresión clásica en la forma de matriz:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

donde el vector de parámetros a estimar, es constante en el espacio y es estimado por:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

El equivalente en GWR es:

$$\boldsymbol{Y} = (\boldsymbol{\beta} \otimes \boldsymbol{X})\mathbf{1} + \varepsilon$$

donde $\otimes$ es un operador de multiplicación lógico en el que se multiplica cada elemento de $\boldsymbol{\beta}$ por el correspondiente elemento de $\boldsymbol{X}$.

Si hay n puntos de datos y k variables explicativas tanto $\boldsymbol{\beta}$ como $\boldsymbol{X}$ tendrán las dimensiones $n \times (k+1)$ y $\mathbf{1}$ es un $(k+1) \times 1$ vector de unos. La matriz $\boldsymbol{\beta}$ ahora consta de n conjuntos de parámetros locales y tiene la estructura siguiente:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1,v_1) & \beta_1(u_1,v_1) & .... & \beta_k(u_1,v_1) \\ \beta_0(u_2,v_2) & \beta_1(u_2,v_2) & .... & \beta_k(u_2,v_2) \\ .... & .... & .... & .... \\ \beta_0(u_n,v_n) & \beta_1(u_n,v_n) & ..... & \beta_k(u_n,v_n) \end{bmatrix}$$

Los parámetros en cada fila de la matriz anterior se estiman por

$$\widehat{\boldsymbol{\beta}}(i) = (\boldsymbol{X^T W}(i)\boldsymbol{X})^{-1}\boldsymbol{X^T W}(i)\boldsymbol{Y} \tag{2.15}$$

donde i representa una fila de la matriz $\boldsymbol{\beta}$ y $\boldsymbol{W}(i)$ es una matriz espacial ponderada $n \times n$ de la forma:

$$\boldsymbol{W}(\boldsymbol{i}) = \begin{bmatrix} w_{i1} & 0 & .... & 0 \\ 0 & w_{i2} & .... & 0 \\ . & . & & . \\ 0 & 0 & .... & w_{in} \end{bmatrix}_{?} \tag{2.16}$$

donde w_{in} es el peso dado para un punto de n datos en la calibración del modelo para la ubicación i.

El estimador en la ecuación (2.15) es un estimador de mínimos cuadrados ponderados pero tiene una matriz de pesos constante; los pesos en GWR varían de acuerdo con la ubicación del punto i. Por lo tanto la matriz de ponderación tiene que ser calculada para cada punto i y los pesos representan la proximidad de cada punto de datos a la ubicación i, en donde los puntos de mayor proximidad llevan más peso en la estimación de los parámetros para la localización i.

Sin embargo, en las ecuaciones (2.15) y (2.16) no hay razón por la que i tiene que ser la ubicación de un punto de datos. Las estimaciones locales de los parámetros, pueden ser derivados para cualquier punto en el espacio, independientemente de si el punto es uno en el que se han observado datos.

Además de la estimación de los parámetros locales, es útil también calcular los errores estándar locales para tener en cuenta las variaciones en los datos utilizados en el cálculo de las estimaciones. En algunos casos, por ejemplo, las estimaciones de los parámetros locales

pueden ser una función del número relativamente reducido de puntos de datos o los puntos de datos podrían tener bajos pesos en la regresión local, ya que se encuentran muy lejos del punto de regresión.

Los errores estándar de las estimaciones de los parámetros GWR se derivan de la siguiente manera: Se expresa el estimador de las estimaciones de los parámetros locales dada en la ecuación (2.14) como:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = \boldsymbol{C}\boldsymbol{y}$$

donde $\boldsymbol{C}$ es la suma de los cuadrados residuales normalizados de la regresión local y se define como:

$$\boldsymbol{C} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)$$

La varianza del estimador está dada por:

$$Var\left[\widehat{\boldsymbol{\beta}}(u_i, v_i) = \boldsymbol{C}\boldsymbol{C}^T\sigma^2\right] \qquad (2.17)$$

$$\sigma^2 = \sum_i \frac{(y_i - \hat{y}_i)}{n - 2v_1 + v_2}$$

$$v_1 = tr(\boldsymbol{S})$$

$$v_1 = tr(\boldsymbol{S}^T\boldsymbol{S})$$

La matriz $\boldsymbol{S}$ se conoce como la matriz sombrero (Hoaglin y Welsch, 1978) que mapea a $\widehat{y}$ sobre y de la siguiente manera:

$$\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$$

donde cada fila $\boldsymbol{r_i}$ de $\boldsymbol{S}$, , está dada por:

$$\boldsymbol{r_i} = \boldsymbol{X_i}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)$$

El término $n - 2v_1 + v_2$ se conoce como los grados efectivos de libertad del residual. El término $2v_1 - v_2$ es equivalente al número de parámetros en un modelo de regresión lineal global y se puede llamar el número efectivo de parámetros en el Modelo GWR local. Debido a que la traza de S y la traza de S^TS son generalmente muy similares, el número efectivo de parámetros en la regresión local por lo general se puede aproximar mediante v_1 que ahorra tener que calcular la traza de S^TS.

Una vez que la varianza de cada estimación del parámetro se obtiene de la ecuación (2.17), los errores estándar se obtienen a partir de:

$$ES(\widehat{\boldsymbol{\beta}}_i) = \sqrt{Var(\widehat{\boldsymbol{\beta}}_i)}$$

donde $\boldsymbol{\beta}_i$es una notación abreviada para $\boldsymbol{\beta}(u_i, v_i)$.

Hasta este punto, simplemente se ha indicado en GWR que $\boldsymbol{W}(u_i, v_i)$ o en términos más convenientes, $W(i)$, es un esquema de ponderación basado en la proximidad de la regresión del punto i a los puntos de datos alrededor de i sin una relación explícita declarada. Ahora consideremos la elección de dicha relación. Primero supongamos un esquema implícito de ponderación del marco OLS en la ecuación (2.12):

$$w_{ij} = 1 \quad \forall\, i, j$$

donde j representa un punto específico en el espacio en el que se observan los datos y i representa cualquier punto en el espacio para el cual se estiman los parámetros; es decir, en el modelo global cada observación tiene un peso igual a la unidad.

Un primer paso hacia la ponderación basada en la localidad podría ser la de excluir desde el modelo de calibración las observaciones que estén más alejadas de una cierta distancia d desde el punto de regresión.

Esto sería equivalente a establecer sus pesos iguales a cero, dando una función de ponderación de:

$$w_{ij} = 1 \text{ si } d_{ij} < d \qquad (2.18)$$
$$w_{ij} = 0 \text{ de otra forma}$$

El cual es el enfoque de ventana móvil; el uso de este esquema de ponderación simplificaría el procedimiento de calibración porque en cada punto de regresión sólo un subconjunto de los datos se utilizaría para calibrar el modelo. Sin embargo, esta función de ponderación espacial presenta discontinuidad. Como los cambios de los puntos de regresión, los coeficientes estimados podrían cambiar drásticamente como un punto de datos que se mueve dentro o fuera de la ventana alrededor de i. Aunque los cambios bruscos de los parámetros en el espacio pueden genuinamente producirse, en este caso los cambios en las estimaciones se producirían como artefactos del arreglo de los puntos de datos, en lugar de representar cualquier proceso subyacente en la relación bajo investigación. Ejemplos del uso de este tipo de GWR están dadas por Fotheringham et al. (1996) y por Charlton et al. (1997).

Una forma de combatir el problema de las discontinuidad de los pesos es especificar w_{ij} como una función continua de d_{ij}, que es la distancia entre i y j. Una selección obvia es:

$$w_{ij} = exp\,[-\tfrac{1}{2}(d_{ij}/b)^2] \tag{2.19}$$

Donde b se refiere al ancho de banda. Si i y j coinciden (es decir, i también es el punto en el espacio en el que se observan los datos), la ponderación de los datos en ese punto será la unidad y la ponderación de los otros datos se reducirá de acuerdo con una curva de Gauss conforme la distancia entre i y j se incrementa. En el último caso, la inclusión de datos en el procedimiento de calibración se convierte en "fraccional". Por ejemplo, en la calibración de un modelo para el punto i, si $w_{ij} = 0.5$, los datos en el punto j contribuyen sólo en la mitad del peso en el procedimiento de calibración. Para los datos lejanos de i la ponderación tenderá a ser prácticamente cero, excluyendo efectivamente éstas observaciones en la estimación de los parámetros para la localización i. Los resultados empíricos del GWR descritos anteriormente se basan en funciones de ponderación o kernel de Gauss o aproximadamente gaussianas.

Se pueden utilizar funciones de ponderación alternativas; algunos ejemplos de la aplicación de estas formas alternativas se pueden revisar en Brunsdon et al. (1996; 1997) y Fotheringham et al (1998).

Cualquiera que sea la función de ponderación específica empleada, la idea esencial de la GWR es que por cada punto de regresión i hay una influencia alrededor de i descrita por la función

de ponderación, de tal manera que las observaciones muestreadas cerca de i tenga más influencia en la estimación de los parámetros que las observaciones más lejanas.

Un aspecto de la GWR es que los parámetros estimados dependen en parte de la función de ponderación o kernel seleccionado. En la ecuación (2.18), por ejemplo, conforme d se hace más grande, más cerca estará la solución del modelo a la respuesta de la regresión OLS y cuando d es igual a la distancia máxima entre los puntos en el sistema, la respuesta de los dos modelos será igual. De manera equivalente, en la ecuación (2.19) cuando b tiende a infinito, los pesos tienden a uno para todos los pares de puntos, de tal forma que los parámetros estimados se hacen uniformes y la GWR se vuelve equivalente a la OLS. Por el contrario, conforme el ancho de banda se hace más pequeño, las estimaciones de los parámetros dependerán cada vez más de las observaciones en las proximidades de i y por lo tanto se incrementa la varianza. Entonces, el problema es cómo seleccionar un ancho de banda adecuado o función de decaimiento en la GWR. Hay una serie de criterios que se pueden utilizar para la selección del ancho de banda.

Considere la selección de b en la ecuación (2.19). Una posibilidad es elegir b con el criterio de los mínimos cuadrados, es decir minimizar:

$$z = \sum_{i=1}^{n} [y_i - \hat{y}_i(b)]^2 \tag{2.20}$$

donde (b) es el valor ajustado de y_i utilizando un ancho de banda de b. Con el fin de encontrar el valor de ajuste de y_i es necesario estimar las $\beta_k(u_i, v_i)$ en cada uno de los puntos de datos y luego combinar estos con los x-valores en estos puntos. Sin embargo, hay un problema con un procedimiento de este tipo: Supongamos que b se hace muy pequeña, así que la ponderación de todos los puntos excepto i misma es insignificante, entonces los valores ajustados en los puntos muestreados tenderán a los valores reales de modo que el valor de la ecuación (2.20) se convierte en cero. Esto sugiere que, bajo este criterio de optimización el valor de b tiende a cero, lo que no es útil; en primer lugar, los parámetros de un modelo de este tipo no estarán definidos en este caso límite y, en segundo lugar, las estimaciones fluctúan ampliamente en todo el espacio con el fin de dar valores locales de ajuste precisos en cada punto de la regresión.

Una solución a este problema es un enfoque de validación cruzada (VC) sugerido para la regresión local por Cleveland (1979) y para la estimación de densidad de kernel por Bowman (1984).

$$VC = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2$$

Se utiliza cuando $\hat{y}_{\neq i}(b)$ es el valor de ajuste de y_i con las observaciones para el punto i omitido en el proceso de calibración. Este enfoque tiene la propiedad deseable de contrarrestar el efecto "envolvente", ya que cuando b se vuelve muy pequeño, el modelo se calibra sólo en muestras cerca de i y no con i.

Una aproximación a la estadística de validación cruzada que es más fácil de calcular se conoce como el criterio de validación cruzada generalizada (GCV) que describe Loader (1999) y que se utilizó por primera vez en el contexto de suavizado de *splines* por Craven y Wahba (1979). La fórmula para la puntuación GCV es:

$$GVC = n \sum_{i=1}^{n} [y_i - \hat{y}_i(b)]^2 / (n - \upsilon_1)^2 \tag{2.21}$$

donde υ_1 es el número efectivo de parámetros en el modelo tal como se define:

$$\upsilon_1 = tr(\boldsymbol{S})$$

Este término evita la calibración envolvente alrededor de los puntos de datos porque υ_1 tiende a n y el denominador de la ecuación (2.21) tendería a cero.

Un método similar para derivar el ancho de banda que proporciona una compensación entre la bondad de ajuste y los grados de libertad es minimizar el criterio de información de Akaike:

$$CIA_c = 2nlog_e(\hat{\sigma}) + nlog_e(2\pi) + n\left\{\frac{n + tr(\boldsymbol{S})}{n - 2 - tr(\boldsymbol{S})}\right\}$$

donde n es el tamaño de la muestra, $\hat{\sigma}$ es la desviación estándar estimada del término del error, y tr ($\boldsymbol{S}$) denota la traza de la matriz sombrero, que es una función del ancho de banda. El CIA tiene la ventaja de ser más general en la aplicación que el estadístico VC porque puede ser utilizado en Poisson y GWR logística, así como en modelos lineales.

Otro criterio de selección de ancho de banda que se ha utilizado en la literatura de la GWR es el criterio de información bayesiano (CIB), Nakaya (2002), a veces conocido como el Criterio de Información de Schwartz (CIS) (Schwartz 1978), el cual se define como:

$$CIB = -2log_e(\mathrm{L}) + klog_e(n)$$

donde L es el modelo de probabilidad, k es el número de parámetros y n es el tamaño de la muestra. Esto es similar a la CIA aunque la penalidad de la "complejidad del modelo" difiere. Aquí, el mismo grado de complejidad (es decir, el mismo valor de k) conlleva una mayor penalidad para muestras más grandes. En consecuencia, en muestras más grandes el uso de la CIB tiende a identificar los modelos con menos parámetros como óptima. Como su nombre lo indica, la CIB se derivó en un contexto bayesiano. Surge de un modelo bayesiano en el que cada uno de un número discreto de modelos candidatos tienen iguales probabilidades *a priori*, pero las distribuciones *a priori* sobre los parámetros del modelo, dado el modelo no son informativas.

Es útil en esta etapa revisar la relación entre el sesgo y la varianza tanto en general y dentro del contexto del GWR. Por lo que, hay que discutir algunas propiedades del estimador de $\boldsymbol{y}$, $\hat{y}$. En cualquier punto en el espacio (u, v), si se da un conjunto de predictores, $\boldsymbol{X}$, y un conjunto de los estimadores de los coeficientes, $\widehat{\boldsymbol{\beta}}$, entonces, $\hat{y} = \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ es una estimación de $\boldsymbol{y}$ en ese punto. Sin embargo, $\widehat{\boldsymbol{\beta}}$ es una estimación de $\boldsymbol{\beta}$ basada en una muestra del espacio **X** y y observaciones. Debido a la aleatoriedad de los términos y, $\widehat{\boldsymbol{\beta}}$ es aleatorio, y por lo tanto también lo es $\hat{y}$. Dos propiedades importantes de la distribución de $\hat{y}$ son su valor esperado $E(\hat{y})$, y su desviación estándar, $DS(\hat{y})$. Cuando, para todo $\boldsymbol{X}$, $E(\hat{y}) = E(y)$, el estimador se dice que es insesgado. En este caso, $DS(\hat{y})$ es una medida útil de la calidad de $\hat{y}$ como estimador de y.

Sin embargo, el sesgo cero no garantiza un estimador óptimo. Un estimador sesgado, es decir cuando el valor esperado de y es diferente que el valor verdadero, puede tener una variabilidad mucho menor que la del estimador insesgado. Por lo tanto, los valores extremos de los posibles errores en la predicción de y son menores y la única ventaja del estimador insesgado es ofrecer que la distribución del error se centra en cero. En este caso, la

distribución del error de predicción (ESP), para el estimador insesgado tendría una cola más larga. En general, por lo tanto, a pesar de su sesgo, se podría preferir al estimador sesgado.

Este es un ejemplo de compromiso entre sesgo y varianza que se presenta en muchos tipos de modelado estadístico. Si los coeficientes de regresión varían de forma continua en el espacio, entonces al utilizar la regresión ponderada por mínimos cuadrados es poco probable que se obtengan estimaciones totalmente insesgadas de $\beta(u, v)$en el punto dado (u, v), porque para cada observación habrá un valor diferente de valor de la regresión, pero la regresión requiere que este valor sea el mismo para todas las observaciones. Lo mejor que se puede esperar es que los valores no varíen demasiado y este óptimo se alcanza si solo se toman en cuenta las observaciones cercanas al punto (u, v) en el cual se desea estimar $\beta(u, v)$. Sin embargo, ya que esto reduce el tamaño de la muestra efectiva para la estimación, el error estándar de $\beta(u, v)$ aumentará. Por lo tanto, la cuestión que surge es qué tan cerca de (u, v) deben estar los puntos para ser considerados: demasiado cerca, con varianza grande y sesgo pequeño o demasiado lejos, con una varianza pequeña y sesgo grande. En un extremo, si se elige un modelo global de modo que $\beta(u, v)$ se supone constante para todo (u, v), y si hay variabilidad en la verdadera $\beta(u, v)$, entonces el sesgo causará problemas. En el otro extremo, si las estimaciones de los parámetros locales se derivan de muestras muy pequeñas de datos, tendrán varianzas grandes y serán cada vez menos confiables.

Este compromiso entre sesgo y varianza proporciona alguna justificación para el uso de la validación cruzada como un medio de la elección de ancho de banda. Una puntuación de validación cruzada es esencialmente la suma de la estimación de la predicción de los errores al cuadrado (PSE_s). Los PSE_s pueden ser considerados como una medida del rendimiento global de una particular combinación sesgo/varianza. No se pueden conocer los PSE_s exactos (si lo hiciéramos, se conocería la $E(y)$ y los valores de β por lo que no se necesitaría ninguna predicción estadística), pero las puntuaciones VC proporcionan una estimación que luego se puede utilizar como una base para la selección.

Un punto final sobre la selección de ancho de banda es que para un determinado conjunto de variables y por lo tanto para un modelo dado, el ancho de banda óptimo cambiará si la estrategia de muestreo cambia. Por lo tanto la elección del ancho de banda no es un parámetro relativo al modelo en sí mismo, pero es parte esencial de la estrategia de calibración para una muestra dada. Por ejemplo, si se han añadido nuevos puntos de datos para el modelo, uno esperaría lograr mejores estimaciones de $\beta(u, v)$y cabría esperar que el

ancho de banda óptimo disminuya. En última instancia, si el tamaño de la muestra se incrementa continuamente, $\hat{\beta}(u,v)$ debe tender a $\beta(u,v)$,pero el ancho de banda debe tender a cero. Sin embargo, esto no significa que alterando el ancho de banda, y observando los cambios en $\hat{\beta}(u,v)$, no se puede obtener una idea de las diferentes escalas de variación de $\beta(u,v)$.

2.3. Modelo de radiación

Simini et al (2012), proponen un "modelo de radiación" para la movilidad y la migración. Este modelo se basa en dos supuestos simples y plausibles: los seres humanos no disfrutan el movimiento, y por tanto eligen la oportunidad más cercana que mejora sus circunstancias. En otras palabras, los individuos se mueven a una nueva ubicación sólo porque es el lugar más cercano que ofrece, por ejemplo, un mejor trabajo.

Por lo tanto, Simini et al., presuponen que la proyección geográfica de una persona, con centro en su ubicación actual, aumenta hasta que se identifique un lugar mejor, pero no más allá. El supuesto fundamental en el modelo de la radiación es que los individuos no buscan necesariamente la mejor oportunidad, sino más bien su prioridad es el destino más cercano.

Por supuesto, que el número de oportunidades que se ofrecen en un lugar es proporcional al tamaño de la población de ese lugar, y que cada oportunidad tiene una puntuación de calidad al azar, Simini et al, fueron capaces de calcular el flujo esperado de individuos entre dos lugares, por lo que la distribución espacial de la población es la única entrada para su teoría.

Compararon las predicciones de la teoría con conjuntos de datos múltiples que van desde los trayectos de viaje-diarios a los patrones de migración a largo plazo, la movilidad de telefonía móvil y los patrones de comunicación. Encontraron que esta teoría sustancialmente supera el modelo de gravedad.

En analogía con la ley de gravedad de Newton, la ley de gravedad asume que el número de individuos T_{ij} que se mueven entre las ubicaciones i y j por unidad de tiempo es proporcional a alguna potencia de la población de la fuente (m_i) y el destino (m_j) y decae con la distancia r_{ij} entre ellos:

$$T_{ij} = \frac{m_i^{\alpha} \times m_j^{\beta}}{f(r_{ij})} \tag{2.22}$$

donde α y β son exponentes ajustables y la función $f(r_{ij})$ se elige para ajustarse a los datos empíricos. Ocasionalmente T_{ij} se interpreta como la tasa de probabilidad de las personas que viajan desde i hasta j, o un acoplamiento efectivo entre los dos lugares. A pesar de su uso generalizado, la ley gravedad carece de una derivación rigurosa, carece de orientación teórica. Los profesionales utilizan una gama de funciones $f(r_{ij})$, ley de potencia o exponenciales y hasta nueve parámetros para ajustar a los datos empíricos. Además como el modelo requiere datos históricos de la circulación para ajustar los parámetros (α, β, ...), es incapaz de predecir la movilidad en las regiones donde están ausentes los datos.

La ley de gravedad tiene discrepancias predictivas sistemáticas. Para dos localizaciones con poblaciones en el origen y en el destino el flujo entre ellas debe ser el mismo, en cualquier sentido, sin embargo, se puede dar el caso que conmuten más personas en una dirección que en otra. El modelo de gravedad predice que el número de viajeros aumenta sin límite a medida que aumentamos la población de destino n_j, sin embargo, el número de pasajeros no puede exceder el tamaño de la población de origen m_i, situación que resalta la inconsistencia analítica del modelo. La ley de gravedad no contempla la variabilidad en el número de viajeros entre dos lugares ya que es un modelo determinista.

Simini et al, parten del hecho que mientras los desplazamientos es un proceso diario, su origen y destino está determinado por la selección de empleo, una decisión tomada durante una mayor escala de tiempo. Usando la partición natural de un país en condados, para los cuales se recolectan los datos de trayecto, suponen que la selección de empleos consta de dos pasos:

1. Un individuo busca ofertas de empleo de todos los condados, incluyendo su condado natal. El número de oportunidades de empleo en cada condado es proporcional a la población residente, n, en el supuesto que existe una oferta de trabajo para cada n trabajos particulares. Se capturan los beneficios de un potencial de oportunidades de empleo con un solo número, z, elegido al azar de la distribución $p(z)$, donde z representa una combinación de los ingresos, las horas de trabajo, las condiciones, etc. Por lo tanto, a cada condado con población n se le asigna $n/n_{\,trabajos}$ números

aleatorios, z_1, $z_2, \ldots, z_{[n/n_{trabajos}]}$, lo que representa el hecho de que cuanto mayor sea la población de un condado, mayores oportunidades de empleo ofrece.

2. El individuo elige el trabajo más cercano a su casa, cuyos beneficios z son superiores a la mejor oferta disponible en su condado. Por lo tanto, no desplazarse tiene prioridad sobre los beneficios, es decir, los individuos están dispuestos a aceptar trabajos inferiores pero que estén más cerca de su casa.

Este proceso, aplicado en proporción a la población residente en cada condado, asigna los lugares de trabajo a cada viajero potencial, lo que a su vez determina los flujos de desplazamientos diarios de todo el país. El modelo tiene tres parámetros desconocidos: la distribución de beneficios $p\,(z)$, la densidad de trabajo, $n_{trabajos}$ y el número total de viajeros, N_c.

Los flujos de viajeros T_{ij} son independientes de $p(z)$ y $n_{trabajos}$, y el parámetro restante, N_c, no afecta la distribución del flujo, lo que hace al modelo libre de parámetros. Como el modelo puede ser formulado en términos de procesos de radiación y absorción, sus autores lo llamaron modelo de radiación.

Para predecir analíticamente los flujos de trayectos, se consideran las ubicaciones i y j con población m_i y m_j respectivamente, a r_{ij} de distancia uno del otro, y se denota con s_{ij}, la población total en el círculo de radio r_{ij} con centro en i, excluyendo la población de origen y destino. El flujo medio, T_{ij}, de i a j, según lo predicho por el modelo de radiación es:

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{\left(m_i + s_{ij}\right)\left(m_i + n_j + s_{ij}\right)} \tag{2.23}$$

$T_i \equiv \sum_{j \neq i} T_{ij}$, es el número total de viajeros que comienzan su viaje desde la ubicación i, que es proporcional a la población de la ubicación de origen, de ahí $T_i = m_i(N_c/N)$ donde N_c es el número total de viajeros y N es la población total en el país.

El modelo es independiente tanto de $p\,(z)$ como de $n_{trabajos}$,, y es una alternativa propuesta a la ley de la gravedad. El modelo de radiación tiene una derivación rigurosa y no tiene parámetros libres. Una diferencia clave entre el modelo de radiación y la ley de la gravedad es

que la variable del modelo de radiación no es la distancia r_{ij}, es s_{ij}. Así, el flujo de trayectos depende no sólo de m_i y m_j sino también de la población s_{ij}, de la región que rodea a la ubicación de origen. Para una densidad de población s_{ij} uniforme $s_{ij} \approx m_i r_{ij}^2$ y $n = m$ el modelo se reduce a la ley gravedad con $f(r) = r^{\gamma}, \gamma = 4$ y $\propto +\beta = 1$.

La densidad de población no uniforme, sin embargo, es la clave para resolver la limitación cuatro: el modelo de radiación predice un orden de magnitud de diferencia en Alabama y Utah, en línea con los datos del censo. De hecho, la densidad de población alrededor de Utah es significativamente inferior a la media de Estados Unidos, por lo tanto las oportunidades de trabajo dentro el mismo radio son diez veces más pequeño en Utah que en Alabama, lo que implica que los viajeros en Utah tienen que viajar más lejos para encontrar oportunidades comparables de empleo.

El modelo de radiación predice el número de viajeros que sale de un lugar con población m a uno con $n \to \infty$ satura a $T_{n \to \infty} = \frac{m^2}{(m+s)} + O\left(\frac{1}{n}\right) \leq m$, resolviendo la solución de la divergencia no física se resuelve la inconsistencia del modelo de gravedad (Simini, et al, 2012).

T_{ij} en el modelo de radiación es una variable estocástica, no sólo predice el flujo promedio entre dos ubicaciones, sino también su varianza.

Los autores para explorar la capacidad del modelo de radiación para predecir los patrones correctos de los desplazamientos estudiaron los flujos de trayectos con más de diez viajeros procedentes del Condado de Nueva York. Los destinos previstos por la ley de gravedad se encontraron a 400 km del origen, perdiendo todos los viajes de media y de larga distancia. El desempeño local de la ley de gravedad fue igualmente pobre: en el Estado de Nueva York sobreestimó burdamente los flujos en las proximidades de la Ciudad Nueva York y subestimó los flujos en el resto del estado. El modelo de radiación estimó en una forma más realista las pautas de movilidad observadas, tanto a nivel nacional como a nivel estatal.

Para mostrar la generalidad del modelo Simini et al pusieron a prueba su rendimiento para cuatro fenómenos socio-económicos: los patrones de viaje por hora, las migraciones, los patrones de comunicación y los flujos de los productos básicos. Se encuentra que el modelo de radiación ofrece una descripción cuantitativa precisa de la movilidad y el transporte que abarca una amplia gama de escalas de tiempo (movilidad por hora, desplazamientos diarios, s

migraciones anuales), la captura de diversos procesos (desplazamientos, la movilidad intra-día, los patrones de llamadas, comercio), recogidos a través de un amplia gama de herramientas (censo, teléfonos móviles, documentos de impuestos) en diferentes continentes (América, Europa). El ajuste con los datos de tan diversa naturaleza es algo sorprendente, lo que sugiere que las hipótesis detrás del modelo capturan los mecanismos fundamentales de decisión que, directa o indirectamente, son relevantes para un amplio abanico de movilidad y procesos de transporte.

2.4.Identificación de los distritos industriales (DI)

En este trabajo se ha elegido la metodología utilizada por el ISTAT (1996), debido a su sencillez y facilidad de aplicación y al tipo de unidad territorial que utiliza (SLT). Una vez obtenidos los SLT, el proceso para la identificación de distritos industriales consta cuatro fases, basadas en el cálculo de coeficientes de concentración anidados.

Una vez que los SLT han sido delimitados, se procede al estudio de su estructura socio-económica con el fin de identificar aquellos que puedan ser considerados como "distritos industriales". Al final de esta fase se obtiene el mapa de los distritos industriales.

La metodología utilizada por el ISTAT (1996), es sencilla y de fácil aplicación, al tipo de unidad territorial que utiliza (SLT). Una vez obtenidos los SLT, el proceso para la identificación de distritos industriales consta cuatro fases, basadas en el cálculo de coeficientes de concentración anidados. Se consideran sistemas especializados en manufactura a aquellos que muestran un coeficiente de localización de la manufactura superior a 1.

$$LQ_j = \left(\frac{W_{ma}}{W_a} \Big/ \frac{W_m}{W} \right) > 1$$

W = Puestos de Trabajo Localizados;
m = industria manufacturera;
a = sistema local de trabajo.

Partiendo de los SLT especializados en manufactura, se calcula la especialización territorial de ocupación manufacturera en Pymes como porcentaje de ocupados en empresas de menos de 250 ocupados (Pymes) en la industria manufacturera. Si el coeficiente es mayor a uno, este ratio es superior a la media nacional, y por tanto el SLT está especializado en Pymes:

$$LQ_j = \left(\frac{W_{250,ma}}{W_{ma}} \Big/ \frac{W_{250,m}}{W_m} \right) > 1$$

Partiendo de los sistemas locales de Pymes, se calcula el porcentaje de ocupados de cada agregado manufacturero en relación con el total de las manufacturas del sistema local, y se divide por el mismo ratio calculado para el agregado de sistemas de Pymes. La industria que maximiza este coeficiente se considera industria principal (industria distrito):

$$LQ_P = \left(\frac{W_{sa}}{W_{ma}} \Big/ \frac{W_s}{W_m} \right) > 1$$

p = industria-distrito;
s = sector.

Una vez identificados los sistemas manufactureros especializados en Pymes, e identificada la industria principal de los mismos, se procede a comprobar si en estos sistemas locales las Pymes ocupan más trabajadores que las grandes empresas. Si esta condición se cumple, se considera que el SLT tiene características de distrito industrial, y el sector con mayor coeficiente de especialización es la "industria-distrito":

$$I_P = máx\left(\frac{W_{250,pa}}{W_{pa}} \right) > 0.5$$

3. Áreas de viaje al trabajo en España: Análisis Estadístico Multivariable Empírico.

3.1 Introducción

Regionalizar, es organizar y subdividir de manera óptima el espacio geográfico lo que aporta una gran variedad de beneficios, entre los cuales se pueden mencionar los siguientes: se optimizan los recursos para cumplir con las necesidades de planeación, se disminuyen los costos (Cova, 2000), se promueve el desarrollo regional (Hemmasi, 1980), se enfocan los recursos disponibles con el fin de maximizarlos, se detecta heterogeneidad social, así como la correlación de patrones de salud pública con circunstancias económicas y sociales, se reducen los efectos de impactos externos o imprecisiones en las bases de datos (Wise et al., 2001), se facilita la visualización e interpretación de la información de datos georreferenciados, mejora y hace más eficiente la cobertura del producto (Hess, 1971), ayuda a implementar topes en el tamaño de la administración de cada región y nivela la carga de trabajo (Hess, 1971), ayuda a ajustar el tamaño de la fuerza de ventas y áreas de servicio (Zoltners, 1983), facilita la evaluación de la efectividad de nuevos proyectos (Hesse, 1971), refuerza el dibujo de áreas de servicio, de ventas o distritos electorales y ayuda a reducir el tiempo de viajes (Kruskal, 1956 y Miller, 1960). En la Tabla 5 se resumen las ventajas de la regionalización.

Conociendo los beneficios que se obtienen mediante la regionalización es difícil explicar por qué no es una práctica más difundida o incluso obligatoria en cualquier proceso de evaluación de proyectos o planeación. Para entender la problemática que implica la regionalización es necesario considerar que existe un amplio abanico de posibilidades en cuanto a metodologías con sus respectivos algoritmos y diferentes formas de proceder, todas ellas con matices particulares que conducen a diferentes resultados. Tomando en cuenta lo anterior se ha procurado resolver este problema y se han realizado trabajos que proponen una clasificación de las técnicas de regionalización (Duque, 2006, Fisher, 1980, Murtagh, 1985, Gordon, 1996 y Williams, 1995). Aun así queda como menciona Johnston (1968), mucho espacio para la subjetividad tanto en las técnicas como en el trazado de las regiones funcionales, lo cual conduce a pensar que el amplio abanico de posibilidades, así como el espacio para la subjetividad son quizás algunas de las causas primarias por las que la regionalización y el trazo de regiones funcionales son todavía procedimientos que no reciben la atención que merecen.
En este capítulo se presenta la investigación realizada que tiene como objetivo plantear y comparar metodologías dirigidas a realizar una segmentación geográfica bien definida e internamente homogénea utilizando la variable residencia-trabajo, para perfilar subdivisiones

geográficas funcionales o regiones analíticas en España. Se utiliza la variable residencia-trabajo, debido a su disponibilidad, condición material frente a otro tipo de interrelaciones no sometidas a constricciones de distancia, su carácter de proceso recurrente, su factor de vinculación entre los municipios de trabajo y de vivienda y como un intento de describir el patrón de actividad alrededor de áreas urbanas en un día de trabajo típico (Feria, 2008).

Tabla 5. Principales ventajas de la regionalización según la literatura

Optimiza recursos para cumplir necesidades de planeación y disminuye costos.	Cova (2000)
Las regiones funcionales pueden ser consideradas para promover desde ahí desarrollo regional, distribución equitativa de recursos entre clases sociales, expansión de la justicia.	Hemmasi (1980)
Enfoca los recursos disponibles con el fin de maximizarlos, detecta heterogeneidad social, correlación de patrones de salud pública con circunstancias económicas y sociales, reduce los efectos de impactos externos o imprecisiones en bases de datos.	Wise et al. (2001)
Facilita la visualización e interpretación de información de datos geo referenciados, mejora y hace más eficiente la cobertura del producto, ayuda a implementar topes en el tamaño de la administración de cada región y nivela la carga de trabajo, facilita la evaluación de la efectividad de nuevos proyectos.	Hess (1971)
Ayuda a ajustar el tamaño de fuerzas de ventas y áreas de servicio.	Zoltners (1983)
Refuerza el dibujo de áreas de servicio, de ventas o distritos electorales.	Kruskal (1956)
Ayuda a la reducción en el tiempo de viajes.	Miller (1960)

Para comparar la evolución morfológica de las regiones obtenidas se utilizaron los censos de población realizados por el Instituto Nacional de Estadística de España durante 2001 y 2011 y se elaboraron matrices; las filas señalan los orígenes y las columnas los destinos diarios entre residencia y trabajo. Cada celda de la matriz representa el conteo de las personas que viajan diariamente al trabajo entre el lugar indicado en la fila y el destino indicado en la columna.

Esta compleja red tiene como punto neurálgico al individuo cuyas decisiones las cuales se reflejan en la economía, en principio a través de un solo individuo y después en función de las de un cúmulo de individuos, además es un reflejo de la relación laboral existente entre municipios, ya que se reconoce el hecho de que los mercados laborales en la economía actual pueden ser considerados como núcleos, en los que convergen las fuerzas que determinan un estado complejo de flujo y rotación de sus elementos, la producción de una región y por lo tanto de la cantidad de ingresos que las personas y familias tienen disponibles para maximizar su bienestar.

Existen, sin embargo, factores que al ser ajustados influyen para maximizar este bienestar. Es la administración pública quien tiene a su disposición herramientas para ajustar y tratar de maximizar el bienestar, a través de políticas públicas que estén orientadas entre otras cosas, a incentivar la demanda real de bienes y servicios, por lo tanto la producción y con ello la fuerza laboral. Es así, como los requerimientos de la fuerza laboral se determinan con estos ajustes; si tales requerimientos disminuyen, ocurren despidos. En este contexto, el individuo económicamente activo vende su capacidad de trabajo para generar ingreso, con el que se detona la demanda del consumidor y esto a su vez deriva en la creación de mayores empleos e ingresos en un proceso complejo auto sostenido (Carmichael, 1978), por lo que resulta entonces fundamental definir unidades geográficas con el propósito de investigar, de formular políticas públicas y planeación territorial del transporte y vivienda entre otros ya que las áreas administrativas en muchas ocasiones no proveen una perspectiva significativa de la realidad funcional del territorio.

En Gran Bretaña y Estados Unidos, esta delimitación es una práctica estándar y establecida; en Estados Unidos desde 1930 con las FUR´s (*Functional Urban Regions*) y en Gran Bretaña desde 1950 con las TTWA´s (*Travel to Work Areas*), para la medición de condiciones del mercado laboral, reporte de cuentas desagregadas de condiciones laborales, identificación de posibles áreas de asistencia para propósitos de política regional industrial, regionalización del gobierno

local, delimitación de áreas asistidas, estudio del fenómeno laboral, ciclo de fases del desempleo (Casado-Díaz, 2000), o delimitación de distritos industriales (Sforzi, 1990).

En el caso de España, la práctica no está estandarizada a pesar de que su Constitución en los artículos 141.3 y 152.3, confiere la posibilidad de asociación a conjuntos de municipios contiguos en entidades territoriales diferentes de la región o provincia a la que pertenecen y en la Ley de Corporaciones Locales (LRBRL Art. 43), se establece que las áreas metropolitanas son entidades locales compuestas de municipios, de aglomeraciones locales grandes, con vínculos económicos y sociales donde la coordinación y planeación son indispensables.
Es necesario en este punto especificar que la división política y administrativa en España, que se recoge en la Constitución de 1978, divide al país en 17 comunidades autónomas, las cuales representan por su alto grado de autonomía política, la principal referencia para muchas políticas y decisiones de orden jerárquico, pero a pesar de esto la regionalización del territorito español se ha abordado desde una perspectiva académica con diferentes metodologías y finalidades.

Por ejemplo, Casado-Díaz, (2000),aplicó la misma metodología que se utiliza en Gran Bretaña, a la región de Valencia con datos derivados del Censo de Población y Vivienda de 1991, produciendo regiones definidas de forma que la mayoría de los trabajos se ocupan por residentes del área y la mayoría de los residentes de la localidad trabajan en el área.

Boix et al. (2004), apoyados en la teoría del distrito industrial Marshalliano, en otro estudio sobre regionalización, se enfocó a la ubicación geográfica de distritos industriales y evaluó la posibilidad de utilización de tres enfoques distintos para llevar a cabo la regionalización, escogiendo el enfoque funcional para captar lo que llamó policentricidad urbana para aplicarlo en España e Italia.

Por su parte Feria (2008), en un trabajo aplicado únicamente a España justificó el uso de la variable residencia-trabajo en un intento de describir los flujos de actividad cotidiana en una zona metropolitana. Boix et al. (2004), agregó que la identificación de las áreas metropolitanas, podría incrementar el bienestar de los residentes, transformando políticas que propicien competitividad, cohesión social, calidad ambiental y gobernabilidad y que el enfoque de región administrativa, es claramente inadecuado para identificar áreas urbanas integradas social y económicamente; utilizó la variable residencia-trabajo, ya que en el enfoque morfológico se presenta el problema de encontrar ciudades muy pequeñas que por su tamaño

no es posible designarlas como áreas metropolitanas. Feria (2008), tomó en cuenta las relaciones socioeconómicas para formar un área metropolitana, en un enfoque que llama funcional, que captura la estructura espacial urbana.

Existen referencias sobre intentos de regionalización en España por parte de fuentes institucionales. Boix et al., (2004) señalaron que entre estas referencias se destacan los intentos por parte de la Dirección General de Urbanismo del Ministerio de Vivienda en 1965 y 1967, que utilizaron como metodología un enfoque propuesto por Boix (2010), quien regionalizó morfológicamente áreas metropolitanas, tomando núcleos con una cantidad mínima de habitantes y a partir de ahí, agrupó municipios de los alrededores que tenían fuertes relaciones socioeconómicas con el núcleo. Un segundo enfoque surgido de una fuente institucional, se encuentra en el III Plan de Desarrollo Económico y Social de 1972, que propuso tres criterios para identificar áreas metropolitanas: estadístico, desarrollo económico y planeación.

A partir de la revisión de la literatura sobre el tema, presentada enseguida se considera que técnicas relativamente novedosas relacionadas como es el análisis estadístico multivariante, pueden ofrecer una alternativa novedosa en relación al procedimiento de regionalización, tanto en lo que se refiere a la facilidad de aplicación, como a la posibilidad de revisar dichas zonas de manera periódica; además, una ventaja de la metodología de regionalización a base de agrupación, conglomerado o cluster, es que en el proceso pueden llevarse a cabo ajustes finos con el fin de que las regiones funcionales se integren por unidades lo más homogéneas posible entre sí, en términos del fenómeno en investigación y a su vez sean lo más heterogéneas de las demás regiones en el mismo sentido.

3.2 Revisión de la literatura

Una vez establecido que los agentes económicos son sometidos a fuerzas, que ejercen una presión distinta en cada región geográfica, se ha evaluado esta hipótesis con metodologías que analizan la información contenida en variables generadas por fenómenos, en los que se consideran la ubicación y la localización de los agentes económicos y sociales. Carey (1966), Atkinson (2011), Boermans et al. (2011) y Plane (2003), suponen que existen factores que actúan a nivel local y que forman e influyen en la actividad económica y social que se desarrolla en cada región; sin embargo no proponen un proceso formal para formar regiones

analíticas en donde la información bajo estudio se encuentre optimizada previa a su análisis. Por lo tanto, es preciso profundizar en el tema de la formación de regiones geográficas analíticas, tomando en cuenta que al trazar regiones geográficas funcionales, las subdivisiones reflejen y contengan la mayor cantidad de la información bajo estudio, que la regionalización sea coherente, comprobable, fiel a la realidad y que sea fácil de interpretar. En la Tabla 6, se resumen las finalidades de los estudios realizados por estos investigadores.

Tabla 6. Reconocimiento de diferencias en los factores que tipifican a diferentes regiones.

Autor	Finalidad del Estudio	Procedimiento	Metodología
Atkinson (2011)	Planeación urbana y potenciar el desarrollo urbano orientado a tránsito en Phoenix, Arizona.	Tipificación de zonas aledañas a estaciones del sistema de ferrocarriles ligeros en el área metropolitana de Phoenix.	Utiliza doce variables sobre características sociales, demográficas y usos de suelo, en áreas aledañas a estaciones del sistema de ferrocarriles ligeros, para potenciar el desarrollo urbano orientado a tránsito.
Boermans (2011)	Identificar determinantes que crean las distribuciones regionales desiguales de inversión extranjera directa en China.	Análisis Factorial, para identificar la información contenida en múltiples variables.	Resumir información incorporada en cuarenta variables. Encuentra cuatro determinantes: Calidad institucional, costos laborales, tamaño de mercado y factores geográficos.
Hongmian (2002)	Identifica factores que afectan la distribución laboral desigual en el área metropolitana de Atlanta encontrando que factores como: Trabajadores de medio tiempo, cuarteles generales de corporaciones, profesionales bien educados y acceso carretero. Resultaron ser importantes determinantes de localización, siendo crecientemente responsables para la suburbanización de negocios y servicios profesionales.		
Plane (2003)	Exponer y dibujar el tránsito migratorio en Estados Unidos.	Análisis factorial aplicado a flujos migratorios estandarizados estratificados por edad entre áreas económicas propuestas por el Bureau de Análisis Económico.	Agrupa datos de flujos de migración del Censo de 1990. Remarca la importancia de etapa del ciclo de vida reconociendo que es un predictor de las decisiones de migración.

Trabajos recientes de regionalización (Berdegué et al.,2011), (Cörvers et al., 2009), (Mitchel y Watts, 2009), (Noronha y Goodchild, 1992), (Casado-Díaz, 2000), (Salom y Casado, 2007), (Boix y Galleto, 2004) y (Casado et al., 2010), utilizaron diversas metodologías sobre tablas de contingencia que resumen la interacción del trayecto que une las áreas administrativas. Los desplazamientos se expresan en la cantidad (o conteo) de las personas que se mueven entre ellos, con la motivación del trabajo o como en el caso de Noronha y Goodchild (1992), con la motivación del estudio. Entonces, el objetivo es formar grupos de regiones administrativas y obtener una división del territorio que refleje los mercados de trabajo. Un resumen de las metodologías empleadas por los autores anteriormente citados se muestra en la Tabla 7.

Para obtener dichas regiones analíticas, existe un amplio abanico de posibilidades en cuanto a técnicas de regionalización, y es Duque (2006), quien enumeró algunas de las características que pueden encontrarse en cualquier metodología usada para definir regiones analíticas y propuso una clasificación. Entre las características generales, menciona que los métodos deben agregar áreas geográficas en un número de regiones mientras se optimiza un criterio particular, que las áreas dentro de una región deben estar geográficamente conectadas, que cada unidad geográfica puede asignarse a una región únicamente, que debe existir conocimiento previo sobre el proceso, de las variables relevantes para la agregación y del número de regiones a diseñar y por último, que existe la restricción de contigüidad espacial o la existencia de un criterio de agregación.

Algunos autores, señalados en la Tabla 7, lograron una regionalización utilizando matrices de movilidad como base sobre la cual se aplica el algoritmo de Coombes et al. (1986), el cual utiliza el ratio dentro-región a interacciones entre regiones, como base para formar los mercados de trabajo. Mitchel y Watts (2009), mencionaron dos deficiencias de este algoritmo; primero la especificación arbitraria de los valores de los parámetros (75% de autonomía) y segundo, que el proceso de desmembramiento aparecería para generar un conjunto final de agrupaciones con numerosos grupos simples, pero también con algunos grupos muy grandes. Este procedimiento se ha utilizado en el Reino Unido para reportar fenómenos laborales y delimitación de las zonas asistidas. También con este algoritmo se han obtenido diversas conclusiones relativas a los distritos industriales, morfología y evolución del mercado de trabajo, Casado-Díaz (2000), Salom y Casado (2007), Boix y Galleto (2004), Casado et al. (2010). Por ejemplo, Salom y Casado (2007), identifican cuatro comportamientos diferentes que presentan los mercados de trabajo entre 1991 y 2001 en Valencia (España). Estas conductas se identifican comparando, en 1991 y 2001, la cantidad de empleos, la cantidad de ocupados y los

índices que cuantifican o no la apertura de un mercado de trabajo, en términos de la suma de trabajadores que entran y salen. También encuentran un aumento de la movilidad que se refleja en la reducción de 192 a 112 TTWA.

Tabla 7. Metodologías de regionalización utilizadas por diferentes autores.

Autor	Metodología	Base de datos	País (Región)
Berdegué et al. (2011)	Cluster Jerárquico	Censos (1992,2002)	Chile
Cörvers et al. (2009)	Cluster Jerárquico de Ward, Distancia Euclídeana estandarizada	Encuesta sobre comportamiento de viajes (1991/1992,2001)	Países Bajos
Mitchel and Watts (2009)	Cluster Jerárquico Intramax	Censo (2001)	Australia (Nueva Gales del Sur)
Noronha y Goodchild (1992)	Procedimiento Heurístico "*Greedy Add*"	*National Center for Education Statistics* (1979,1981)	Estados Unidos
Casado-Diaz (2000)	Coombes et al. (1986)	Censo (2001)	España (Valencia)
Salom y Casado (2007)	Coombes et al. (1986)	Censos (1991 y 2001)	España (Valencia)
Boix a y Galleto (2004)	Coombes et al. (1986)	Censo España (2001), Censo Italia (1997)	España, Italia
Casado et al. (2010)	Coombes et al. (1986)	Censo (2001)	España (Valencia)

Boix y Galleto (2004), también desarrollaron un estudio relativo a la economía geográfica que utiliza una adaptación del algoritmo de Coombes et al. (1986); un algoritmo heurístico no jerárquico con cuatro etapas principales y una etapa final de calibración o ajuste fino que forma regiones contiguas. Boix y Galleto (2004), identificaron 806 mercados laborales de España, que agruparon 8.100 municipios y 784 para Italia, que agruparon 8.600 municipios y midieron la estructura socioeconómica a través de un conjunto de coeficientes de acumulación industriales anidados, de cada mercado de trabajo identificado. Encontraron zonas que

podrían ser consideradas como distritos industriales en analogía directa con el distrito industrial de Marshall, tanto en España como en Italia y los compararon.

Utilizando el mismo algoritmo que Boix y Galleto (2004), Casado-Díaz (2000), lo aplicó a una matriz de movilidad extraída del Censo de Valencia (España), argumentando que el objetivo de una regionalización de un área de mercado, es definir las unidades geográficas en las que se produce la mayor parte de la interacción entre los trabajadores que buscan empleo y los empleadores que contratan mano de obra (Casado-Díaz, 2000). El algoritmo está calibrado de tal manera, que su objetivo es formar el número máximo de áreas que estén compuestas por al menos 3,500 trabajadores y que la mayoría de los puestos de trabajo dentro de la zona estén ocupados por residentes de esa zona (demanda de autocontención). Encontró 27 áreas de mercado de trabajo local, que agrupan 539 municipios, con estas características. Segmentó la muestra por subgrupos (sexo, ocupación, actividad) y encontró que los mercados relativamente independientes varían en tamaño de un subgrupo a otro. Casado-Díaz (2000), mencionó que esto podría explicarse debido a factores como los ingresos, las responsabilidades de la familia, la propiedad de automóviles o el número de horas de trabajo.

Casado et al. (2010), con el argumento de que cualquier territorio amplio es claramente fragmentado en áreas delimitadas relativamente autónomas, donde un grupo de trabajadores ofrecen sus habilidades y un grupo de empleadores demandan las mismas, adaptó el algoritmo de Coombes et al. (1986), para proponer cuatro diferentes regionalizaciones con regiones contiguas. Estas regionalizaciones se derivan permitiendo un compromiso *trade-off* entre los parámetros de autonomía tanto de oferta como de demanda. Dependiendo de cómo se aplica rigurosamente el requisito de tener 75% de autonomía y el tamaño mínimo requerido, el número de regiones obtenidas puede variar entre 2127 a 2237 en el enfoque más relajado.

Otro procedimiento heurístico para formar regiones funcionales, lo propuso Noronha y Goodchild (1992), pero este método podría quedar atrapado en óptimos locales, estrechando la cantidad de configuraciones a probar y generando configuraciones que muy probablemente mejoren la función objetivo, *greedy and heuristics*, perdiendo así el óptimo global porque el algoritmo sólo tiene en cuenta las mejoras en la función objetivo. Esto podría resolverse mediante el procedimiento de recocido simulado, *Simulated annealing*, propuesto por Kirkpatrick et al. (1983).

Entre los autores que han utilizado un enfoque de regionalización relacionado con el artículo de Casado se puede citar a Berdegué et al. (2011), Cörvers et al. (2009) y Mitchel y Watts (2009), quienes usaron un análisis jerárquico de conglomerados en sus diversas variaciones para proponer una división funcional en Chile, Países Bajos y Australia respectivamente. Cörvers et al. (2009), por ejemplo, antes de la aplicación del análisis de conglomerados (distancia euclidiana, Ward, 1963, variable-z estandarizada) derivaron de la matriz de movilidad normal, otra matriz que captura las distancias funcionales entre grupos. Evaluaron la coherencia de su regionalización en términos de cuatro indicadores económicos: nivel de ingresos, precios de la vivienda, el empleo y las tasas de desempleo, y lo compararon con la regionalización administrativa utilizada por el gobierno de los Países Bajos. Relacionaron las regiones administrativas y funcionales mediante medidas de desempeño económico regional, para tener una idea de cuál de las regionalizaciones era más coherente y por lo tanto debía ser la preferida. Al realizar pruebas de hipótesis de proporciones no se detectaron diferencias estadísticamente significativas, encontrando así que la división funcional no superaba la división administrativa, por lo que no había mucho que ganar en la formulación de políticas mediante el uso de esta división funcional particular.

Con el mismo enfoque metodológico Berdegué et al. (2011) analizaron los cambios socioeconómicos en Chile que se capturan en las variables de demografía, empleo, pobreza, educación y otros. Para calcular la evolución de estas variables, primero realizaron una regionalización mediante clusters jerárquico sobre una matriz de los desplazamientos de trabajo entre todos los municipios del país. Encontraron 103 grupos de regiones funcionales en seis categorías diferenciadas por tamaño (urbana, rural, rural-urbano, entre otras) y calcularon diferentes variables para cada una de las seis categorías en dos años (1992 y 2002). Encontraron cambios entre ambos años, en variables como la tasa de dependencia, la incidencia de la pobreza, la concentración de la renta, la tasa de población económicamente activa de la población total y otras veinte variables relacionadas, tales como el empleo, demografía, educación y grupos étnicos.

Mitchel y Watts (2009), para formar regiones funcionales en Australia utilizaron una variación del análisis jerárquico de conglomerados llamado Intramax sobre una matriz de viajes al trabajo. El método Intramax maximiza la proporción de la interacción total, que tiene lugar dentro de la agregación de las unidades de datos básicos que forman los elementos de la diagonal de la matriz y por lo tanto minimiza la proporción de los movimientos a través de su frontera en el sistema como un todo (Masser y Brown, 1975). El patrón de región funcional

descrito en su documento reduce la dispersión intrarregional en las tasas de desempleo, lo que quiere decir que la técnica tiende a agrupar áreas con mayor homogeneidad. Una aplicación, para 2001 del censo de los datos de población y vivienda de la Oficina Australiana de Estadística, reveló 24 regiones, formadas por 197 áreas estadísticas locales, encontrando que las tasas medias de desempleo regionales eran más altas y las tasas de participación de la fuerza laboral eran menores cuando se aplicó esta nueva geografía. La homogeneidad del comportamiento mejorado de las regiones funcionales también reduce la dispersión intrarregión, medida por la desviación estándar de las tasas de desempleo, lo que sugiere que la fuerza de los flujos de desplazamiento entre zonas contiguas captura la interacción económica entre ellos.

Mitchel y Watts (2009), justificaron su trabajo a través de la evaluación de los índices globales de autocorrelación espacial, que son medidas formales de cómo los elementos de observación cercanos y distantes están relacionados. Las estadísticas globales se pueden descomponer para proporcionar medidas locales de asociación espacial (LISA's), que revelan las agrupaciones estadísticamente significativas por arriba de los valores medios (puntos calientes) y estadísticamente significativos por debajo de la concentración promedio (puntos fríos) de los fenómenos que se investigan. Los mapas LISA's, a nivel de zona postal, revelaron una considerable heterogeneidad espacial en los resultados de la fuerza de trabajo y resaltaron una motivación clave para el desarrollo de la nueva geografía.

La segmentación de la base de datos permite un análisis más profundo de los fenómenos que experimentan las áreas; por ejemplo, Coombes et al. (1988), llevaron a cabo una regionalización de West Midlands (Reino Unido), segmentando la muestra en subgrupos para evitar la generalización de las diferencias sustanciales en el comportamiento del viaje al trabajo de los diferentes sectores de la fuerza de trabajo. En este mismo sentido, Rouwendal (2004), afirmó que es poco realista suponer que todos los trabajadores son idénticos en gustos, como en ingresos.

Green et al. (1986), encontraron que la TTWA agregada, subestima la longitud y la diversidad de las pautas de movilidad de los hombres y sobreestima la distancia del trayecto de las mujeres. Rouwendal (2004), en este aspecto afirmó que las mujeres tienen en promedio trayectos más cortos que los hombres, debido a que un viaje largo suele ser más problemático para ellas.

Beckman (2008), segmentó su muestra teniendo en cuenta un número finito de clases de los trabajadores relacionados con el tamaño del hogar y no con el género. Es decir, sostiene que las distancias de trayecto tienen también no solo relación con el género sino tambiéncon las características del hogar. Acerca de este enfoque Rouwendal (2004), sugirió que la economía urbana podría predecir una clara relación entre las características del hogar (incluidos los ingresos) y la distancia de los desplazamientos, lo que implica que las diferencias entre los trabajadores dan lugar a discrepancias en el comportamiento de los desplazamientos.

Owen y Green (2000), destacaron que también hay patrones espacialmente concentrados entre los grupos étnicos, señalando la posibilidad de que este fenómeno podría ser válido para los diferentes grupos de ocupación, sectores de actividad, género o incluso por edad.

Green (1986), demostró que hay diferentes patrones de desplazamientos al trabajo; utilizó una base de datos segmentada por género y encontró diferentes grados de índices de autocontención de la oferta y de la demanda entre hombre y mujer. Salom y Casado (2007), indicaron que es muy difícil desunir los factores asociados al territorio de los relacionados con las características de los individuos, dado que se entremezclan, destacando las diferencias en relación a la movilidad entre los diversos subgrupos, como el género, composición del hogar y los diferentes niveles de ingresos y educación. Estas desigualdades las señalaron muy bien Berdegué et al. (2011), quienes mencionaron que se reflejan a través de las decisiones de las personas acerca de su lugar de vida, sus capacidades, los sistemas de género, las etnias y otros factores.

Casado-Díaz (2000), reconoció cómo la relación entre el lugar de trabajo y residencia varía tanto territorialmente, como en función del sexo, sector de actividad y ocupación; explicó que las diferencias entre los subgrupos podrían explicarse por factores como los ingresos, las responsabilidades familiares, la propiedad de automóviles y el número de horas de trabajo; señaló un conjunto de áreas de mercado laboral específicas superpuestas, por sexo, edad, posición socioeconómica, actividad profesional e industrial, producto de diferente acceso entre los grupos de transporte y otros recursos.

Cuando las regionalizaciones se realizan utilizando datos agregados tales divisiones son “promedio” de distancia (Green et al. 1986). Un análisis completo requiere el conocimiento de oferta de los medios de transporte públicos y de cómo satisface esta oferta las necesidades de la población ocupada, aunque la utilización del numero de viajes diarios al trabajo entre

municipios como fuente primaria de información trae implícitas las barreras geográficas más importantes, ríos, caminos mayores, líneas de ferrocarril, espacios abiertos, tal y como lo señaló Openshaw, (1998).

A continuación se presenta la revisión bibliográfica subdividida en función de cómo proceden los algoritmos o cuál es la base teórica para proceder a la realización de la regionalización.

3.3. Regionalización mediante Algoritmos Convencionales.

En esta sección se revisan los trabajos de investigación donde se han formado determinadas regiones, en las cuales sus elementos fueron lo más homogéneos posible entre sí y se explica la forma en cómo procede cada algoritmo de regionalización.

En estos procedimientos las regiones funcionales se forman iterativamente, y en cada iteración se optimiza la información bajo investigación la cual se encuentra resumida en un conjunto de variables. En este tipo de procedimientos no se especifica ningún tipo limitante de contigüidad espacial, sino que la contigüidad espacial se aplica *a posteriori*.

Fisher (1980), comparó tres medidas de similitud, distancia de Manhattan, distancia euclídea y correlación, entre los miembros y comparó dos métodos iterativos no jerárquicos de regionalización, K-Means y K-Centroide, caracterizados por funciones objetivo orientadas a optimizar las regiones manteniendo el principio de homogeneidad interna.

El primer método aplicado por Fisher (1980), es una ampliación del método presentado por MacQueen (1967), llamado K-Means cuyo principal propósito es dividir una población de dimensión N en K conjuntos. El proceso forma regiones que son razonablemente eficientes en el sentido del criterio de la varianza, medida desde un punto de la región llamado centroide (media), hasta cada uno de los elementos que integran la región. La metodología K-means ofrece la posibilidad de tomar una partición *a priori* que refleje una hipótesis en particular o una partición aleatoria, iniciando posteriormente un proceso de reubicación iterativo, que verifica si es posible mejorar la partición mediante la reubicación de cada unidad a regionalizar de acuerdo al criterio de la varianza antes mencionada. Si no hay mejora en la varianza con el cambio de cada una de las unidades no se realiza ningún cambio. Después de cada reubicación,

el centroide de la región receptora se recalcula. El ciclo iterativo termina cuando se examina la ubicación de todas las unidades básicas.

El segundo método evaluado por Fisher (1980), se denomina K-Centroide y fue introducido por Diehr (1971). Ofrece la posibilidad de elegir entre dos criterios de optimización alternativos, en los que para regionalizar, primero se escoge el centro de cada región encontrado por medio de un algoritmo de búsqueda de tres pasos en el que los centros de las n regiones son seleccionados de acuerdo al criterio de minimización de la suma de las distancias desde el centroide hasta el resto de las unidades a regionalizar y una vez elegidos los centroides, el resto de las unidades a regionalizar se asignan al centroide más cercano.

De acuerdo a Lankford (1969), el método funciona bien cuando los datos tienen un patrón muy claro, pero cuando los grupos están altamente dispersos la agrupación se vuelve inestable. Lankford (1969), por su parte también utilizó algoritmos de aglomeración convencional y en un intento de evaluar su desempeño comparó tres algoritmos de agrupación: centroide, Ward y limitado por vecindad; evaluó los tres algoritmos con datos predefinidos por él mismo, e identificó el que mejor reconoce el patrón.

El uso de algoritmos limitados por vecindad requiere que se suministren datos adicionales sobre conjuntos vecinos y regionaliza igual que el ojo humano. Lanckford, menciona que el empleo de conjuntos limitados por vecindad, para determinar conectividad, libera al algoritmo de muchos problemas y no tiene preferencias inherentes por grupos esféricos o del mismo tamaño como el algoritmo de Ward (1963). El algoritmo tampoco requiere ninguna distribución particular de los datos.

El algoritmo de Ward (1963), es una rutina que examina la matriz de distancias entera entre los elementos y une los elementos que producen el mínimo incremento a la suma de las distancias cuadradas inter grupales. Las distancias inter grupales y la matriz de distancias se actualizan en cada paso. Las distancias entre los elementos que son unidos se sustituyen por la media del grupo o centroide, lo que produce el problema conocido como “encadenamiento”, que es la unión de elementos hacia una sola dirección, ya que la media o el centroide, se va moviendo a medida que se van uniendo elementos al grupo.

El método de Ward (1963), mediante la minimización de la suma de las desviaciones al cuadrado desde la media del grupo, puede mantener el tamaño aproximadamente igual,

mantener grupos de alta densidad, desarrollar grupos que son esféricos en tamaño y según Lankford (1969), el algoritmo es más eficiente que el centroide, ya que puede trabajar con datos con un patrón menos claro y producir una agrupación limpia, aunque ya que emplea la media del grupo para actualizar la matriz de distancias en cada paso, sufre hasta cierto punto del mismo problema de encadenamiento que el método del centroide.

Tabla 8. Algoritmos convencionales

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Fisher (1980)	Compara dos métodos iterativos no jerárquicos (K-Means y K-Centroid), de regionalización, caracterizados por funciones objetivo orientadas a optimizar regiones funcionales.	K-Means	Mantiene el principio de homogeneidad interna. De fácil programación y computacionalmente económico.	K-Means es muy sensible a *outliers* o valores atípicos ya que son elegidos como núcleos al inicio del algoritmo. Se debe especificar el número de regiones *a priori.*
		K-Centroid	Funciona eficientemente en datos con un patrón claro o densamente agrupado.	Surge problema de encadenamiento cuando los datos se encuentran dispersos.
MacQueen (1967)	Describe un proceso para dividir una población con dimensión *n* en *k* conjuntos. Las particiones son razonablemente eficientes en el sentido de a varianza intraclase.	K-Means	Que el proceso es fácilmente programable y económico computacionalmente lo que hace posible procesar muestras muy grandes.	Sensibilidad a valores extremos. Únicamente es posible utilizar la medida de Distancia Euclídea para medir diferencia entre elementos. Problema de encadenamiento ya que el centroide cambia de posición solamente en una dirección a medida que se añaden elementos.
Diehr (1971)	Introduce un método para regionalizar en el cual se escoge primero el centro de cada región de acuerdo al criterio de minimización de la suma de las distancias desde el centroide hasta el resto de las unidades a regionalizar.	Modificación del método de MacQueen: K-Means	Las regiones formadas son internamente homogéneas en función varianza interna. Económico computacionalmente	Presenta problema de encadenamiento. Depende únicamente de una medida de similitud, la distancia euclídea.

Tabla 8. Algoritmos convencionales *(continuación)*

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Lankford (1969)	Compara algoritmos de agrupación que son aplicables a problemas de regionalización.	Ward (1963)	Mantiene el tamaño aproximadamente igual, mantiene grupos de alta densidad, desarrolla agrupaciones esféricas.	Sufre hasta cierto punto del mismo problema del encadenamiento que centroide. No existe metodología objetiva para determinar el tamaño, número y forma de las regiones.
		Centroide	Funciona bien cuando los datos tienen un patrón claro o densamente unido.	No existe ningún método objetivo para determinar el nivel en el que seleccionar regiones. Problema del encadenamiento debido a que el centroide se mueve a medida que los elementos se suman al grupo solamente en una dirección.
		Limitado por vecindad	No tiene preferencias inherentes, como el algoritmo de Ward (1963), por grupos esféricos o del mismo tamaño	Se tiene que suministrar información adicional sobre conjuntos vecinos.

Debe destacarse que problemas metodológicos importantes son el tamaño de los grupos y su número y su forma; puesto que la agrupación producida es un conjunto de particiones, no existe absolutamente ningún método objetivo para determinar el nivel o tamaño de las regiones.

3.4. Maximización de Compactación Regional.

Al trazar regiones funcionales, existen decisiones que deben ser tomadas por el investigador de manera subjetiva y arbitraria, como por ejemplo las restricciones de homogeneidad infra regional, las cuales Openshaw (1998), fija arbitrariamente en un 75%. Estas restricciones son generalmente expresadas como porcentaje y buscan que las regiones funcionales sean lo más homogéneas posible.

Las decisiones pueden ser tomadas a partir de fijar límites en medidas de correlación intra área para lograr la maximización de la homogeneidad social en la áreas obtenidas (Martin, 2001), o como en el trabajo de Hess et al. (1965), donde se utilizaron medidas de similitud de pares, que reflejan el espacio métrico (distancia Euclídea, distancia de Manhattan o block) o como Kaiser (1966) que usó medidas de igualdad de la población en una zona quien tomó como criterio de agrupación, la minimización de las diferencias o la maximización de la similitud de la población o el mismo tamaño de distrito en términos del tamaño de la población.

Con más detalle Kaiser (1966), utilizó para regionalizar, una función objetivo que contiene dos componentes ponderados. Al primero le llamó "Igualdad Poblacional" (*population e equality*). En este componente la población de cada región obtenida por el método debe ser tan cercana como sea posible al ratio entre la población total y el número de regiones a diseñar, es decir intenta distribuir al total de la población en regiones compuestas por la misma cantidad de personas. La compactación es el segundo componente. Para simplificar el método propuso estandarizar o expresar en unidades generales y no en alguna unidad de medida en particular, utilizando el área de un círculo del mismo tamaño que el distrito a ser creado y así evitar la integración numérica de la fórmula del momento de inercia de cada distrito, ya que se hubiera requerido varios cientos de horas de labor extremadamente tediosa para producir información de relativamente poca importancia y consideró que el error producido en la aproximación al utilizar el área de un círculo, (el círculo es la figura plana con el menor momento de inercia) produce un error que no es significativo. La mayor parte del algoritmo de optimización consiste en dos tipos de iteraciones, la primera implica mover un área de su región original a otra región, únicamente si la función objetivo presenta una mejora y la segunda, el intercambio de todo par de áreas que pertenezcan a diferentes regiones. La convergencia se alcanza cuando todos los posibles movimientos e intercambios no incrementen la función objetivo.

Weaver (1963), desde un punto de vista legal planteó que en los Estados Unidos, las Cortes Federales tienen jurisdicción para revisar la constitucionalidad de las particiones territoriales y los distritos electorales, y puesto que la circunscripción electoral usualmente afecta el balance político de una legislatura, una Corte que lleve a cabo una partición territorial con fines electorales, es muy probable que se convierta en sujeto de crítica y apelaciones por parte de los partidos. Weaver (1963), mencionó que tal criticismo podría crear la percepción de que la Corte está actuando por motivación política con el deseo de beneficiar a un interés partidario en particular. Para evitar los efectos colaterales volátiles, las repercusiones políticas y limitar la

discrecionalidad en la creación de nuevos distritos electorales, adoptó una formula mecánica o algoritmo que delimita los distritos de manera no discrecional, una vez que se determinan los principios generales de representación popular. En la etapa final del proceso, se toman en cuenta los principios de contigüidad e igualdad poblacional y mencionó que aunque se sigan estos dos principios las delimitaciones pueden ser dibujadas de muchas maneras, cada una con diferentes repercusiones políticas.

Para asegurar la compactación regional, Weaver (1963), al igual que Kaiser (1966), minimizó el momento de inercia de una figura plana; en general el momento de inercia es una medida de física que en términos estadísticos es una suma de cuadrados, en alguna unidad de medida, desde el centro de masa hasta todos los centros de los elementos que conforman un cuerpo plano.

Kaiser (1966), propuso la integral del momento de inercia ya que consideró que el número de puntos en el terreno de un distrito es infinito. De acuerdo a la propuesta de Kaiser (1966) y Weaver (1963), si las personas se encontraran igualmente distribuidas en la región, el centro de gravedad de la población se encontraría en el centro de la región y al minimizar esta medida se asegura que la conformación obtenida de todos los distritos esté compuesta por unidades lo más aglomeradas posible; así con esta definición de compactación se evita que los distritos producidos por el procedimiento tengan formas alongadas y además se crean distritos que coinciden con las áreas de alta densidad de población.

La medida de compactación que escogió Weaver (1963), hace posible que se tomen ciertas similaridades matemáticas entre el problema de la circunscripción electoral y el problema de asignación de órdenes de compra de clientes a almacenes con localizaciones específicas de manera que se minimicen costos. Weaver, consideró que los dos problemas son muy parecidos, por lo que utilizó el mismo algoritmo, el cual asigna unidades geográficas (clientes) a distritos electorales (almacenes) de manera que se minimice la medida de compactación (costos).

Bacao (2005), se enfrentó al problema de diseño de zonas, específicamente al diseño de distritos electorales y señaló que el método para su diseño tiene que ser capaz de proveer zonas contiguas, compactas e igualmente pobladas. Por lo que propuso medidas que capten estos tres requerimientos y un algoritmo que obtiene múltiples soluciones al problema (cada solución es codificada en una cadena de caracteres y cada una tiene diferentes atributos

métricos). Dos operadores (mutación y combinación) cambian cada solución; la solución que mejor resuelve el problema es seleccionada como óptima. El algoritmo converge una vez que se han obtenido 5000 generaciones de soluciones, sin mejoras producidas por las mutaciones y combinaciones.

En estos trabajos la formación de las regiones tiene como eje principal la compactación de la información bajo estudio y proponen una optimización de esa información para que las regiones funcionales producidas estén perfiladas con las características deseadas.

Tabla 9. Otros métodos de regionalización/compactación

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Martin (2001)	Reexamina el procedimiento de diseño de zonas automatizado adoptado para la creación de los resultados geográficos del Censo 2001 del Reino Unido.	Algoritmo que con función objetivo en la que se especifica tamaño de población de las áreas producidas, contigüidad.	Produce regiones contiguas e iguales en términos de tamaño de población.	El algoritmo propuesto es enumerativo por lo que está limitado en términos de tamaño del problema.
Openshaw et al. (1998)	Buscan profundizar, ampliar y confirmar los resultados obtenidos por Martin (1997) con un método para automatizar el diseño geográfico del Censo 2001 del Reino Unido, al utilizar algoritmos de diseño de zona más sofisticados que los utilizados por Martin y datos espaciales con resolución más fina.	Optimización no lineal en función objetivo, sujeta a restricciones de homogeneidad y tamaño, es una modificación de método usado por Martin (1997) con modificaciones para manejar restricciones y diferente función objetivo.	Método flexible que permite añadir restricciones explícitas a las zonas que son producidas.	La decisión sobre el nivel que deben tener las restricciones de la función objetivo optimizada, deben ser arbitrarias y subjetivas.
Weaver (1963)	Plantea un método para particionar objetivamente, basado en fundamentos matemáticos, para formar distritos electorales de manera que se limite la discrecionalidad en la partición con el fin de evitar beneficios parciales a partidos políticos.	Algoritmo que optimiza minimizando la compactación regional formando distritos electorales mediante la asignación de poblaciones de unidades geográficas, multiplicadas por su distancia hasta un centro de gravedad de población del distrito.	Produce regiones contiguas, compactas y con regiones muy similares en términos de tamaño de población.	Es necesario especificar el número de regiones a obtener, esta cantidad en la investigación actual se desconoce.

Tabla 9. Otros métodos de regionalización/compactación (*continuación*)

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Kaiser (1966)	Por disposición de La Suprema Corte de los Estados Unidos propone un procedimiento para formar distritos legislativos iguales en términos de cantidad de población.	Optimización de función objetivo ponderada con restricciones de compactación e igualdad en términos de cantidad de población. Procedimiento que cuantifica el área de las regiones en cada paso para buscar regiones compactas.	Produce regiones compactas y con el mismo tamaño en términos de población.	Para represar y simplificar su procedimiento estandariza la medida de compactación (momento de inercia) utilizando el área de un círculo y así evita la integración matemática del momento de inercia de cada distrito.
Bacao(2005)	Propuesta de utilización de un algoritmo (genético) que lleva acabo mutaciones y combinaciones para resolver el problema de distritación electoral.	Algoritmo que inicia con una partición aleatoria y para obtener nuevas soluciones realiza dos operaciones, combinaciones y mutaciones; elige la solución que mejor resuelve el problema.	Produce zonas contiguas, compactas e igualmente pobladas.	Necesaria matriz de contigüidad.

3.5. Regiones Sembradas

La utilización de procedimientos supervisados para la formación de regiones funcionales implica que se tiene que conocer información de antemano, como por ejemplo el número de zonas que se tienen que producir con la regionalización, como es el caso de los autores Taylor (1973) y Openshaw (1977b), quienes al principio de su algoritmo seleccionan las regiones núcleo en un procedimiento llamado regiones sembradas. Openshaw (1977b), utilizó cuatro modelos con los que optimizó tres matrices, matriz de contigüidad, matriz de interacción y matriz de costos de viaje, para demostrar que los efectos de las metodologías de zonificación o regionalización en el desempeño de los modelos de interacción espacial no son triviales.

Openshaw, reconoció que los patrones de interacción intra zonales e interzonales, son críticamente dependientes en la selección de barreras zonales y tamaños relativos y propuso un modelo predictor de la interacción en el que utilizó variables como número de personas que viajan desde la zona i hasta la zona j y número de personas que finalizan su trayecto en la zona j, así como el tiempo de viaje entre dichas regiones. En el modelo la cantidad de interacción predicha es directamente proporcional a la cantidad de personas que tienen como origen la zona i, inversamente proporcional al tiempo de viaje entre la zona i y la zona j y directamente proporcional a la cantidad de personas que terminan su viaje en la zona j. Para resolver el problema para modelos de interacción espacial utilizó un procedimiento de optimización no lineal sujeto a restricciones; propuso en su procedimiento que el proceso de regionalización produzca regiones óptimas para que los parámetros y desempeño de los modelos de predicción de interacción sean lo más ajustados y eficientes posible y utilizó el algoritmo de Taylor (1973), para demostrar que los efectos son grandes en los parámetros de los modelos de predicción de interacción producidos de regionalizaciones aleatorias.

Taylor (1973), desarrolló un modelo de optimización, que enfocó particularmente en las consecuencias de diferentes regionalizaciones para varios partidos políticos y en la optimización, es decir en el trazo de distritos para maximizar un cierto factor, de manera que la regionalización resulte lo más provechosa posible. Taylor también mencionó que en el caso de la circunscripción electoral, el número final de clases o regiones debe ser conocido de antemano, por lo que es propicio un enfoque de agrupación jerárquica de un nivel. El algoritmo inicia seleccionando aleatoriamente una cantidad fija de áreas núcleo, y posteriormente posiciona el resto de áreas con el núcleo más cercano. Dichas áreas son inmediatamente contiguas a los núcleos y continúa seleccionando hasta que todas las áreas estén posicionadas con uno de estos núcleos. El algoritmo regionaliza sujeto a restricciones, una de las cuales es la contigüidad, puesto que los partidos no pueden operar en circunscripciones ampliamente fragmentadas. Establece una función objetivo con la que pretende minimizar las desviaciones de representación proporcional de los partidos que es producida por una solución, no enfatiza en la homogeneidad de la población, y el rango de población de distritos no debe ser muy grande por razones prácticas de organización y para asegurar algún grado de comparación en la representación de distritos.

Tabla 10. Procedimiento Regiones Sembradas

Autor	Finalidad	Metodología	Ventajas	Desventajas
Taylor (1973)	Propone un algoritmo para simular y determinar las características de cualquier número de planeaciones o particiones, particularmente las consecuencias de estas planeaciones para varios partidos políticos. Menciona que el papel alternativo del ordenador es trazar múltiples combinaciones de distritos para maximizar u optimizar cierto factor.	Algoritmo de regionalización para sistemas bipartidistas con restricciones de contigüidad, proporcionalidad y rango de población.	Las regiones obtenidas son aproximadamente del mismo tamaño en términos de votantes. El algoritmo antes de regionalizar verifica que la región se ajuste en términos de proporcionalidad para partidos políticos.	El algoritmo es simplemente enumerativo y por lo tanto limitado en términos de tamaño de problema. No es eficiente debido a que enumera todas las combinaciones posibles. Establece límites de manera subjetiva. Aplicado a sistema bipartidista únicamente.
Openshaw (1977b)	Con cuatro modelos para explicar interacción espacial demuestra que los diferentes sistemas para formar zonas, tienen efectos en valores de parámetros de modelos explicativos de interacción espacial. Propone un método de regionalización óptimo para producir modelos lo más ajustados posible de predicción de interacción.	Los modelos de interacción espacial utilizando datos espacialmente agregados toman en cuenta patrones de comportamiento espacial.	Resume el problema de diseño de zonas o regionalización en cómo se representa el espacio en un modelo espacial, tal como los modelos de interacción espacial.	La regionalización la usa únicamente como un paso *a priori* a producir un modelo de predicción de interacción. Ajusta las regiones producidas para que el modelo de interacción sea más ajustado, lo que podría ser considerado según Openshaw y Gerrymandering, o distritación para beneficio de intereses de un partido político.

3.6. Análisis Factorial para trazar regiones funcionales

En la literatura sobre regionalización existen ejemplos en los que se ha utilizado el análisis factorial para formar regiones analíticas. La justificación sobre el uso de esta técnica es que tiene la capacidad de descubrir comunalidades ocultas que no se aprecian explícitamente en los datos y por tanto no pasa por alto información relevante mediante decisiones subjetivas. Existen ejemplos empíricos (Goddard 1970, Hemmasi 1980, Illeris 1968, Holsam 1980,

Mitchelson 1994 y Wheeler 1989), de formación de zonas auto delimitadas y auto contenidas a partir de sistemas complejos de vinculación geográfica, y que buscan un patrón de vinculación que elimine ruidos menores y se concentre en los elementos troncales del sistema. Las regiones resultantes por sus vínculos internos, las hacen agrupaciones diferenciadas por el alto grado de actividad interna, contiguas o no fragmentadas, sin omisión por duplicación o traslape y sin necesidad de una matriz de contigüidad. Aunque el uso del análisis factorial está limitado debido a que es preciso que las variables cumplan con los supuestos de normalidad, homocedasticidad y linealidad.

Goddard (1970), hizo énfasis en la ubicación de zonas auto delimitadas y auto contenidas en el centro de Londres. Dichas zonas se infieren utilizando datos de traslados en el sistema de taxis. En la misma línea metodológica, Hemmasi (1980), presentó los resultados de un trabajo de regionalización con datos de flujos de migración bruta permanente, los cuales son la única información sobre vinculación geográfica disponible para Irán, con propósitos de planeación e identificación de puntos geográficos nodales, desde donde proyectar políticas de desarrollo regional en un marco de desagregación nacional de la economía en un sistema de regiones, con el fin de lograr desarrollo socioeconómico nacional y equidad interregional. Illeris (1968), aplicó un análisis factorial, a una matriz del número de llamadas telefónicas entre distritos en Dinamarca, y los factores resultantes del análisis factorial se interpretaron como indicadores de centros regionales y sus zonas de influencia; así encontró lugares con un alto grado de centralidad en términos de tráfico telefónico con sus alrededores y delimitó las zonas de influencia de estos lugares centrales.

Holsam (1980), regionalizó usando una red de transporte australiana al aplicar una técnica de análisis factorial propuesta por Cattell (1965), en la que las soluciones factoriales son el resultado de factorizar las matrices de correlación entre los factores obtenidos del análisis de componentes principales, donde se utilizó la estructura simple de rotación oblicua; el autor concluyó su estudio con la aplicación de un análisis de componentes principales a esta matriz y obtuvo una solución de regionalización, basada en la similaridad de accesibilidad o perfiles de destinos de los nodos. Mitchelson (1994), utilizó flujos de información entre ciudades americanas, para medir la participación de centros metropolitanos en los flujos de información internacional en los Estados Unidos, con la idea de que las ciudades son como un sitio de interconexión donde confluyen los actores.

Tabla 11. Análisis Factorial para el trazo de regiones funcionales

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Goddard (1970)	Examina el problema de medir la relación entre patrones de movimiento y la ubicación de actividades dentro del centro de la ciudad de Londres. Utiliza análisis factorial y correlación en los datos para determinar la estructura subyacente del sistema de flujos de taxis que se encuentran resumidos en una matriz origen destino.	Análisis Factorial sobres matriz de viajes origen destino.	Técnica que elimina el ruido de flujos menores y se concentra en elementos básicos del sistema.	Los datos incluyen solamente traslados en taxi, ignorando los vínculos hechos a pie, vehículos particulares o transporte público. No realiza ninguna transformación a los datos.
Hemmasi (1980)	Subraya el problema que implica la tarea de reconciliación entre un conjunto de subdivisiones administrativas imprecisas y arbitrariamente definidas y un conjunto de regiones de planeación multa-propósito.	Análisis de una matriz de origen destino por medio del análisis de componentes principales.	La técnica identifica los flujos dominantes para regionalizar.	En análisis factorial solamente es posible utilizar la correlación como medida de similitud.
Illeris (1968)	Localización de centros regionales y sus zonas de influencia a nivel nacional en Dinamarca	Análisis factorial aplicado a una matriz de número de llamadas telefónicas entre 62 distritos. Se excluyen las llamadas interdistrito.	Regiones resultantes son agrupaciones diferenciadas por la fortaleza de sus vínculos internos.	No propone ninguna transformación a los datos, no menciona ninguna prueba de normalidad de las variables ya que el análisis factorial es limitado en cuanto a la normalidad que presentan las variables.
Holsam (2010)	Critica el uso de análisis factorial y el de componentes principales en investigación geográfica ya que evoca críticas debido a la frecuentemente encontrada dificultad que implica definir factores y componentes. Amplía la discusión sobre la utilización de análisis factorial de alto orden y demuestra su utilidad a través de su aplicación a datos de una red de transportación aérea.	Con una matriz que contabiliza origen destinos. Aplica análisis factorial	Forma Regiones autocontenidas y auto delimitadas que minimizan los ruidos menor y se concentran en los ejes principales del sistema	El análisis factorial es limitado en cuestión de normalidad de las variables

Tabla 11. Análisis Factorial para el trazo de regiones funcionales (*continuación*)

Autor	Finalidad del Estudio	Metodología	Ventajas	Desventajas
Mitchelson (1994)	Revela patrones sistemáticos, regionaliza flujos de información con análisis de componentes principales, en una red de ciudades, e identifica los principales nodos de influencia y las ciudades secundarias sobre las que se ejerce.	Utilizando datos de envió de paquetería de la compañía Federal Express entre ciudades de todo el mundo.	No es necesario tomar decisiones subjetivas pues el análisis factorial identifica los principales patrones en los flujos.	El análisis factorial está limitado en cuanto a los supuestos de normalidad, homocedasticidad y linealidad de las variables.
Wheeler (1989)	Utiliza datos de envíos de paquetería nocturna para examinar los flujos de información entre 48 zonas metropolitanas de Estados Unidos, con el fin de analizar el patrón de flujos de información intermetropolitana.	Analiza con componentes principales una matriz origen destino de flujos de información.	Aísla patrones de flujo espacial comunes, para conocer la estructura de los flujos.	Solamente es posible utilizar una sola medida de similaridad (Correlación de Pearson).
Palm (2002)	Examina patrones internacionales de comunicaciones, documentando una cercana relación entre indicadores internacionales de conectividad incluyendo, volumen de comercio internacional, turismo y migración	Aplica análisis factorial a una matriz en la que se resume la cantidad de minutos de llamadas telefónicas entre 136 orígenes y 136 destinos internacionales.	El análisis factorial permite que las variables saturen en varios factores de manera simultánea, así es posible verificar las relaciones entre las variables (inversa o directamente proporcionales).	La utilización de llamadas telefónicas implica que debe existir lenguaje en común, acceso a equipo, mensaje apropiado para el teléfono, costos adecuados

3.7. Datos

En este estudio se utiliza la base de microdatos o registros primarios de los censos de población y vivienda elaborados por el Instituto Nacional de Estadística (INE) para los años 2001 y 2011. La información está dividida en municipios, que son en la actualidad la mínima

unidad administrativa geográfica en España; esta restricción particiona la tarea del diseño de áreas en aproximadamente 8400 municipios o unidades geográficas separadas y cada municipio se procesó y resumió en una matriz de conteo de personas que viajan entre municipios por razones laborales.

Las encuestas realizadas durante los Censos del 2001 y del 2011, permiten tener en detalle un conjunto de variables asociadas a las personas: género, ocupación, lugar de residencia expresado en código postal, entre otros, al hogar: tamaño del hogar, número de miembros y otros y al desplazamiento : tiempo de desplazamiento, medio de desplazamiento, municipio de trabajo o estudio expresado en código postal.

El censo se realiza con periodicidad de diez años por lo que es posible detectar cambios generales o evolución en la información y por lo tanto en el espacio demográfico, causados por mejoras y desarrollo de nuevas infraestructuras de transporte o apreciar en qué medida la crisis económica ha alterado la estructura de movilidad o cómo ha cambiado la situación del mercado laboral.

El censo recoge los datos relativos a la duración del viaje, la actividad de cada persona expresada en términos del Directorio de la Clasificación Nacional de Actividades de España (CNAE) y la ocupación en términos del Directorio de la Clasificación Nacional de Ocupaciones de España (CNO).

La utilización de microdatos permite realizar exploraciones muy específicas y presentar una síntesis. En este estudio se utiliza una muestra segmentada en nueve estratos en función de cuatro ocupaciones: trabajadores manuales calificados y no calificados, trabajadores no manuales, supervisores, profesionales y técnicos, tres actividades: agricultura, manufactura y construcción y del género: Masculino y femenino.

Despues de separar a la población por género, se llevó a cabo un recuento de los viajes al trabajo entre municipios, resumiendo el conteo en nueve matrices, una para cada estrato, realizando este proceso para los años 2001 y 2011. En la Tabla 12 se resumen las variables analizadas.

Tabla 12. Breve descripción de las bases de datos utilizadas en este estudio

Estrato	Censo 2001	Censo 2011
	Clave de Clasificación Censal:	Clave de Clasificación Censal:
Mujeres	06	06
Hombres	01	01
	Clasificación Nacional de Actividades (CNAE-93):	Clasificación Nacional de Actividades(CNAE-09)
Agricultura	01	01
Manufactura	15-37	10-33
Construcción	45	41,42,43
	Clasificación Nacional de Ocupaciones (CNO-94):	Clasificación Nacional de Ocupaciones (CNO-11):
Administradores, Profesionales y Técnicos	10-35	11-38
Supervisores y Trabajadores no Manuales	40-53	41-59
Trabajadores Manuales Calificados	60-78	61-78
Trabajadores Manuales no Calificados	80-97	81-97

Fuente: Instituto Nacional de Estadística, España; Censos de Población y Vivienda 2001 y 2011

Con esta muestra se captan los principales desplazamientos de movilidad cotidiana, a pesar de que se trabaja con un segmento de la información de los censos y como ya se mencionó la segmentación de la base de datos permite un análisis más profundo de los fenómenos que experimentan las áreas

Es necesario mencionar que la muestra obtenida es una descripción de comportamientos y decisiones personales e individuales frente a la necesidad de movilidad obligada por razones de trabajo y son una medida de la demanda de movilidad, de medios de transporte que requiere una población ocupada.

Es importante destacar que después de aplicar la metodología las matrices resultantes fueron cuadradas y no simétricas; en las filas se señalaron los orígenes de la población y en las columnas los destinos diarios entre su residencia y lugar de trabajo. Cada celda de la matriz representa el conteo de las personas que viajan diariamente al trabajo entre el lugar indicado en la fila y el destino indicado en la columna. Esta compleja red tiene como punto neurálgico al individuo y a sus decisiones, las cuales se reflejan en la economía, en principio a través de la decisión de un individuo y después en función de las decisiones de un cúmulo de individuos, además de ser un reflejo de la relación laboral existente entre municipios, ya que se reconoce

el hecho de que los mercados laborales en la economía actual pueden ser considerados como núcleos en los que convergen las fuerzas que determinan un estado complejo de flujo y rotación de sus elementos, la producción de una región y por lo tanto de la cantidad de ingresos que las personas y familias tienen disponibles para maximizar su bienestar.

Se debe señalar que en el Censo de Población y Vivienda realizado por el INE del 2001 y del 2011, la información sobre el lugar de residencia y el lugar de trabajo para municipios con menos de 20,000 habitantes ha sido recodificada para salvaguardar el secreto estadístico, por lo que no es posible hacer un análisis de esos municipios y según Jofre-Monseny (2009), la población promedio de los municipios españoles es de 5,000 habitantes de acuerdo a lo que señaló cuando comentó sobre el trabajo de Viladecans-Marsal (2004), en el cual se analizó el papel que tienen las economías de aglomeración en la ubicación industrial a nivel municipal en España. Por otra parte Feria (2008), menciona que ciudades con más de 100,000 habitantes muestran mayor fortaleza demográfica en un sistema metropolitano y que es el tamaño promedio de la mayoría de los municipios españoles.

3.8. Metodología

El proceso de producir regiones analíticas debe permitir que puedan llevarse a cabo ajustes para que las características deseadas como compactación, contigüidad (Garfinkel 1970), homogeneidad social, restricciones impuestas por tamaño, forma o cuestiones electorales (Rossiter 1981), así como por las barreras naturales propias de la geografía, por ejemplo, redes de carreteras, montañas, ríos o lagos (Zolteners y Sinha, 1983), que actúan como limitantes de conectividad territorial (Horn 1995), se vean reflejadas en las regiones obtenidas. En este trabajo se planificó y realizó un análisis comparativo de los resultados obtenidos con distintas metodologías, con las que se procuró maximizar la homogeneidad interna dentro de las regiones y la heterogeneidad entre distintas regiones, para poderlas considerar diferentes y por lo tanto, unidades analíticas distintas e individuales. Para tal efecto, se utilizaron datos de flujos de traslado al trabajo entre municipios para analizar la red formada por ellos, con la finalidad de conocer cómo se distribuyen las relaciones entre ellos y para cuantificar qué papel juega cada municipio dentro de la red. Una vez resumida esta relación en una matriz de viajes origen-destino, que muestra el número de viajeros de cada uno de los municipios hasta todos los demás, obtenidos del Censo de Población 2001 y 2011 del Instituto Nacional de Estadística (INE) de España, el objetivo del análisis consistió en identificar cuáles son los polos de

atracción y a partir de éstos generar áreas de influencia utilizando algoritmos que minimizen la autocontención de las áreas generadas, así como un continuo geográfico.

La metodología elegida para el análisis comparativo fue el análisis de cluster jerárquico, con lo que a partir de la matriz de viajes origen-destino, se pudieron formar grupos o regiones homogéneas en base a la cantidad de personas que comparten destinos en su viaje al trabajo y lugares de residencia. Estas regiones se integraron por grupos de municipios, considerando como criterio de similaridad entre municipios los destinos (trabajo) y orígenes (residencia) que las personas tienen en su vida diaria, es decir se consideró que dos municipios son similares si las personas que viven en esos dos municipios tienen destinos similares en su viaje al trabajo. En concreto, se consideraron siete métodos de aglomeración para cluster jeráquico: completo, simple, mediana, centroide, Ward, vinculación intergrupos y vinculación intra-grupos)

El método de cluster jerárquico es idóneo para determinar el número óptimo de conglomerados así como el contenido de los mismos y comienza con el cálculo de la matriz de distancias entre cada uno de los elementos (casos o variables); a continuación, en esta matriz se buscan los dos elementos más similares y se agrupan en un conglomerado, el cual es indivisible a partir de ese momento, así se van agrupando los elementos en conglomerados cada vez más grandes y heterogéneos hasta llegar al último paso en el que todos los elementos están agrupados en un único conglomerado global.

La versatilidad del análisis de conglomerados jerárquicos radica en la posibilidad de utilizar distintos tipos de medidas para estimar la similaridad (o distancia) existente entre los casos o variables, y también la posibilidad de transformar la métrica original de las variables, así como la posibilidad de seleccionar entre una gran variedad de métodos de aglomeración (enlace simple, enlace completo, método de vinculación intergrupos, método de vinculación intra-grupos, método de pérdida de inercia mínima o método de aglomeración de Ward, método de aglomeración de centroides y método de agrupación de medianas), los cuales calculan de diferente manera la distancia o similaridad entre dos conglomerados (agrupaciones de dos o más casos), una vez que se han fusionado dos o más casos.

Una vez resumida la información de cuantos viajes al trabajo se hacen entre municipios de España, en una matriz origen-destino y en donde se especifica el conteo de personas que viajan desde y hasta cada municipio, se realizó la transformación de los datos como paso previo a aplicar el análisis de cluster jerárquico, teniendo en cuenta que para obtener

resultados más ajustados y confiables es preciso que los datos no contengan *outliers*, que la distribución de los datos se asemeje lo más posible a una distribución normal o cuando menos que no presenten sesgo y que los datos con los que se evaluará la similaridad o distancia entre casos, en este trabajo, municipios, se encuentren medidos en la misma escala.

La diagonal principal de la matriz origen-destino fue suprimida, ya que en ella se resume la cantidad de personas que no salen del municipio en el que habitan para trabajar, las que por su gran cantidad representarían datos extremos y aportarían poca información para la formación de regiones funcionales.

Posteriormente con el fin de simetrizar la distribución, toda la matriz se escaló en una unidad antes de transformar los datos por la inversa negativa al cuadrado $(-1 \,/\, (1+x)^2)$, Vidal-Díaz (2003), y por último se transformaron los datos a puntuación z, con la idea de tener las variables expresadas en la misma unidad de medida y poder comparar municipios con población menor con los municipios que tienen mayor densidad de población y por lo tanto mayor número de viajes (grandes ciudades generan y atraen más interacción que las pequeñas).

La medida escogida para cuantificar la similaridad entre municipios fue el coseno del ángulo formado por los vectores de las variables en el espacio dimensional de los casos (Mongay 2005), es decir si los destinos de dos municipios se encuentran perfectamente correlacionados, los vectores de los dos destinos dibujados en el plano cartesiano de los municipios-origen, tendrán la misma dirección, por lo que el coseno del ángulo formado por los dos vectores será la unidad; así los destinos de dos municipios estarán más correlacionados cuanto más se aproxime a la unidad el coseno o coeficiente de similaridad. Se utilizó este coeficiente de similaridad debido a que se quieren capturar municipios similares y agruparlos en función de su forma o destinos en común y no de niveles, es decir no medir la similaridad en función de la cantidad de personas que se mueven. La semejanza se determina por la forma de los componentes y no por el tamaño de los componentes en la estructura de similaridad, (Fisher, 1980).

Para la construcción de los clusters, se utilizaron los métodos de aglomeración de vinculación simple, vinculación completa, vinculación de promedio del grupo, vinculación de promedio Intra-grupos, pérdida de inercia mínima o método de aglomeración de Ward, centroides y mediana.

Para decidir cuál de las estrategias de vinculación o métodos de aglomeración es el más adecuado, se utilizó la matriz cofenética, (Mongay, 2005), en la que se compararon los coeficientes de similaridad de cada uno de los posibles pares de municipios contra las distancias entre todos los pares de municipios obtenidas de la matriz original; los coeficientes de similaridad para cada método se obtienen del historial de aglomeración, el cual tiene $N-1$ etapas, de cada uno de los métodos de aglomeración se compararon con los coeficientes de similaridad expresados en la matriz de similaridad (o distancia) inicial producida al inicio del procedimiento del cluster.

Para determinar el método que mejor se ajusta a los datos originales se midió el coeficiente de correlación, en este estudio, el método de vinculación de promedio intra-grupos mostró la mayor correlación, $r = 0.84$, con las distancias de los datos originales.

Al trazar regiones funcionales es preciso llevar a cabo "*robustness checks*", con diferentes conformaciones para evaluar cuál de las opciones forma agrupaciones lo más homogéneas posible de manera interna y lo más diferente posible entre distintas regiones. En este trabajo se utilizó el análisis de varianza de un factor para evaluar las regiones.

Basándose en la idea fundamental del análisis de cluster, la cual es formar grupos de tal manera que los elementos en un grupo compartan un mismo perfil, mientras que los elementos en otro tengan un perfil totalmente distinto, es posible verificar si existen diferencias significativas entre los conglomerados obtenidos. Para esto, es posible utilizar la variable que contiene la información sobre el conglomerado al que pertenece cada sujeto y como variables dependientes cada una de las variables incluidas en el análisis que en este caso son los conteos de personas que tenían como destino cada municipio. El análisis de varianza de un factor permitió valorar si los conglomerados son diferentes entre sí y qué variables contribuyen a hacerlos diferentes.

3.9. Evidencia empírica

3.9.1. Clusters por género

Los clusters de mayor tamaño en el estrato compuesto unicamente por Mujeres en el año 2001, es el formado por los municipios Barakaldo, Getxo-Leioa, Santurtzi, Sestao y Erandio, Figura 7, ubicados en la provincia Bizkaia. Muestra una región funcional contigua sin necesidad

de realizar ningun ajuste alguno, además es compacto en lo que respecta a distancia en kilómetros y es una región importante en fuentes de empleo en el año señalado.

Cabe mencionar que los patrones de desplazamiento de las mujeres son muy diferentes de los de los hombres; los patrones de desplazamiento para mujeres muestran que sus traslados son de menor distancia en comparación a los desplazamientos de los hombres, situación que puede ser explicada debido a la mayor tendencia que tienen las mujeres a ocupar trabajos de medio tiempo, lo que disminuye el incentivo para trasladarse a grandes distancias, mientras que para hombres los conglomerados tienden a ser mayores, lo que debe ser considerado para investigaciones futuras.

Figura 7. Cluster Bizkaia Mujeres 2001

Los municipios que pierden su lugar en un cluster y se aislan, probablemente no sean generadores de empleos; el estado de aislamiento puede ser un indicio de pérdida de atracción y generación de fuentes de empleo.

El cluster, Figura 8, formado por los municipios Icod de los Vinos, San Cristóbal de La Laguna, Realejos, Santa Cruz de Tenerife y Tacoronte, es el que ocupa el segundo lugar de mayor tamaño para Mujeres en el año 2001.

Figura 8. Cluster de la provincia S.C. Tenerife
Mujeres 2001

La conformación de los clusters para Mujeres en el año de 2001 en todo el país se muestra en la Figura 9.

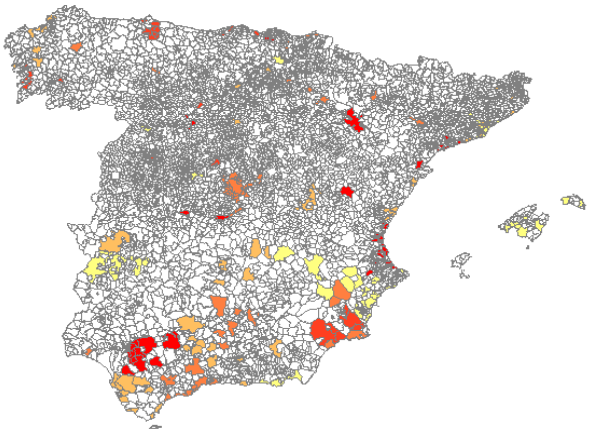

Figura 9. Clusters de España Mujeres 2001

El cluster de mayor tamaño para el estrato conformado por los Hombres en el año 2001, Figura 10, está formado por los municipios Barakaldo, Basauri y Santurtzi, cuyo trayecto más largo es de 21 Km. en la provincia de Bizkaia. Este cluster tiene la particularidad de estar delimitado por la Ría de Bilbao, información que se encuentra en la variable analizada en este trabajo y reflejada por la metodología.

Figura 10. Cluster de la provincia Biskaia
Hombres 2001

El resto de clusters para Hombres no presenta gran variación de tamaño. El mapa para la conformación de Hombres en el 2001 se puede ver en la Figura 11, que muestra una conformación para 294 clusters.

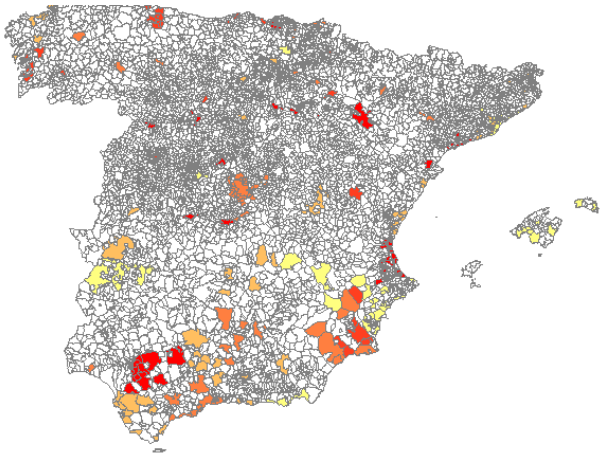

Figura 11. Clusters de España Hombres 2001

La provincia de Álava muestra los estratos de los dos años analizados, solamente un cluster formado por un municipio Vitroria-Gateiz, indicando que fue un generador y receptor de empleos por sí solo para los nueve estratos; es el cluster que tiene la mínima cantidad de municipios en todos los años y que no aparece integrado en ningún otro cluster.

3.9.2. Cluster de Manufactura

En la Tabla 13 y en la Figura 12, se muestra que el mayor cluster en el año 2011, en el estrato Manufactura, es el formado por veinte municipios: Badalona, Barcelona, Castelldefels, Cornellà de Llobregat, Esplugues de Llobregat, Gavà, Hospitalet de Llobregat, Molins de Rei, Prat de Llobregat, Rubí, Sabadell, Sant Adrià de Besòs, Sant Cugat del Vallès, Sant Feliu de Llobregat, Sant Joan Despí, Santa Coloma de Gramenet, Sant Vicenç dels Horts, Cerdanyola del Vallès, Terrassa y Viladecans. La mayor distancia en kilómetros en este cluster se registra entre los municipios Castelldefels y Terrassa; es el mayor cluster de todos los estratos en ambos años analizados y respeta el supuesto de contigüidad sin necesidad de hacer ajustes posteriores a la aplicación de la metodología.

Tabla 13. Cantidad de clusters por estrato y clusters de mayor tamaño 2001.

Género:	Número de Clusters	Municipios Agrupados	Cluster Mayor
Mujeres	253	311	6
Hombres	294	316	3
Industria:			
Agricultura	107	191	5
Manufactura	262	293	3
Construcción	190	302	8
Ocupación:			
CNO 10-38	146	250	5
CNO 40-59	170	307	11
CNO 60-78	236	305	7
CNO 80-97	146	299	12

Nota: Ocupaciones 10-38 Administradores Profesionales y Técnicos, Ocupaciones 40-59 Supervisores y Trabajadores no Manuales, Ocupaciones 60-78 Trabajadores Manuales Calificados, Ocupaciones 80-97 Trabajadores Manuales no Calificados.Elaboración Propia

Figura 12. Cluster de la provincia de Barcelona Manufactura 2011

3.9.3. Cluster de Construcción.

Como se muestra en la Tabla 14 y en la Figura 13, el cluster mayor en el estrato de Construcción en el año 2001, está formado por los municipios: Alcalá de Henares, Alcorcón, Coslada, Fuenlabrada, Getafe, Leganés, Móstoles y Parla, todos ellos en la provincia Madrid; en el año 2011, se reduce a seis clusters. En el año de 2011 el cluster con mayor cantidad de municipios está formado por: Agüimes, Ingenio, Palmas de Gran Canaria, San Bartolomé de

Tirajana, Santa Lucía de Tirajana y Telde, ubicados en la provincia de Las Palmas. Este cluster también resulta contiguo.

Figura 13. Cluster de la provincia de Construcción Madrid 2011

Tabla 14. Cantidad de Clusters por estrato y clusters de mayor tamaño 2011.

Género:	Número de Clusters	Municipios Agrupados	Cluster Mayor
Mujeres	381	394	3
Hombres	292	394	12
Industria:			
Agricultura	145	312	9
Manufactura	182	384	20
Construcción	223	388	6
Ocupación:			
CNO 10-38	190	394	19
CNO 40-59	184	394	6
CNO 60-78	219	393	9
CNO 80-97	157	387	13

Nota: Ocupaciones 10-38 Administradores Profesionales y Técnicos, Ocupaciones 40-59 Supervisores y Trabajadores no Manuales, Ocupaciones 60-78 Trabajadores Manuales Calificados, Ocupaciones 80-97 Trabajadores Manuales no Calificados.Elaboración Propia

3.9.4. Clusters de Trabajadores manuales no calificados

El cluster formado con la mayor cantidad de municipios en el estrato de Trabajadores Manuales no Calificados del año 2001, está formado por los municipios siguientes: Cornellà de Llobregat, Esplugues de Llobregat, Hospitalet de Llobregat, Molins de Rei, Prat de Llobregat, Sant Andreu de la Barca, Sant Boi de Llobregat, Sant Feliu de Llobregat, Sant Joan Despí, Santa Coloma de Gramenet, Sant Vicenç dels Horts y Viladecans, todos ellos ubicados en la periferia de Barcelona, Figura 14.

Figura 14. Cluster de la provincia de Barcelona
Trabajadores manuales no calificados 2001

3.9.5. Cluster de Administradores, Profesionales y Técnicos.

En el caso del estrato Administradores, Profesionales y Técnicos de 2011 el cluster compuesto con la mayor cantidad de municipios está formado por los siguientes: Alcorcón, Aranjuez, Arroyomolinos, Boadilla del Monte, Fuenlabrada, Galapagar, Getafe, Leganés, Majadahonda, Móstoles, Navalcarnero, Parla, Pinto, Pozuelo de Alarcón, Rozas de Madrid, Torrelodones, Valdemoro, Villaviciosa de Odón e Illescas, ubicados al oeste de Madrid, Figura 15. Este cluster a pesar de ser el segundo más grande de todos los estratos en ambos años, resulta contiguo sin necesidad de hacer ningún ajuste posterior a la aplicación de la metodología empleada.

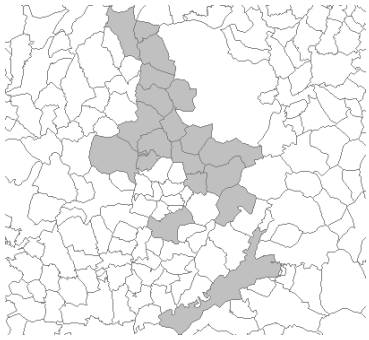

Figura 15. Cluster de la provincia de Madrid
Administradores, Profesional y Técnicos 2011

3.9.6. Cluster de Agricultura

La conformación de Agricultura en 2011 para España, se muestra en la Figura 16, con 145 clusters en total; se agruparon 312 municipios y el cluster de mayor tamaño está formado por nueve municipios que son también contiguos.

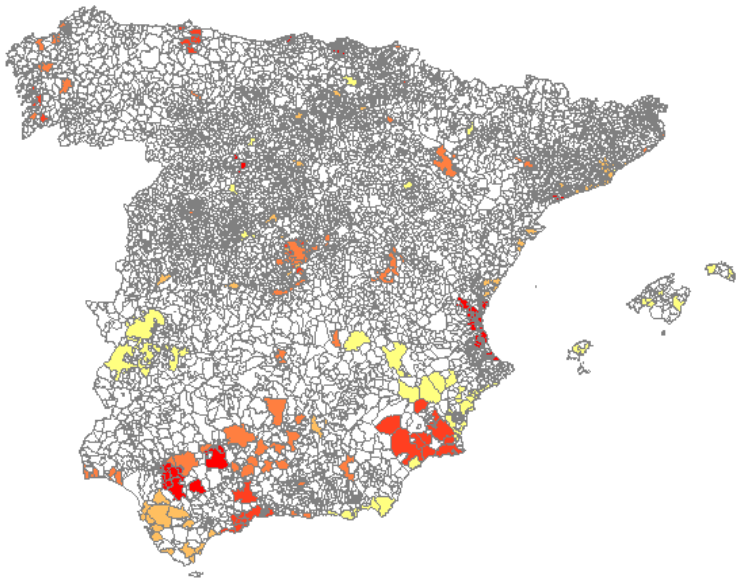

Figura 16. Clusters de España
Agricultura 2011

3.10. Mercados Locales Laborables

Para medir y comparar la movilidad y los procesos de articulación y expansión de los mercados laborales locales (MLL), se analizaron cuatro variables para los años 2001-2011: autonomía de la oferta, autonomía de la demanda, empleo y cantidad de trabajadores. El primer índice es el porcentaje de trabajadores residentes en un MLL que trabajan dentro de los límites del mismo, mientras que el segundo es el porcentaje de los puestos de trabajo disponibles en el MLL que están ocupados por los trabajadores que residen en el mismo MLL. En ambos casos, a medida que el índice es más alto, el MLL será más cerrado y la interacción funcional con otro MLL es menor.

En 2001, los MLL formados por mujeres y trabajadoras manuales calificadas muestran índices de autonomía de oferta que son comparables a los índices de autonomía de la demanda o

incluso mayores, lo que indica oferta de puestos de trabajo y capacidad para atraer a los trabajadoras de otras regiones, situación que se invierte en 2011.

En 2011, los índices de oferta son más bajos que los índices de demanda en los MLL formados por trabajadoras, lo que señala zonas deficitarias en el empleo y con predominio de salidas. Esta situación se puede observar en los MLL formados por trabajadores masculinos manuales calificados. Una situación inversa se da en el MLL formado por los trabajadores agrícolas donde los resultados son áreas muy excedentes y con capacidad de atracción de otras regiones en 2011 y zonas deficitarias y con predominio de mayor salida en 2001, lo que indicó una recuperación de los empleos agrícolas en las tierras del interior, tal como lo muestran los flujos de inmigración.

Figura 17. Comportamiento del sector de mujeres en España 2001-2011

Los MLL formados por trabajadores empleados en la construcción, trabajadores manuales no calificados y los segmentos de manufactura muestran índices de autonomía que son representativos de los mercados deficitarios en ambos años. Aunque el déficit aumenta en 2011, y se captura mediante una diferencia más amplia entre los índices de autonomía de 2001 y 2011, la autonomía de la oferta es menor que la autonomía de la demanda, lo que indica el déficit de puestos de trabajo predominando las salidas sobre las entradas.

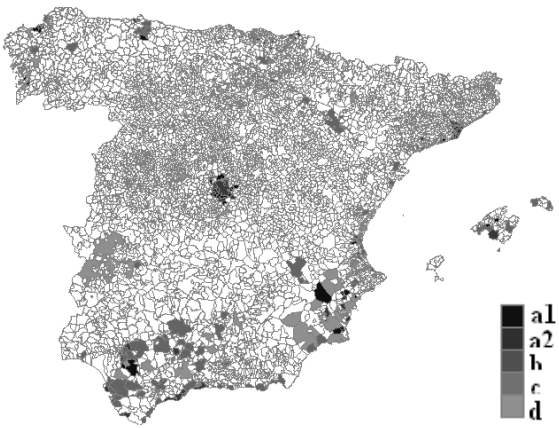

Figura 18. Comportamiento del sector de trabajadores manuales calificados en España 2001-2011

Los niveles de autonomía de la oferta y de la demanda de los mercados de trabajo locales formados únicamente por mujeres trabajadoras son más altos en comparación a los mostrados por los trabajadores de sexo masculino en ambos años, indicando que el MLL formado por las mujeres trabajadoras es más cerrado y que la interacción funcional con otro MLL es menor para las mujeres que para los hombres. Esto indica que las trabajadoras tienden a residir dentro del mismo MLL en mayor porcentaje que los trabajadores varones. En ambos años se puede observar que el segmento con la mayor tendencia a desplazarse es el de profesionales administrativos y técnicos seguidos por el de hombres. El segmento con la menor propensión a desplazarse es el segmento femenino, seguido del de la construcción.

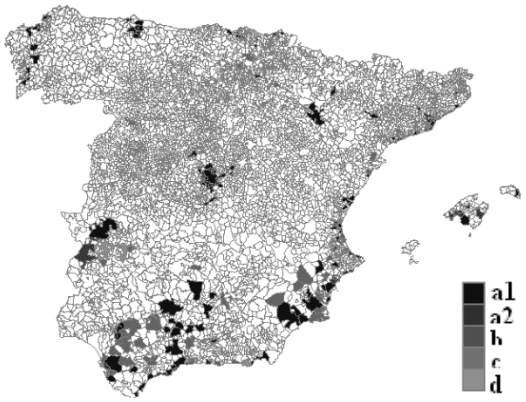

Figura 19. Comportamiento en España del sector de trabajadores no calificados 2001-2011

En general, hay un aumento en la movilidad como consecuencia de la disminución de la autonomía entre 2001 y 2011, lo que se muestra en la disminución de los índices de oferta y demanda para los segmentos mujer, hombre, agricultura y construcción. Los segmentos de manufactura, trabajadores manuales no calificados y los profesionales y técnicos administrativos muestran una disminución en la movilidad.

Al analizar los resultados del estudio del MLL se observa que hay un incremento en los flujos intermunicipales en todo el territorio que apuntan a un aumento global en la movilidad, mientras que los índices de autonomía muestran una reducción en la concentración de los flujos, lo que significa que la interacción entre las regiones urbanas aumenta.

Junto a esta tendencia, se realizó una clasificación del MLL teniendo en cuenta el cambio en las siguientes variables: Empleo, cantidad de trabajadores, y la autonomía de la oferta y la

demanda, para ambos años. Esta clasificación es similar al trabajo de Salom y Casado (2007) e identifica cuatro comportamientos diferentes del MLL entre 2011 y 2001:

A. El MLL que "reduce sus índices de autonomía de demanda y de oferta" en un contexto de fuerte crecimiento del empleo y de trabajadores como es el caso de los niveles de autonomía en 2001, que por lo general son relativamente bajos, con los dos siguientes matices:

   A.1. El MLL con "pequeños índices de autonomía", indica la existencia de flujos cruzados y una tendencia a la interrelación funcional con las zonas vecinas. En el primer caso, es más representativo de un MLL situado en un entorno urbano o metropolitano que tiene vínculos cercanos de tipo no-polarizado con el MLL circundante. El ejemplo más claro en el segmento femenino son: Getafe, Leganés y Pinto, donde tres MLL en 2001 y en 2011 se fusionan para convertirse en un mercado único.

   A.2. El MLL, donde los índices de autonomía de la demanda son más bajos que los índices de la autonomía de la oferta, señala la existencia de una importante oferta de puestos de trabajo y la capacidad para atraer flujos de trabajadores de las zonas vecinas, ya que están cerca de las zonas más dinámicas en cuanto al crecimiento del empleo (tipo A), se convierten en parte de ellos y desaparecen como mercados autónomos entre 2001 y 2011. El caso de San Cristóbal de la Laguna y Tacoronte que se convierten en un MLL junto con Orotava en 2011, es el ejemplo más claro se encuentra en el segmento femenino.

B. El MLL que reduce significativamente sus niveles de autonomía en un contexto de estancamiento o recesión del número de puestos de trabajo y de los trabajadores, es decir que los niveles de autonomía de la oferta son inferiores a la autonomía de la demanda, se identifican como áreas con déficit de puestos de trabajo y un predominio de salidas. El caso del segmento de mujeres en Petrer en Alicante en 2011, muestra este comportamiento.

C. El MLL que apenas reduce sus niveles de autonomía a pesar de experimentar pérdidas de empleo y reducción del número de trabajadores, empieza a altos niveles de autonomía y por estar lejos de un MLL más dinámico mantiene la misma estructura. El caso más claro es el segmento de la construcción en Vitoria-Gasteiz en el País Vasco.

D. El MLL que apenas modifica sus niveles de autonomía aunque experimenta un alto crecimiento de los puestos de trabajo y trabajadores, parte de fuertes niveles de autonomía. Este tipo de MLL a pesar de modificarse ligeramente mantiene su patrón articuladamente bajo, relacionado con la abundancia de las ciudades de tamaño medio. Cuando este MLL inicia de muy altos niveles de autonomía a pesar de que se reducen, mantienen un modelo territorial no articulado, relacionado con la abundancia de las ciudades de tamaño medio. Los casos más representativos son el MLL del segmento de los Profesionales Administrativos y Técnicos de Palma, Algeciras, Barcelona y Zaragoza.

3.11. Conclusiones

Se realizó una investigación orientada a la formación de grupos de unidades geográficas determinados midiendo los vínculos laborales compartidos, los cuales quedan resumidos en los viajes diarios, residencia-trabajo, realizados por las personas que viven y trabajan en esas unidades geográficas, así dichos grupos pueden ser considerados como regiones funcionales (MLL).

Para formar las regiones funcionales existe una gran variedad de metodologías, cada una de ellas con matices que las hacen proceder de diferente forma y por lo tanto conducen a conformaciones de agrupaciones distintas.

Se aplicó un Análisis de Conglomerados Jerárquicos el cual inicia por comparar con índices de similaridad a todos los elementos a ser clasificados, posteriormente los agrupa iniciando por los más parecidos y dejando al final los menos semejantes, de ahí su nombre de jerárquicos, la similaridad de todos los elementos agrupados en cada etapa queda expresada en una gráfica llamada dendrograma donde se pueden observar cuáles grupos se forman de manera natural por la similaridad entre sus elementos y por cuáles elementos está formado cada grupo. Para asegurar que los grupos observados en el dendrograma sean óptimos, es decir los grupos sean internamente homogéneos y externamente heterogéneos, se evaluaron multiples conformaciones de agrupaciones distintas utilizando el procedimiento de ANOVA de un factor escogiendo la conformación que producían los factores con la mayor diferencia entre sí.

Se encontró que en términos de replicabilidad el análisis de cluster jerárquico facilita la comparación en la morfología de las regiones funcionales con el paso del tiempo, debido a la

facilidad de uso y aplicación y al no estar limitado en términos de normalidad y homocedasticidad que presenten las variables con las que se pretenden caracterizar a los elementos; además el análisis de cluster no es una técnica inferencial sino algorítmica por lo que no es necesario que las variables tengan relaciones lineales. La metodología elegida limita el uso de decisiones arbitrarias y subjetivas y ayuda a que las regiones o clusters estén optimizadas en función de la información bajo estudio, que en este caso son los vínculos sociales, económicos y laborales existentes entre municipios. La técnica utilizada también produce una regionalización coherente, fiel a la realidad y fácil de interpretar sin la necesidad de contar con una matriz de contigüidad. Sin embargo, el cluster jerárquico presenta la desventaja de que no es posible añadir restricciones explícitas a las zonas que son producidas, además, en el caso de regiones geográficas la contigüidad tiene que ser revisada de manera posterior al procedimiento.

El estudio se realizó con datos del censo de población de España y se comparó la morfología de los mercados laborales en los años 2001 y 2011. La muestra se separó en nueve estratos relacionados con variables asociadas a las personas censadas como: Género (Hombres, Mujeres), Ocupación (Trabajadores Manuales Calificados, Trabajadores Manuales No Calificados, Supervisores, Profesionales) y Sector Industrial (Manufactura, Construcción, Agricultura). Los Municipios fueron utilizados como unidades geográficas para la formación de los mercados laborales (MLL), se utilizó la variable residencia-trabajo como medida de vinculación laboral entre municipios. Se evaluaron múltiples configuraciones de agrupaciones y para este estudio, el método de vinculación de promedio Intra-Grupos mostró la mayor correlación con las distancias de los datos originales.

Para cada mercado laboral local (MLL), se determinaron cuatro variables para los años 2001-2011: autonomía de la oferta, autonomía de la demanda, empleo y cantidad de trabajadores; se compararon, encontrando que en general, hay un aumento en la movilidad como consecuencia de la disminución de la autonomía entre 2001 y 2011, lo que se muestra en la disminución de los índices de oferta y de demanda para los segmentos Mujer, Hombre, Agricultura y Construcción.

Los segmentos de Manufactura, Trabajadores Manuales No Calificados y los Profesionales y Técnicos Administrativos muestran una disminución en la movilidad. Hay un incremento en los flujos intermunicipales en todo el territorio que apuntan a un aumento global en la movilidad.

Los índices de autonomía muestran una reducción en la concentración de los flujos, lo que significa que la interacción entre las regiones urbanas aumenta. Junto a esta tendencia, se realizó una clasificación del MLL, teniendo en cuenta el cambio en las variables: Empleo, cantidad de trabajadores, y la autonomía de la oferta y la autonomía de la demanda, para ambos años, identificando cuatro comportamientos diferentes del MLL entre 2011 y 2001, congruente con la clasificación señalada en el trabajo de Salom y Casado (2007).

Se encontraron MLL con pequeños índices de autonomía que se fusionan para convertirse en un mercado único, por ejemplo, en el segmento femenino: Getafe, Leganés y Pinto, así como MLL donde los índices de autonomía de la demanda son más bajos que los índices de la autonomía de la oferta y absorben mercados vecinos, como es el caso de San Cristóbal de la Laguna y Tacoronte, que se convierten en un MLL junto con Orotava en el segmento femenino y MLL donde los niveles de autonomía de la oferta son inferiores a la autonomía de la demanda, que se identifican como áreas con déficit de puestos de trabajo y un predominio de salidas.

También se encontraron MLL que apenas reducen sus niveles de autonomía a pesar de experimentar pérdidas de empleo y reducción del número de trabajadores como en el segmento de la construcción en Vitoria-Gasteiz en el País Vasco, así como MLL que apenas modifican sus niveles de autonomía a pesar de experimentar un fuerte crecimiento de los puestos de trabajo y de los trabajadores tal es el caso del segmento de los Profesionales, Administrativos y Técnicos de Palma, Algeciras, Barcelona y Zaragoza.

En resumen regionalizar, que en el presente trabajo consistió en organizar y subdividir de manera óptima el espacio geográfico del territorio Español con el fin de obtener los mercados laborales locales para diferentes segmentos, permite tener una idea consistente del comportamiento de los mismos y es un instrumento útil para la planificación de políticas que permitan alcanzar grados de desarrollo, sino igualitarios cuando menos a niveles aceptables en todas las regiones del país.

4. Evaluación del desarrollo económico medido con variables asociadas a educación en zonas comerciales derivadas de economías de aglomeración y de ubicación.

4.1. Introducción

Debido a que en cualquier proceso de producción el tiempo y los recursos por lo general son limitados, es preciso identificar segmentos focales en los cuales se han de priorizar y enfocar dichos recursos. Se intuye que es posible obtener mayores rendimientos al dirigir los recursos en determinadas áreas o mercados específicos, mientras que si se aplican en otros incluso podría ser contraproducente, por lo que varios autores se han enfrentado a la identificación de áreas de oportunidad y la justifican con un conjunto de hipótesis teóricas entre las que destacan: optimizar los recursos disponibles con el fin de sacar mayor provecho de ellos (Mulhern, 1994), elevar la rentabilidad (Meadows 1998), mejorar la ventaja competitiva (Mulhern, 1994), direccionar los recursos disponibles hacia grupos de consumidores más lucrativos (Meadows 1998), ayudar a asignar los esfuerzos hacia grupos específicos de clientes potenciales (Bass 1968), direccionar los recursos con iniciativas específicas para cada región o segmento según el grado de problemática (Govind 2014), fortalecer la posición en el mercado (Haley 1968), proveer conocimientos adicionales del mercado (Bloom 2005), ayudar a la planeación de estrategias (Sarabia 2012) y proveer una guía para profesionales de la salud para combatir el consumo de productos hedónicos dañinos (cigarro, drogas, alcohol, ludopatía, etc.) (Govind 2014)

En todos estos trabajos se realiza una segmentación en regiones (Cui 2000), grupos de personas (Sarabia, 2012), o climas determinados (Govind 2014), con la intención de clasificar en grupos mutuamente excluyentes a un cúmulo de elementos que aparentemente son iguales. Una vez realizada esta clasificación los autores se enfocan en el o los grupos que sean de su interés o en los que resulte más conveniente orientar los recursos y tiempo, lo que implica aceptar que distintos segmentos tendrán respuestas diferentes al mismo estímulo. Este concepto aplicado en economía geográfica implica que los mismos estímulos pueden tener efectos distintos en regiones distintas.

Partridge (2008), reconoce este fenómeno en una investigación enfocada a identificar los factores regionales que detonan el crecimiento del empleo y sugiere que existe heterogeneidad espacial en los diferentes lugares que incluye en su muestra. Sugiere además que dicha heterogeneidad puede surgir debido a que los mercados laborales locales varían en

su estructura, contexto social e historia (Lloyd and Shuttleworth, 2005) y otras características que son particulares de cada mercado laboral, las cuales, es casi imposible que se repitan en dos o más regiones.

Mathur (1999), también reconoce el fenómeno de la heterogeneidad especial y debido a este fenómeno recomienda estrategias de desarrollo económico que no sean uniformes a lo largo de todo el territorio.

De acuerdo al trabajo de Partridge (2008), la estimación de la heterogeneidad espacial puede generar nuevas hipótesis a evaluar. Por ejemplo, en su trabajo Govind (2014), plantea la hipótesis de que la variación regional de las tasas de consumo de alcohol *per cápita* en San Francisco puede deberse a la variación de la presencia de personas sin hogar de cada región. Acemoglu (2001), por su parte plantea que los censos muestran una gran relación positiva entre educación promedio y los salarios individuales, lo que quizás sea más prominente en algunas regiones que en otras.

Esto pone en evidencia que es necesario profundizar a fin de descubrir la verdadera naturaleza de las relaciones a nivel local, lo que contribuiría al desarrollo exitoso de políticas económicas locales, y para esto se requiere del conocimiento de los procesos socioeconómicos y de las dinámicas locales, Partridge (2008).

A partir de la revisión de la literatura sobre el tema, es necesario reconocer la heterogeneidad espacial en las características de una comunidad o puntos de observaciones y entonces aplicar un modelado estadístico *ad hoc* que tome en cuenta la no estacionariedad espacial que puede verse enmascarada al utilizar un método estándar como los mínimos cuadrados ordinarios.

Desde hace algunos años se ha venido aplicando exitosamente la regresión geográficamente ponderada (GWR) a procesos que varían en el espacio. Los resultados de la GWR es un conjunto de estimaciones de los parámetros locales para cada relación que se puedan asignar para producir una superficie de parámetros a través de la región de estudio. De esta manera la GWR ofrece una valiosa información sobre la naturaleza de los procesos que se están investigando y reemplaza las formas globales tradicionales de los modelos de regresión.

El objetivo de la presente investigación es modelar el comportamiento del crecimiento del empleo a través del territorio español, en función de variables relacionadas con el nivel de educación con el fin de detectar áreas de oportunidad.

El crecimiento de empleo se estudia en relación a las siguientes variables explicativas: razón de empleados: administradores, profesionales y técnicos por municipio, razón de empleados: trabajadores manuales no calificados por municipio, diferencia entre la razón de empleados hombres profesionales activa y la razón de mujeres profesionales empleadas con respecto a la población económicamente activa, razón de trabajadores empleados: supervisores no manuales por municipio, razón de trabajadores empleados: trabajadores manuales calificados por municipio, razón de trabajadores en empleos relacionados con manufactura por municipio, razón de trabajadores en empleos relacionados con la pesca, acuicultura y actividades de los servicios relacionados por municipio, razón de la población con educación de tercer nivel (diplomatura, licenciatura y doctorado) por municipio, diferencia entre la razón de hombres ocupados con respecto a la población económicamente activa y la razón de mujeres ocupadas con respecto a la población económicamente activa y, razón de trabajadores con empleo en el sector de la construcción por municipio.

4.2. Revisión de la literatura

Existe evidencia empírica que sostiene que el proceso de consumo, y por lo tanto el productivo, son afectados por una gran variedad de factores que actúan a nivel regional, al grado de que Govind (2014), argumentó que el precio de alcohol en San Francisco varía inversamente con la temperatura y directamente con la precipitación pluvial y los cielos nublados. Además Govind (2014), sugirió que este fenómeno es más prominente en algunas partes de la ciudad que en otras. Con base en esta observación, investigó como las condiciones climáticas afectan las tasas de consumo de productos o actividades hedónicas negativas (tabaco, ludopatía, drogas y otras.) e identificó los lugares donde el clima influye con mayor predominio al consumo del tabaco. Es decir localizó distritos en donde existen excepciones focalizadas o desviaciones de las tendencias globales (Grose, 2006), y alude a la diferencia de porcentajes de personas sin hogar en diferentes distritos para explicar este fenómeno.

Lo anterior sugiere la utilidad de considerar las características únicas de las regiones, puesto que su estructura económica, social, política, histórica y otras, son elementos influyentes que

determinan las actividades productivas y de consumo que se desarrollan en cada zona. Entre dichas influencias Hawkins et. al. (1980), mencionó el paisaje físico (topografía, clima, recursos naturales) y el psicológico (la historia, estructura económica, religiosa, legal, poblacional) como condiciones a ser consideradas. En este mismo sentido Kahle (1986), sugirió el clima y los recursos compartidos como fuerzas de cohesión o elementos a nivel regional que empujan a las inmediaciones a estar juntas y Cui (2000), mencionó la diversidad geográfica y la disparidad económica como factores que producen diferencias significativas entre vecindades, resaltando el hecho de que una estrategia única a lo largo de todas la circunscripciones no parece aconsejable (Govind, 2014).

Además de las diferencias topográficas, climáticas y económicas, Hawkins (1980), mencionó las diferencias culturales predominantes a nivel local y Sarabia (2012), ejemplificó estas diferencias al reconocer la existencia de un grupo cultural significativo de consumidores escandinavos el cual, considera aspectos ecológicos que rodean a las actividades de la empresa; según Sarabia (2012), dicho grupo se caracteriza por ser más consciente de campañas con preocupaciones ecológicas, procesos de producción, uso de materiales, accesorios de tienda, aspectos experimentales amigables con el medioambiente y relacionados con preocupaciones ambientales y también mencionó que es posible motivar la intención de compra de este grupo resaltando las actitudes ecológicas de la empresa. Los aspectos culturales también fueron mencionados por Cui (2000), quien alude a la herencia cultural como una condición que hace a cada mercado único. En este caso Chaudhuri (2005), demostró que residentes en las áreas con más apego cultural hacia valores tradicionales son menos propensos a hacer de la adquisición un objetivo, afectando a nivel regional el grado de consumo, por ende el nivel de producción; así los planificadores deberían considerar subculturas geográficas como una variable potencialmente útil cuando se desarrollan estrategias de mercado (Hawkins 1980), o de producción.

Beane (1987), explicó que este tipo de segmentación implica reconocer que las personas y sus necesidades varían geográficamente y que el término “geográficamente” puede tomar varios significados (país, densidad de población o clima). En un trabajo anterior al de Beane (1987), realizado por Hawkins (1980), justificó este enfoque argumentando que en cada región existen subculturas y que los miembros para pertenecer a una subcultura deben de compartir ciertos patrones (clima, legislación, religión y otras), ya que eventualmente el proceso de consumo y productivo es afectado, por lo que las influencias a nivel regional se convierten en un elemento de interés.

Hawkins (1980), argumentó que podría haber una gran variación entre regiones y señala que una sola planeación con cobertura nacional podría no ser tan efectivo como una segmentada geográficamente de acuerdo a la subcultura predominante en cada región.

Govind (2014), recomendó una estrategia diferenciada por regiones y argumentó que entender las variaciones regionales en el consumo proveerá mayor conocimiento sobre los consumidores, los cuales serán útiles para la planeación de políticas públicas hechas a la medida para cada región.

En esta misma línea de investigación se encuentra el trabajo de Cui (2000), quien encontró y enumeró diferencias estadísticamente significativas en el poder adquisitivo, actitudes, estilos de vida, uso de medios, y patrones de consumo, de las personas pertenecientes a cuatro regiones en China; argumentó que no hay que pasar por alto la diversidad entre las personas autóctonas de cada región.

Por su parte Mulhern (1994), también realizó un estudio de segmentación geográfica; buscó el aislamiento de una región de manera que esté habitada mayoritariamente por el segmento bajo su investigación y posteriormente verificó evidencia estadística que explique el comportamiento de compra de ese segmento del mercado en particular.

En un trabajo posterior, Chaudhuri (2005), también encontró diferencias entre las personas que conformaban su muestra de acuerdo a la región a la que pertenecían, lo que según el desafía las suposiciones comunes de que todos los consumidores son similares.

En la actualidad, la segmentación ocupa un lugar vital en el análisis cuantitativo de mercados, al grado que Simkin (2008) se refierió a ella como esencial para una estrategia de negocios efectiva, en la cual es preciso considerar la creación de una estrategia distinta para cada mercado objetivo en específico.

Una estrategia que implique la segmentación de mercados evade la falta de exactitud de la planeación en masa (Anable 2005) cuyo propósito es dirigirse a la mayor cantidad de agentes económicos posible; estima el tamaño efectivo de su mercado objetivo y presenta una estrategia de planeación apropiada para cada segmento Cui (2000).

La premisa detrás de la segmentación geográfica es desagregar la información lo más posible para que el fenómeno sea medido a un nivel más certero, a su vez trata de maximizar la información bajo estudio.

En su trabajo Ter Hofstede (1999), encontró que la segmentación que realizó en términos de percepciones de consumidor y actitudes no se superpone con las barreras políticas de los países.

Mittal (2004) también se refirió a este fenómeno y agregó que el patrón de variabilidad regional es probable que no coincida con las barreras políticas o zonas tales como los estados o condados. Sin embargo las personas y los medios a menudo se identifican con sus estados (Kahle 1986), por lo tanto agrupaciones de municipios podrían constituir regiones más significativas, incluso si no existe ninguna razón histórico-cultural que los una.

Govind (2014), apoyó este enfoque realizando un estudio a nivel de código postal, argumentando que lo ayuda a analizar diferencias regionales sin la restricción de barreras pre especificadas y Cui (2000), en este sentido propuso la investigación sistemática de las variaciones regionales, ya que estas pueden ayudar a planear nuevas introducciones de producto y estrategias de expansión y así poder superar barreras entre regiones. En este mismo sentido Hawkins (1980), mencionó que la consideración sistemática de la posible influencia geográfica, es un intento de concienciar a los administradores de su importancia y de la habilidad que se debe tener en localizar con precisión en qué punto del proceso de toma de decisión ocurre ésta influencia; con esto se debe lograr que el desarrollo de una estrategia sea más completo y efectivo.

Chaudhuri (2005), también hizo referencia a este proceso al mencionar que los valores culturales variantes entre regiones pueden explicar las diferencias en las variables, por ejemplo del marketing.

Tödtling y Wanzenböck (2003), en su estudio mostraron que en Austria hay marcadas diferencias regionales en la actividad de puesta en marcha de nuevas empresas, en términos tanto de la intensidad y características. Señalaron que la actividad de puesta en marcha en las zonas industriales viejas y en zonas rurales era sustancialmente más baja que el promedio, y las características estructurales también eran menos positivas. En términos de política económica, concluyeron que, si bien hay un mejoramiento general del entorno por la creación

de empresas, una mayor diferenciación regional de recursos financieros, apoyo informativo y asesoramiento es deseable, ya que no sólo las condiciones para la formación de la nueva empresa, sino también la intensidad y características de la creación de empresas varían considerablemente entre las regiones.

López-Bazo y Motellón (2012), utilizaron datos a nivel micro para analizar el efecto del capital humano en las diferencias salariales por regiones. Los resultados para el grupo de las regiones españolas confirmaron que existen diferencias en cuanto a la dotación de capital humano, pero también variaron en gran medida los beneficios que las personas obtuvieron a partir de este capital en las diferentes regiones. La heterogeneidad regional de los beneficios fue especialmente aguda en el caso de la educación, en concreto cuando se consideró qué efecto tenía en la capacidad de la inserción laboral de las personas. Estas diferencias en las dotaciones y, especialmente, en los beneficios del capital humano representaron un porcentaje significativo de las diferencias salariales por regiones.

Høgni y Lars (2012), examinaron la geografía desigual y las relaciones entre el crecimiento del capital humano, así como el incremento del empleo total de los municipios daneses. Encontraron que el sector público contribuye, con el tiempo, a la disminución de la distribución espacial desigual del capital humano, mientras que el sector privado aumenta la desigualdad espacial.

En otro estudio Høgni y Lars (2012), examinaron en qué medida las diferentes competencias y capacidades laborales se relacionaron con el crecimiento del empleo municipal utilizando nueve categorías estratificadas, como representantes de diferentes niveles educativos del capital humano. Dividieron los municipios en cuatro categorías espaciales que fueron desde lo urbano al periférico, y concluyeron que existe una fuerte distinción espacial de las estructuras de educación con un sesgo urbano, y determinaron cuales categorías educativas distintas del capital humano académico, pueden impulsar el crecimiento del empleo en el ámbito municipal.

En los estudios de aglomeración, los efectos de las diferentes externalidades regionales relacionadas con la transferencia de conocimientos siguen siendo muy poco claros. A fin de explicar las agrupaciones de innovación, los investigadores destacan la contribución de la transferencia localizada de conocimientos y, en concreto al calcular la función de producción del conocimiento de los efectos indirectos de la investigación interregional; no obstante, se

presta menos atención a otras causas de la heterogeneidad espacial. En los trabajos aplicados, la asociación espacial en datos está econométricamente relacionada con la evidencia de transferencia de investigación. Guastella y Van Oort (2015), argumentaron que en un entorno de función de producción del conocimiento, el omitir la heterogeneidad espacial puede producir estimaciones sesgadas de la transferencia de investigación. Como prueba empírica, calcularon una función de producción del conocimiento espacial a partir de datos regionales de la UE-25, incluyendo una tendencia espacial para controlar la variación espacial no explicada en la innovación; determinaron que al tener en cuenta las características geográficas, se debilita en gran medida la evidencia de la transferencia de investigación interregional.

En resumen, los estudios académicos o el diseño de políticas en un contexto regional requiere el reconocimiento explícito de la heterogeneidad espacial en las características de una comunidad, y como afecta las variables objetivo (Kamar, et al, 2007). Reconocer esta variación espacial implica de algún modo aplicar un modelado estadístico que tome en cuenta esta consideración.

La técnica de modelado estadístico más comúnmente utilizada en las ciencias sociales es la regresión. En las aplicaciones estándar de la regresión, una variable dependiente se relaciona con un conjunto de variables independientes siendo sus principales resultados la estimación de los parámetros (β_s) que relacionan cada variable independiente con la dependiente. Un problema importante con esta técnica cuando se aplica a datos espaciales es suponer que los procesos que se estudian son constantes en el espacio, es decir, suponer un modelo de ajuste global.

Al proporcionar sólo una medida “global” para todo el espacio, los enfoques estándar como los mínimos cuadrados ordinarios (OLS) u otros modelos econométricos espaciales tienden a comprometer la heterogeneidad espacial en favor de estimaciones promedio y de eficiencia.
Los científicos regionales normalmente juntan los datos espaciales a través de regiones y localidades para examinar el impacto de varias variables. Los enfoques estándar como los mínimos cuadrados ordinarios (OLS) o la econometría espacial ganan eficiencia al utilizar todos los datos, pero ocultan la heterogeneidad regional, ya que las respuestas marginales de las variables explicativas se presuponen por lo general fijas en el espacio, es decir, para cada variable, hay un coeficiente de regresión para toda la muestra. Este enfoque ignora uno de los principios fundamentales de la ciencia regional; lo relacionado con la localización espacial.

Los científicos regionales esperan no sólo que las variables explicativas se diferencien a través del espacio, sino también que las respuestas marginales a los cambios en las variables explicativas, puedan variar a través del espacio. En el lenguaje de regresión, tanto las X y las β_s variarían espacialmente, no sólo las primeras, como los enfoques estándar implícitamente suponen.

Existen al menos tres razones para sospechar que las relaciones entre las variables cambian en el espacio.

1. La primera y más simple es que inevitablemente habrá variaciones espaciales en las relaciones observadas causadas por las variaciones del muestreo aleatorio. La contribución de esta fuente de no estacionariedad espacial no suele ser de gran interés en sí misma, sino que necesita ser reconocida y representada si se han de identificar otras fuentes más interesantes de no estacionariedad espacial. Es decir, sólo se está interesado en variaciones relativamente grandes en las estimaciones de los parámetros que son poco probable que se deban a la variación debida al muestreo.

2. La segunda razón se debe a que las relaciones podrían ser intrínsecamente diferentes a través del espacio. Por ejemplo, hay variaciones espaciales en actitudes o preferencias de las personas o hay diferentes aspectos contextuales administrativos, políticos o de otro tipo que producen diferentes respuestas a los mismos estímulos sobre el espacio. Es difícil conjeturar un ejemplo de esta causa de no estacionariedad espacial en la geografía física, donde las relaciones que se miden son gobernadas por las leyes de la naturaleza. La idea de que el comportamiento humano pueda variar intrínsecamente en el espacio es consistente con las creencias posmodernistas en la importancia del lugar y localidad como marcos para la comprensión de este tipo de comportamiento. Se ha criticado el análisis cuantitativo en la geografía por tener poca relevancia en las situaciones del "mundo real" donde las relaciones son muy complejas y posiblemente altamente contextuales. Indicadores estadísticos locales abordan estas críticas reconociendo tal complejidad y tratando de describirlo (Fotheringham 2006).

3. La tercera razón por la que las relaciones pueden exhibir no estacionariedad espacial es que el modelo a partir del cual las relaciones se miden es burdo con una mala especificación de la realidad y que una o más variables relevantes o bien se omiten en el modelo o están representados con una forma funcional incorrecta. Este punto de vista,

está más en línea con la escuela positivista de pensamiento, en la cual se asume que se puede hacer una declaración global de la conducta y por tanto es aplicable a las relaciones tanto en la geografía física, como en la humana, pero que la estructura del modelo no es lo suficientemente bueno para poder realizar esto. En pocas palabras, ¿se pueden eliminar todos los efectos contextuales mediante una mejor especificación de los efectos a nivel individual? (Hauser, 1970). Si los errores de modelo son la causa de la inestabilidad paramétrica, el cálculo y el mapeo posterior de estadísticas locales es útil para comprender la naturaleza de la mala especificación en forma más clara.

Los estudios académicos o el análisis de la política regional pueden tergiversar los patrones reales al ignorar la heterogeneidad espacial de las respuestas. El análisis convencional se basa generalmente en un promedio de la muestra "global" para las β_s. Aunque un promedio global es un punto de referencia útil para hacer afirmaciones generales sobre las respuestas a una variable, es evidente que no puede reflejar la respuesta real para muchas regiones. Por lo tanto, una pregunta importante es: ¿Qué cantidad de información y objetividad se pierde con los enfoques tradicionales y cuánto pierde el análisis de la política correspondiente?, otra pregunta sería, ¿cómo cuantificar estas cuestiones?

La regresión geográficamente ponderada (GWR) representa un enfoque prometedor para comenzar a abordar estas cuestiones (Brunsdon, Fotheringham y Charlton 1998; Fotheringham, Brunsdon y Charlton 2002). El Enfoque GWR estima una muestra localmente variable para cada punto de observación potencial o para otros lugares deseados, produciendo un conjunto separado de parámetros de regresión para cada punto de observación. Estos parámetros reflejan la heterogeneidad de la muestra mediante la estimación de diferentes respuestas marginales de una variable explicativa a través del espacio.

La regresión geográficamente ponderada, permite el modelado de procesos que varían en el espacio. Los resultados de la GWR es un conjunto de estimaciones de los parámetros locales para cada relación que se pueden asignar para producir una superficie de parámetros a través de la región de estudio. De esta manera la GWR ofrece una valiosa información sobre la naturaleza de los procesos que se están investigando y reemplaza las formas globales tradicionales de los modelos de regresión.

El Enfoque GWR es todavía relativamente nuevo en la literatura, aunque hay un aumento de aplicaciones empíricas, entre las cuales se pueden citar las siguientes:

Eckey et al (2007), utilizaron la técnica de la regresión geográficamente ponderada para analizar detalladamente los procesos de convergencia en Alemania. Estimaron una velocidad individual de convergencia para cada región en base a los coeficientes locales de las ecuaciones de regresión. Obtuvieron diferentes velocidades de convergencia para las regiones. En particular, encontraron que las regiones de Baviera tienen un largo período de vida media y los distritos del norte de Alemania un corto período de vida media. Este enfoque proporcionó evidencia de que las regiones del sur de Alemania con alta productividad laboral y una baja tasa de desempleo serían las regiones más prósperas de Alemania. Sobre la base del desarrollo económico, a la larga, habrá una brecha entre el norte y el sur de Alemania. Los coeficientes sustancialmente diferentes muestran que un modelo de convergencia global, que ha sido estimado por muchos investigadores (Kosfeld y Lauridsen, 2004; Funke y Niebuhr, 2005a, b;.Kosfeld et al, 2006), podrían ser mejorados mediante un enfoque de regresión geográficamente ponderada. Además encontraron que la velocidad de convergencia es substancialmente inferior en el sector de la producción que en el sector de servicios.

(Huang et al, 2010), señalaron que hay que incorporar los efectos temporales en el modelo de regresión geográficamente ponderada (GWR), mediante un modelo GWR de regresión extendido, geográficamente y temporalmente ponderado. Utilizaron un modelo (GTWR), desarrollado para hacer frente tanto a la no estacionariedad espacial y temporal en los datos del mercado de bienes raíces. Con el fin de probar su rendimiento mejorado, GTWR se comparó con los mínimos cuadrados ordinarios globales, en términos de bondad de ajuste y otras medidas estadísticas, utilizando un estudio de caso de las ventas de vivienda residencial en la ciudad de Calgary, Canadá, de 2002 a 2004. Los resultados mostraron que hubo beneficios sustanciales en modelar tanto la no estacionariedad espacial como la temporal de forma simultánea.

La estimación de población a partir de datos obtenidos por detección remota mediante el modelo global ordinario de regresión lineal (OLS) no puede hacer frente al problema de no estacionariedad espacial, por lo que Lo (2013), aplicó un modelo local de regresión geográficamente ponderada (GWR), utilizando cuatro variables de uso de suelo de la zona: uso de alta densidad urbana, uso urbano de baja densidad, tierras de cultivo y los bosques, para la estimación de la población en la ciudad de Atlanta, Georgia en el nivel de sección censal, y encontró que el modelo GWR local de cuatro clases podrían ayudar a mejorar la exactitud de la estimación de la población con respecto al modelo global en un 28%.

El problema de modelar la ocurrencia de incendios forestales a partir de indicadores socioeconómicos y demográficos, junto con la cubierta vegetal y las estadísticas agrícolas, fue abordado por Koutsias et al (2015), en un análisis geoestadístico realizado a nivel provincial en el sur de Europa. El modelo global no fue suficiente para describir completamente los factores causales subyacentes en el modelado de ocurrencia de los incendios forestales por lo que se utilizó un enfoque GWR para superar el problema de la no estacionariedad. Los resultados confirmaron la importancia de las actividades agrarias, abandono de la tierra, y los procesos de desarrollo como factores determinantes de la ocurrencia de incendios. La identificación de regiones con diferentes relaciones espaciales puede contribuir a la mejor comprensión del problema de los incendios, especialmente en grandes áreas geográficas, mientras que al mismo tiempo se reconoció su carácter local, lo que es muy importante para la gestión y la política de incendios.

4.3. Datos

Nuestra muestra se basa en los resultados de todas las personas entrevistadas por el Instituto Nacional de Estadística de España (INE) durante los Censos de Población y Vivienda en los años 2001 y 2011. Dichos censos resumen en más de 250 variables las características que tienen las personas y sus viviendas en todo el territorio peninsular y no peninsular de España, esto incluye Ceuta, Melilla, Islas Baleares y Canarias. Desafortunadamente, por razones de anonimato los censos del 2001 y 2011, únicamente reportan la información de los municipios que cuentan con más de 20,000 habitantes, en el 2001 fueron 317 y en 2011, 394 municipios. Debido a que esta investigación compara el año 2001 y el 2011, los municipios del 2011 que no fueron reportados en 2001 se descartaron.

Con la información de las personas que habitan en cada municipio es posible expresar las características de cada uno de ellos en términos del capital humano entre ellas: tasas de dependencia, diferencias laborales entre hombres y mujeres, proporción de trabajadores dedicados a la manufactura, construcción, servicios, proporción de habitantes que cuentan con estudios terminados de tercer nivel.

Se obtuvieron las proporciones con respecto del total de habitantes que cuentan con educación de tercer grado terminada, la proporción de trabajadores dedicados a la manufactura, servicios, construcción, agricultura según el Directorio de Nacional de Actividades (CNAE); la proporción de trabajadores manuales calificados, no calificados,

supervisores no manuales, administradores, profesionales y técnicos según el Directorio Nacional de Ocupaciones (CNO). Esto es posible por el nivel de información que se encuentra codificada en el Censo, lo que permite conocer en detalle la información de cada persona entrevistada.

La descripción de las variables utilizadas en este estudio se resumen a continuación:

Variable dependiente:

Crecimiento (decrecimiento) de puestos de empleo en el período 2001-2011 en cada municipio. (CRECEMP).

Variables independientes:

Razón de empleados: administradores, profesionales y técnicos por municipio, (PROF).

Razón de empleados: trabajadores manuales no calificados por municipio, (MANNC).

Diferencia entre la razón de empleados hombres profesionales activa y la razón de mujeres profesionales empleadas con respecto a la población económicamente activa, (DIFPEAP).

Razón de trabajadores empleados: supervisores no manuales por municipio, (SUPNM).

Razón de trabajadores empleados: trabajadores manuales calificados por municipio, (MANC).

Razón de trabajadores en empleos relacionados con manufactura por municipio, (MFAC)

Razón de trabajadores en empleos relacionados con la pesca, acuicultura y actividades de los servicios relacionados por municipio, (PESC).

Razón de la población con educación de tercer nivel (diplomatura, licenciatura y doctorado) por municipio. (EDU3).

Diferencia entre la razón de hombres ocupados con respecto a la población económicamente activa y la razón de mujeres ocupadas con respecto a la población económicamente activa, (DIFOCUP).

Razón de trabajadores con empleo en el sector de la construcción por municipio, (CONST).

4.4. Metodología

En España el Censo de Población y Vivienda se lleva acabo por el INE periódicamente cada 10 años, por lo que es posible cuantificar la evolución de la información, y verificar el impacto sobre las variables económicas. En este caso se determinó la variación porcentual de la cantidad de empleos entre el año 2001 y 2011 de cada uno de los 317 municipios con los que cuenta la muestra. Posteriormente se planteó mediante un modelo de regresión lineal múltiple, la variación en la cantidad de empleos en función del capital humano que habita en cada municipio. Se eliminaron del modelo las variables para las cuales el valor de β no es significativo con un nivel de significancia $\alpha = 0.10$.

Las variables eliminadas fueron las que cuantifican la diferencia entre la tasa de desempleo de hombres y de mujeres, la densidad de población, la proporción de ocupados (personas con empleo) cuya actividad está relacionada con servicios (CNAE 60-75), la proporción de ocupados cuya actividad está relacionada con la agricultura (CNAE 1), la tasa de dependencia, la proporción de la población total que pertenece a la PEA y la proporción de la población cuya actividad está relacionada con la construcción (CNAE 80).
El modelo final planteado de regresión lineal está expresado en la siguiente ecuación:

$$CRECEMP = \beta_0 + \beta_1 PROF + \beta_2 MANNC + \beta_3 \mathrm{DIFPEAP} + \beta_4 \mathrm{SUPNM} + \beta_5 \mathrm{MANC} + \beta_6 \mathrm{MFAC} + \beta_7 \mathrm{PESC} + \beta_8 \mathrm{EDU3} + \beta_9 \mathrm{DIFOCUP} + \beta_{10} \mathrm{CONST} + \varepsilon_i \quad (3.11)$$

Todas las estimaciones de los parámetros en el modelo resultaron significativos ($P < 0.05$), excepto el parámetro de la variable que resume el porcentaje del empleo dedicado a la pesca, el cual es significativo a un nivel de significancia $P < 0.10$.

Cabe señalar que el modelo es significativo y explica un 58% del comportamiento de la variable dependiente, con un coeficiente de determinación, $r^2 = 0.5894$. Es decir el modelo logra capturar buena parte del comportamiento de la variable, pero aún queda una proporción de más del 40% que no es explicada.

Además, se observa un alto grado de variación en la tasa de crecimiento (decrecimiento) de empleo a lo largo de los 317 municipios de la muestra; toma valores en un rango que va desde un máximo de crecimiento de empleo 58.05% en el municipio de Barabate, Cadiz, Provincia de

Andalucía, hasta una tasa mínima de decrecimiento de empleo de -75.44% en el municipio de Arraste/Mondragón, Guipúzcoa, en el País Vasco.

Es decir, existe una indicación general de variación espacial en la muestra, que sugiere el uso de la regresión geográficamente ponderada como técnica para describir las variaciones del crecimiento de empleo nivel local. Esta metodología permite describir variaciones espaciales en las relaciones de las variables del modelo.

Se procesaron los datos mediante una regresión por mínimos cuadrados y luego con una regresión geográficamente ponderada, utilizando validación cruzada para determinar el ancho de banda que fue de 350 Km y se consideró un kernel gaussiano.

4.5. Evidencia empírica.

En esta sección se analiza el grado de impacto que tiene sobre el crecimiento del empleo el capital humano de las personas que habitan en cada municipio. El capital humano fue capturado por la proporción de personas de cada municipio que cuentan con las características expresadas en las variables independientes seleccionadas. Es de esperarse que exista diversidad en las respuestas locales a lo largo del espacio, es decir se espera que las variables explicativas tengan efectos diferenciales en el crecimiento del empleo a lo largo del espacio (Xu, 2014).

A continuación se presentan los resultados del modelo de GWR con mapas de los coeficientes de educación de tercer nivel, distancia entre la proporción de hombres y mujeres profesionales, trabajadores manuales no calificados, supervisores no manuales, distancia de desempleo entre mujeres y hombres.

Es de esperarse que la preparación académica y la experiencia laboral, así como las derramas a nivel regional asociadas a estos factores, tengan una influencia positiva en el crecimiento de la economía y por lo tanto en el incremento de puestos de empleo tanto a nivel global como regional.

La metodología GWR nos permite detectar si las influencias de las variables explicativas utilizadas son uniformes en todos los municipios de la muestra o existen impactos

diferenciados a niveles regionales, los cuales son producto de las características particulares y únicas de cada municipio.

La técnica GWR postula que entre más cercanos geográficamente sean dos municipios la influencia mutua que es ejercida es mayor y viceversa. Esta influencia es contabilizada con la ponderación que se asigna a cada municipio que varía de acuerdo a la cercanía geográfica y que posteriormente se resume en los gradientes regionales producidos por el modelo geográficamente ponderado.

Los coeficientes de las variables independientes son presentados en mapas a continuación. El periodo que se utilizó para este análisis es muy particular, ya que se encuentra entre los años 2001 y 2011, periodo que coincide con ajustes económicos y laborales en España de gran magnitud, producto de la desaceleración económica que afectó a la mayor parte del mundo.

Análisis futuros podrían ayudar a desarrollar políticas de desarrollo económico en especial con respecto al gasto o inversión en el ramo educacional y de preparación del capital humano.

4.5.1. Distancia entre la proporción de hombres y mujeres profesionales, DIFPEAP, como determinante del crecimiento del empleo.

El parámetro global de la variable DIFPEAP, que cuantifica la diferencia entre las proporciones con respecto de la población económicamente activa de hombres profesionales y de mujeres profesionales, muestra una relación negativa con el crecimiento del empleo, CRECEMP. Es decir estas dos variables tienen una relación inversamente proporcional, lo que indica que cuando la diferencia entre la proporción de hombres y mujeres profesionales es menor, el crecimiento del empleo es mayor.

A nivel local el parámetro tiene variaciones que van desde un mínimo en el Municipio de Chiclana de la Frontera, Cádiz, Andalucía, hasta un máximo en el municipio de Cambrils, Tarragona, Cataluña. Destaca que el gradiente de la relación entre CRECEMP y la variable DIFPEAP, es mayor en el sur y centro de España. Este gradiente regional de mayor magnitud implica que en esas regiones el impacto en el crecimiento del empleo es mayor que por ejemplo, en Tarragona cuando hay un cambio en la variable DIFPEAP. En la Figura 20 se muestra el mapa de la región en cuestión, Centro-Sur de España, en la que se puede observar la variación espacial de cada parámetro estimado por la regresión ponderada

geográficamente. Como se puede ver la región Centro-Sur de España es en la zona en la que el gradiente de la relación entre las dos variables presenta mayor nivel de variación.

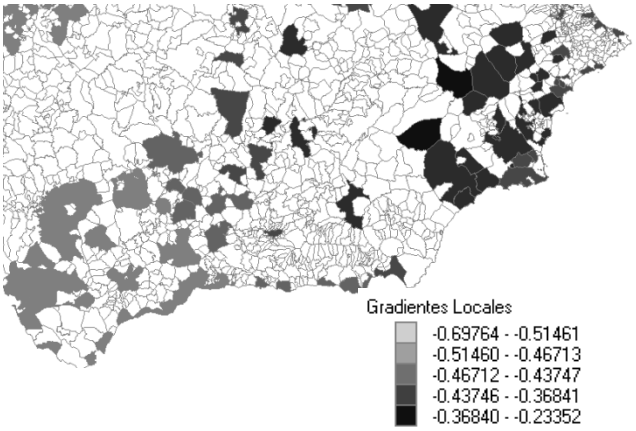

Figura 20. Distancia entre la proporción de hombres y mujeres profesionales, DIFPEAP, como determinante del crecimiento del empleo.

4.5.2. Educación de tercer nivel, EDU3, como determinante del crecimiento del empleo.

El gradiente de la relación entre las variables crecimiento de empleo, CRECEMP, y educación de tercer nivel, EDU3, indica una relación directamente proporcional en todos los municipios; es decir, cuando la proporción de personas con educación de tercer nivel es mayor en el municipio se puede esperar que la generación de empleos sea mayor en ese municipio, como se puede observar a nivel global en todos los municipios de la muestra, sin embargo existen variaciones espaciales en los parámetros calculados a nivel local. Como se puede observar en la Figura 21, los gradientes locales o parámetros locales estimados, son menores en las localidades que se encuentran más cercanas a la costa del mediterráneo y se van haciendo mayores a medida que aumenta la distancia con respecto a la región costera.

En la costa norte de España, los gradientes también son menores aunque su magnitud es mayor que en la costa mediterránea, lo que indica que existen factores distintos que afectan al crecimiento del empleo en las costas norte y mediterránea de España. Esta tendencia se puede deber a que el crecimiento del empleo en las regiones costeras depende principalmente de los ingresos del sector turístico y a la exportación de la imagen de España como destino de turismo de playa.

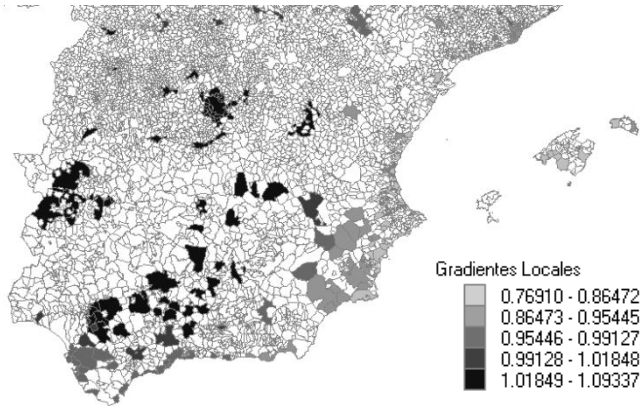

Figura 21. Educación de tercer nivel, EDU3, como determinante del crecimiento del empleo.

4.5.3. Trabajadores manuales no calificados, MANNC, como factor de crecimiento del empleo.

El gradiente global del modelo que captura la relación entre la variable crecimiento del empleo y la proporción de trabajadores manuales no calificados, muestra que su relación es positiva y directamente proporcional, lo que indica que como factor de crecimiento del empleo, la proporción de trabajadores manuales no calificados influye positivamente en la variación del número de empleos entre los años 2001 y 2011 en todos los municipios que componen la muestra.

Sin embargo, a nivel local se pueden ver variaciones en los valores de los parámetros locales, destacando los casos de las zonas metropolitanas, Barcelona y Zaragoza, en los que los gradientes toman valores de lo más pequeño, indicando que en esas dos ciudades la variable dependiente, crecimiento del empleo, no tiene variaciones particularmente importantes ante los cambios de la variable proporción de trabajadores manuales no calificados.

El caso de la zona metropolitana de Madrid, Figura 22, es particular, ya que al variar la proporción de trabajadores manuales no calificados, la variable crecimiento del empleo tiene variaciones más importantes en los municipios al sur de la zona metropolitana: Getafe, Pinto, Fuenlabrada, Móstoles Alcorcón, Rivas-Vaciamadrid, los cuales, se caracterizan por ser una zona con un gran porcentaje de su territorio dedicado a la industrial y manufactura.

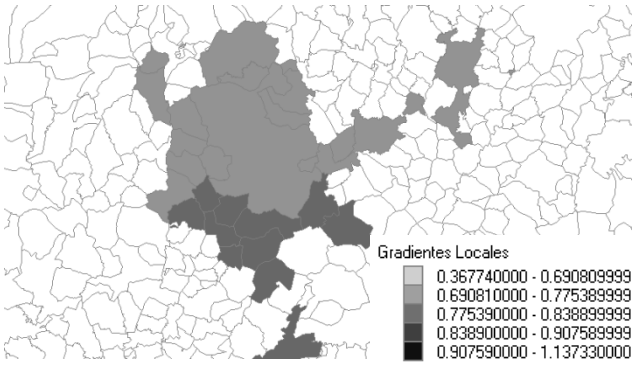

Figura 22. Trabajadores manuales no calificados, MANNC, como factor de crecimiento del empleo en Madrid.

4.5.4. El segmento de los supervisores no manuales, SUPNM, como determinante del crecimiento del empleo.

Para determinar el grado de impacto que tiene sobre el crecimiento del empleo el capital humano que acumulan los supervisores no manuales, SUPNM, se calculó la proporción de personas del total de habitantes que se encuentran empleadas en trabajos con ocupaciones en áreas de supervisión no manual en cada municipio de la muestra. Posteriormente se relacionó esta proporción con el crecimiento porcentual de la cantidad de empleos en cada municipio a través de un modelo de regresión geográficamente ponderada.

A nivel global el gradiente del modelo que captura la relación entre el crecimiento de empleo y el capital humano acumulado por trabajadores de supervisión no manual, muestra una relación positiva en todos los municipios de la muestra, es decir a mayor proporción de supervisores no manuales en un municipio, mayor es su crecimiento de empleo.
Sin embargo a nivel local los parámetros que resumen esta relación se hacen más pequeños a medida que los municipios se acercan más a la costa, Figura 23, excepto en las Islas Canarias y en los municipios conurbados de Barcelona, lo que implica que las derramas de valor agregado de los supervisores no manuales medidas en términos de crecimiento de empleo generan más impacto en las áreas no costeras de España.

Este fenómeno puede deberse a que las habilidades de los supervisores no manuales generan más riqueza en los municipios cuya actividad económica se encuentra diversificada en un número mayor de actividades. Es posible que en la costa las actividades económicas de los municipios dependan en gran parte de la industria turística, esta situación no se refleja en los

casos de Las Islas Canarias y de Barcelona que a pesar de ser zonas costeras, su actividad económica se encuentra diversificada en una gran variedad de actividades.

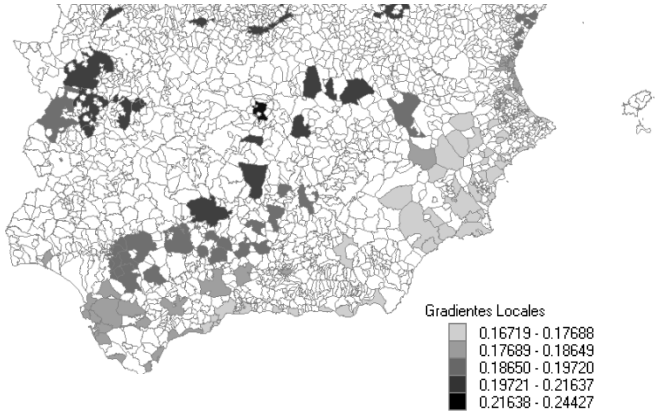

Figura 23. Supervisores no manuales, SUPNM, como determinante del crecimiento del empleo.

4.5.5. Trabajadores manuales calificados, MANC, como determinante del crecimiento del empleo.

Nuevamente a nivel global la relación entre las dos variables es positiva y significativa, lo que sugiere una relación directamente proporcional entre el crecimiento del empleo y la proporción de trabajadores manuales calificados. A nivel local también todos los coeficientes son positivos en los 317 municipios de la muestra. La relación positiva se hace más fuerte a medida que los municipios se acercan más al centro de la península y se aleja tanto de la costa norte como de la costa mediterránea, Figura 24.

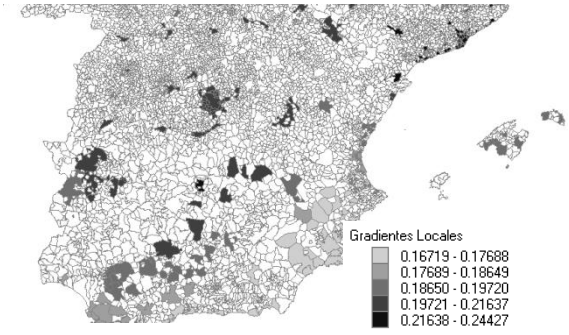

Figura 24. Trabajadores manuales calificados, MANC, como determinante del crecimiento del empleo.

4.5.6. Influencia del Capital Humano empleado en manufactura, MFAC, como factor de crecimiento del empleo.

A nivel global el coeficiente que resume la relación entre el capital humano empleado en manufactura y la variable crecimiento de empleo, tiene signo positivo y es significativo ($P < 0.01$), lo que sugiere que entre mayor sea la proporción de empleo en manufactura en el municipio, mayor será el crecimiento del empleo.

A nivel regional el mapa, Figura 25, distingue diferencias en el impacto de esta variable sobre el crecimiento del empleo entre el este y el oeste de España, lo que conduce a pensar que la industria manufacturera se encuentra alejada de la costa del mediterráneo (con excepción de Barcelona y alrededores) y que su impacto es más prominente en las regiones del Oeste de España.

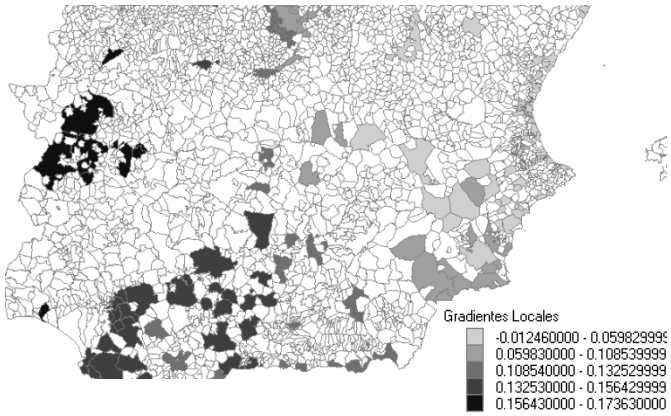

Figura 25. Capital Humano empleado en manufactura, MFAC, como factor de crecimiento del empleo.

4.6. Conclusiones.

La heterogeneidad espacial se espera que surja porque los mercados laborales locales varían en su estructura, en su contexto social y en su historia, en formas que no se capturan fácilmente mediante las variables explicativas en la regresión global estándar. La heterogeneidad espacial en la dinámica de crecimiento puede hacer estimaciones globales engañosas en términos de resultados locales. Por ejemplo, aceptar hallazgos con respecto al papel de las variables clave en el crecimiento económico, puede ser el resultado de estimaciones globales, que enmascaran significativamente la variación local, incluso en la dirección de la influencia. Además, las estimaciones estándar pueden sugerir que no hay efectos marginales cuando en realidad el factor estimula el crecimiento en algunas áreas

mientras lo reduce en otras, produciendo un efecto promedio aproximadamente igual a cero. Dejando aparte la importancia de descubrir la naturaleza verdadera de las relaciones, el desarrollo de políticas económicas locales exitosas, requiere un conocimiento de los procesos socioeconómicos locales y de la dinámica de su crecimiento.

Una aproximación reciente que ha ganado popularidad para conocer potencialmente la heterogeneidad geográfica en los procesos socioeconómicos es la regresión geográficamente ponderada, GWR, que en contraste con el enfoque de la regresión global, puede estimar coeficientes separados para cada área. En la estimación de cada región, las características de las áreas individuales se incluyen en una submuestra que se pondera mediante su proximidad espacial. La variación suavizada ponderada espacialmente en las estimaciones de los parámetros revelan las extensas diferencias regionales en las respuestas marginales locales.
En este contexto, se modeló con el enfoque de la regresión geográficamente ponderada, el crecimiento del empleo en España en función del capital humano, descrito por las siguientes variables dependientes: Razón de empleados: Administradores, profesionales y técnicos por municipio. Razón de empleados: Trabajadores manuales no calificados por municipio- Diferencia entre la razón de empleados hombres profesionales activa y la razón de mujeres profesionales empleadas con respecto a la población económicamente activa. Razón de trabajadores empleados: Supervisores no manuales por municipio. Razón de trabajadores empleados: Trabajadores manuales calificados por municipio. Razón de trabajadores en empleos relacionados con manufactura por municipio. Razón de trabajadores en empleos relacionados con la pesca, acuicultura y actividades de los servicios relacionados por municipio. Razón de la población con educación de tercer nivel (diplomatura, licenciatura y doctorado) por municipio. Diferencia entre la razón de hombres ocupados con respecto a la población económicamente activa y la razón de mujeres ocupadas con respecto a la población económicamente activa. Razón de trabajadores con empleo en el sector de la construcción por municipio.

Se encontró que la distancia entre la proporción de hombres y mujeres profesionales tiene mayor influencia en el crecimiento de empleo en el centro-sur de España; el crecimiento del empleo con respecto a la educación de tercer nivel en las regiones costeras, muestra una relación menor a la que se encuentra tierras adentro, lo cual se explica en función al sector turístico como fuerte componente en la generación de empleos en las zonas costeras; se encontró que las derramas de valor agregado de los supervisores no manuales medidas en términos de crecimiento de empleo generan más impacto en las áreas no costeras de España,

fenómeno que puede deberse a que las habilidades de los supervisores no manuales generan más riqueza en los municipios cuya actividad económica se encuentra diversificada; se encontró que la relación entre el crecimiento del empleo y la proporción de trabajadores manuales calificados a nivel local es mayor a medida que los municipios estén más cerca del centro del país y se detectó una mayor influencia en el oeste de España del capital humano empleado en manufactura como factor de crecimiento del empleo.

Los resultados obtenidos permiten recomendar el modelado estadístico mediante la regresión geográficamente ponderada, ya que al obtener desagregaciones locales de un estadístico global se pueden obtener mapas de la respuesta y con esto conocer en nuestro caso, el comportamiento del crecimiento del empleo en función del capital humano a través de todo el país.

5. Identificación de Distritos Industriales en México.

5.1. Introducción.

La premisa básica de los beneficios colaterales que se consigue con la aglomeración y la proximidad geográfica entre grupos de empresas, como la difusión de conocimiento, el fortalecimiento de las relaciones entre empresas y el establecimiento de instituciones de apoyo, sigue presente, particularmente en la Unión Europea y en los Estados Unidos de Norteamérica. El concepto de distrito industrial fue propuesto por Alfred Marshall y ha sido la base de la teoría italiana del distrito industrial.

Alfred Marshall (1890), estableció que de la misma manera que las empresas pueden gozar de economías de escala en su interior es posible también de gozar de economías de escala externas, es decir introduce el concepto de externalidades positivas.

En su concepto original de distrito industrial, Marshall definió una región en la que la estructura empresarial está compuesta por empresas pequeñas, de propiedad local de tal forma que las decisiones de la inversión y la producción también son locales. Las economías de escala son relativamente bajas, previniendo la aparición de grandes empresas. Dentro del distrito, el comercio se tramita sustancialmente entre compradores y vendedores, y a menudo implican contratos o compromisos a largo plazo. Aunque Marshall no lo mencionó explícitamente, los vínculos y la cooperación con empresas localizadas fuera del distrito se suponen mínimos.

Lo que hace al distrito industrial especial, es la naturaleza y calidad de la mano de obra local que es interna al distrito y altamente flexible. Las personas se mueven de una empresa a otra, y los propietarios, así como los trabajadores viven en la misma comunidad, donde se benefician del hecho de que los secretos de la industria están en el aire.

Los trabajadores están comprometidos con el distrito y no con la empresa. La emigración de trabajadores es mínima, mientras que la inmigración ocurre con motivos de crecimiento. El distrito es visto como una comunidad relativamente estable que permite la evolución de una fuerte identidad cultural local y una especialización industrial compartida.

El distrito marshalliano también abarca un conjunto relativamente especializado de servicios adaptados a industrias del distrito de productos únicos. Estos servicios incluyen experiencia técnica en ciertas líneas de productos, maquinaria y comercialización, y los servicios de mantenimiento y reparación. Incluyen a las instituciones financieras locales dispuestas a tomar riesgos a más largo plazo, ya que tienen información interna y confianza en los empresarios locales.

Todas estas características se incluyen en el concepto de aglomeración, que sostiene que la adherencia a un lugar no depende de la preferencia de localización de las empresas o de los trabajadores, sino que queda determinada por las economías externas a disposición de cada empresa a partir de su relación espacial con otras empresas y proveedores de servicios.

En los distritos de Marshall no es necesario que cualquiera de sus actores sea consciente de que cooperan entre sí, para que el distrito exista y funcione como tal, pero en su variante Italiana, la que se ha extendido en Europa y los Estados Unidos, se ha argumentado que los esfuerzos de cooperación concertada entre los actores del distrito y la construcción de estructuras de gobierno con el fin de mejorar la competitividad del distrito aumenta su fortaleza.

El concepto de distrito industrial se hizo relevante de nuevo en la economía cuando la recesión golpeó al mundo en los años 1970 y 1980. A pesar del aumento del desempleo y el declive económico general, algunas regiones prosperaron. Estas regiones se encontraban en diferentes partes del mundo, participaban en una variedad de industrias, incluyendo industrias avanzadas, como las de mano de obra tradicional.

La Organización para la Cooperación y el Desarrollo promueve a los clusters como el enfoque a seguir para un crecimiento económico y desarrollo industrial regional sustentable (OECD, 1999).

México estableció el Fondo de Apoyo para la Micro, Pequeña y Mediana Empresa (FONDO PYME) que es un instrumento que busca apoyar a las empresas en particular a las de menor tamaño y a los emprendedores con el propósito de promover el desarrollo económico nacional, a través del otorgamiento de apoyos de carácter temporal a programas y proyectos que fomenten la creación, desarrollo, viabilidad, productividad, competitividad y

sustentabilidad de las micro, pequeñas y medianas empresas, sin embargo no está orientado al desarrollo de distritos industriales.

El objetivo del presente ensayo es identificar aglomeraciones o clusters que tengan las características de distritos industriales en México. La información parcial con la que se cuenta en este sentido en el país debe actualizarse, ya que por cuestiones económicas e incluso por la lamentable permeabilidad del crimen en algunas zonas del país, industrias enteras se han movido de localización e incluso cerrado para localizarse en otros países, como el caso de la industria maquiladora.

En México tampoco se cuenta con información censal de trayectos de viaje al trabajo por lo que una aportación de la presente Tesis Doctoral será la estimación de los flujos de viajes entre todos los municipios del país utilizando el Método de Radiación propuesto por Simini et al (2012).

5.2. Revisión de la literatura.

Alfred Marshall (1890), estableció que de la misma manera que las empresas pueden gozar de economías de escala en su interior es posible también de gozar de economías de escala externas, es decir introduce el concepto de externalidades positivas.

El vínculo entre los conceptos de Marshall y la teoría de la localización lo estableció Hoover (1937), en su trabajo sobre la industria zapatera y del curtido en los Estados Unidos. Las investigaciones indican un buen número de factores que influyen en el surgimiento de los distritos industriales, como las instituciones, la trasmisión del conocimiento, etc, pero en el estudio de Hoover, se hizo evidente que los requerimientos de materia prima y la necesidad de tener un abastecimiento seguro de pieles en la industria del calzado fueron condiciones para que las curtidurías se localizaran cerca, por lo que el autor propuso los efectos de la proximidad espacial de las empresas.

Schimitz (1999, 1995), propusieron la noción de eficiencia colectiva como elemento de cohesión del distrito industrial. La eficiencia colectiva tiene dos dimensiones principales: los efectos incidentales, es decir, el resultado de la presencia de externalidades y los efectos intencionales, resultado de estrategias gubernamentales. Shimitz estudió el caso de la

industria zapatera en Sinos Valley, Brasil y concluyó que las empresas que conforman los clusters se mantienen competitivas debido a las eficiencias colectivas que resultan en buena parte de las externalidades o efectos positivos que se derivan de la proximidad geográfica.

Markusen (1996), propuso una tipología de los distritos industriales:

1. Distritos marshallianos y su variante italiana, que son los tradicionales clusters de empresas pequeñas aglomeradas geográficamente. Prevalecen las microempresas porque gozan de economía de escala y la mayor parte de las transacciones ocurren dentro del distrito.

2. Distritos *Hub-and-Spoke,* es un tipo de distrito industrial que está presente en regiones donde un número de empresas e instalaciones clave actúan como anclas o centros para la economía regional, con los proveedores y las actividades conexas extendidos a su alrededor como radios de una rueda. Ejemplos de ellos son el centro de Seattle y New Jersey, Estados Unidos; Toyota City, Japón; Ulsan y Pohang, Corea del Sur; San José dos Campos, en Brasil. Una sola empresa grande, por ejemplo Boeing en Seattle o Toyota en Toyota City, compra a proveedores locales y externos y vende principalmente a clientes externos, que puede ser una empresa grande, por ejemplo, Boeing al ejército o Toyota a masas de consumidores. Investigaciones realizadas en este tipo de distritos se han realizado para Seattle (Gray, Golob y Markusen 1996), para el centro de Nueva Jersey (Fineberg et al 1993) y para San José dos Campos y Campinas, Brasil (Diniz y Razavi 1994).

3. Distritos Plataforma-Satélite, están compuestos de subsidiarias o corporaciones multinacionales ausentes, cuyos centros de operación o casa matriz no se encuentran físicamente en dichos distritos. Por ejemplo, Vancouver (Canadá) tiene un distrito industrial de la industria fílmica, donde gran número de estudios tienen sus subsidiarias, pero no las casas matrices. Hay un grado bajo de desarrollo de las empresas locales.

4. Distritos Estado-céntricos, se han desarrollado como resultado de la presencia de una entidad gubernamental que es el factor de desarrollo en la correspondiente zona geográfica. En Brasil, la presencia de la Universidad de Campinas ha fomentado el crecimiento de la ciudad, así como Washington D.C. ha crecido debido a la presencia del gobierno de los Estados Unidos.

Se ha discutido teóricamente las definiciones de cluster, distrito industrial y milieux. Si bien se usan a veces como sinónimos, se pueden encontrar un buen número de artículos que intentan establecer diferencias entre los tres conceptos. Por ejemplo, Rabellotti (1997, 1995) trató de determinar las características de un "distrito industrial modelo", Martin y Sunley (2003), estudiaron el concepto de cluster y Amara, Landry y Ouimet (2005), estudiaron las áreas innovadoras "milieux innovateurs". En el presente trabajo se tomaran como sinónimos cluster, distrito industrial y milieux.

La popularidad de los clusters se debe en buena medida a los investigadores italianos y el análisis del desarrollo regional de la Tercera Italia. Piore y Sabel 1984; Bellandi 1989; Goodman 1989; Sforzi 1989, investigaron los rasgos característicos de los distritos italianizantes. A diferencia de la pasividad de las empresas de Marshall, los distritos italianizantes exhiben frecuentes e intensivos intercambios de personal entre clientes y proveedores y la cooperación entre las empresas de la competencia para compartir riesgos, estabilizar mercados, y compartir la innovación. Los trabajadores compartidos están involucrados en actividades de diseño y de innovación. Las asociaciones de comercio proporcionan gestión compartida de infraestructura, capacitación, marketing, ayuda técnica o financiera así como marcan el ritmo en la estrategia colectiva. Los gobiernos locales y regionales son importantes en la regulación y promoción de las industrias básicas. La confianza entre los miembros del distrito es fundamental en su capacidad de cooperar y actuar colectivamente (Harrison 1992; Saxenian 1994), aunque los críticos argumentan que el poder de las grandes corporaciones que forman a los distritos industriales italianos se ha subestimado (Harrison 1994).

Al evaluar el crecimiento, la estabilidad, la equidad, y la política de los distritos industriales italianos, la modalidad italiana debe distinguirse de los casos de Silicon Valley y del Condado de Orange y de sus predecesores marshallianos. En términos de crecimiento y estabilidad, siempre y cuando las economías de aglomeración permanecen y no son replicadas en otros lugares, tanto los distritos industriales de Marshall como los italianos mantienen buenas perspectivas a largo plazo.

Piore y Sabel (1984), argumentaron que la historia de la industrialización ha mantenido abierta una alternativa importante al sistema de producción en masa, denominada producción artesanal, sistema que se basa en el uso flexible de maquinaria de propósito general por los trabajadores especializados, capaz de producir una amplia gama de productos para los

actuales mercados caracterizados por un alto grado de cambios. Los cada vez más segmentados mercados fuerzan a las empresas a seguir un enfoque estratégico y buscar la especialización y la flexibilidad. Las tecnologías más recientes también permiten más flexibilidad, en cuanto a la escala de la producción. Existe también mayor flexibilidad con respecto a los insumos laborales y el tipo y calidad de los productos fabricados. Mencionaron que los distritos industriales son buenos proveedores de puestos de trabajo con estabilidad a largo plazo. Piore y Sabel se refirieron a la evidencia de Japón, Alemania e Italia donde las empresas con especialización flexible están normalmente agrupadas.

Bull, Pitt, y Szarka (1991), compararon tres comunidades europeas textiles, Como (Italia), Leicester (Reino Unido) y Lyon (Francia), elegidos sobre la base de la gran cantidad de pequeñas empresas que operan en ellos y utilizando los datos recogidos por los autores a través de un cuestionario postal y entrevistas individuales. Descartando el enfoque más tradicional de pequeñas empresas, que los ve como unidades individuales y autónomas, en el documento se aplicó el modelo de distrito industrial en las tres áreas. Por lo tanto, las empresas fueron estudiadas en su relación tanto entre sí, como con toda la comunidad. Se encontró que cada área posee una organización industrial muy distinta, y un rendimiento muy variable. El estudio mostró claramente la fuerza de un distrito industrial como Como, así como las dificultades que enfrenta una comunidad industrial (en oposición a un distrito) como Leicester. Las malas relaciones entre las empresas locales podrían explicar estas dificultades, al igual que la posición de toda la comunidad de empresas frente a los agentes externos, especialmente las grandes cadenas minoristas.

En los Estados Unidos, los nuevos distritos industriales similares a los de Europa, surgieron en la electrónica y en la computación en Silicon Valle y en el complejo aeroespacial, electrónica y comunicaciones relacionados con lo militar en el condado de Orange, y la ruta 128 al oeste de Boston. La tasa de crecimiento de estas regiones ha sido fenomenal, el Condado de Orange aumento la tasa de empleo a un ritmo de 186% entre 1970 y 1990, más de tres veces la tasa nacional.

Estos casos exitosos han sido motivo que los investigadores sajones también tengan interés por las agrupaciones industriales (Feldman, Francis y Bercovitz, 2005; Harrison, 1992; Porter, 1998; 2000; Quadrio-Curzio y Fortis, 2002).

Porter (1998), define al cluster como concentraciones geográficas de empresas interconectadas, suministradores especializados, proveedores de servicios, empresas de sectores afines e instituciones conexas (por ejemplo, universidades, institutos de normalización, asociaciones comerciales) que compiten pero que también cooperan. En su carácter de masas críticas de inusual éxito competitivo en áreas de actividad determinadas, es una actividad característica de todas o casi todas las economías nacionales, regionales e incluso metropolitanas, en especial las de los países más avanzados.

El concepto unificador sostiene que las empresas, a menudo con la ayuda de los gobiernos regionales y las asociaciones de comercio conscientemente forman una red para resolver los problemas de los ciclos y el exceso de capacidad y responder a las nuevas demandas de flexibilidad (Amin y Thrift, 1992). En la versión americana, la rigidez en las viejas ciudades industriales ciudades tiende a alentar la formación de estas aglomeraciones para arraigarse en nuevas localizaciones (Markusen 1991; Scott 1988b; Storper y Walker 1989).
Estudios realizados encuentran efectos positivos en los distritos industriales, mientras que otros señalan efectos negativos.

Baptista and Swann (1998), investigaron las posibilidades de innovación de las empresas ubicadas en grupos o regiones industriales. Estudiaron 248 empresas de manufactura en el Reino Unido y realizaron un análisis estadístico de la asociación entre la probabilidad de la innovación y la fuerza del clúster. Los resultados encontrados mostraron que una empresa tendrá más probabilidad de innovar si el empleo de su propio sector en la región es fuerte, mientras que el efecto de un nivel fuerte de empleo no tiene efecto significativo en las empresas que no se localizan en cluster.

Bell (2005), investigó los mecanismos de los efectos de las redes desde un clúster en el rendimiento de la empresa, así como estudió la influencia de estos mecanismos diferentes en las empresas ubicadas dentro y fuera del cluster. Destacó la importancia de modelar simultáneamente redes múltiples que pueden influir diferencialmente en los resultados de las empresas. Modeló la capacidad de innovación de las empresas de fondos de inversión canadienses en función de su ubicación geográfica dentro o fuera del cluster industrial de Toronto y de su centralidad en las redes de relaciones gerenciales e institucionales. Concluyó que la localización en el cluster industrial, así como la centralidad en la red de relaciones empresariales favorece la innovación empresarial, mientras que la centralidad en la red de relaciones institucionales no lo hace.

El conocimiento en una empresa es un enfoque reciente para entender la relación entre las capacidades y los resultados de la misma. En concreto, este enfoque sugiere que la generación, acumulación y aplicación del conocimiento, pueden ser el origen de un mayor rendimiento.

Otras investigaciones han conceptualizado el conocimiento de la organización en términos de existencias y flujos de conocimiento en la empresa.

La relación entre existencias y flujos de conocimiento en la organización y el desempeño de la empresa en la industria de la biotecnología fue abordado por Deeds y Decarolis (1999). En su investigación señalaron a la ubicación geográfica de la empresa, las alianzas con otras instituciones y organizaciones y los gastos en Investigación y Desarrollo como representativos de flujos de conocimiento mientras que los productos en desarrollo, citas y patentes de la empresa son indicativos de las existencias de conocimiento. A través de un análisis factorial, desarrollaron una medida agregada de la ubicación de varias variables. Un modelo de regresión sugirió que la ubicación es un predictor significativo del desempeño de la empresa como son los productos en desarrollo y citas.

Molina-Morales (2001), investigó el desempeño de las empresas del distrito industrial en forma comparativa usando el caso de la industria cerámica española. Los resultados empíricos de la investigación mostraron un rendimiento significativamente mayor en las empresas del distrito industrial en comparación con las empresas externas.

Glassmeier (1991), sostuvo que el énfasis actual en las redes de producción como único depósito de innovación potencial tecnológica que se realiza a través de la cooperación y no de la competencia entre las empresas, carece de una apreciación detallada de las redes históricas, y en particular de su frágil carácter en tiempos de crisis económica. Señaló que si bien las redes pueden y deben promover la innovación dentro del marco tecnológico existente, la experiencia histórica sugiere que su estructura fragmentada y atomizada está sujeta a la desorganización y desintegración durante los períodos de cambio tecnológico. Estudió el caso de la industria relojera Suiza y en contra de investigaciones anteriores que se basaron en modelos de competencia oligopólica para explicar la causa de como éste país perdió el control de la industria relojera mundial, concluyó que la experiencia suiza debe entenderse en función

de cómo la organización de la producción, la industria, la cultura, y la sociedad experimentan grandes dificultades para adaptarse a cambios tecnológicos.

Un aspecto que no se toma en cuenta en las economías de aglomeración, es que las empresas además de beneficiarse de la externalidad también contribuyen a ésta, (Shaveer, 2000). Esta idea sugiere que si las empresas son heterogéneas se diferencian en los beneficios netos que reciben de la aglomeración. Shaveer (2000), sostienen que las empresas con las mejores tecnologías, capital humano, programas de formación, proveedores o distribuidores obtendrán poco, pero competitivamente perderán cuando sus tecnologías, empleados, y el acceso a sus programas de apoyo se difunde hacia sus competidores. Por lo tanto, estas empresas tienen poca motivación a agruparse geográficamente a pesar de la existencia de las economías de aglomeración. Por el contrario, las empresas con más débiles tecnologías, capital humano, programas de capacitación, proveedores o distribuidores, tienen poco que perder y mucho que ganar; por lo tanto, estas empresas están motivadas a pertenecer a un clúster. Indicaron como resultado importante que cuando las empresas son heterogéneas la selección de una aglomeración sería adversa y apoyaron lo anterior examinando la elección del lugar y la supervivencia de las inversiones extranjeras recientes en las industrias manufactureras en los Estados Unidos.

Stuart y Sorenson (2000), sostuvieron que las industrias se agrupan porque los empresarios tienen dificultades para aprovechar los lazos sociales necesarios para movilizar recursos esenciales cuando residen lejos de esos recursos. Los factores que permiten la iniciativa empresarial de alta tecnología, no promueven necesariamente los resultados de la empresa. En los análisis empíricos, investigaron los efectos de la proximidad geográfica a las empresas de biotecnología establecidas, fuentes de conocimientos de biotecnología (mano de obra altamente cualificada) y los capitalistas de riesgo en las tasas específicas de ubicación y en el funcionamiento de las empresas. Concluyeron que las condiciones locales que promueven la creación de nuevas empresas difieren de las que maximizan el rendimiento de las empresas de reciente creación.

Feser y Bergman (2000), señalan cuatro factores recurrentes en la literatura sobre clusters: entrelazamientos formales entre compradores y proveedores, proximidad geográfica, instituciones compartidas que tienen interés en la industria local y evidencia de cooperación informal y competencia simultánea. Identificaron 23 clusters de industria de manufactura en Estados Unidos, pero no encuentran consistencia con los patrones de aglomeración, lo cual

sugiere que si bien la metodología utilizada detecta elementos de aglomeración y formación de redes no identifica los clusters de forma consistente.

Hill y Brennan (2000), proporcionaron una metodología cuantitativa para identificar las fuerzas que conforman a los clusters. En Cleveland-Akron identificaron diez clusters que aglomeran 18% del empleo regional.

Colgan y Baker (2003) establecieron una metodología cualitativa e identificaron siete grupos de cluster en Maine. Su metodología es empírica y no es robusta en las comparaciones interregionales.

El Istituto Nazionale di Statistica Italiano, ISTAT, definió en 1996 una metodología cuantitativa que intenta aproximar la esencia de la definición del distrito industrial y sus características básicas. La metodología consta de dos partes: la identificación de la unidad territorial de análisis y la identificación de los distritos industriales. La unidad territorial de referencia para el estudio del distrito es el Sistema Local de Trabajo (ISTAT 1997). Una vez obtenidos los Sistemas Locales de Trabajo, se utiliza una batería de coeficientes de concentración anidados, de naturaleza socioeconómica, para identificar cuáles de estas unidades muestran características de distrito industrial (ISTAT 1996; Sforzi e Lorenzini 2002).

5.3. Datos

En este trabajo se utilizó la base de datos del Instituto Nacional de Geografía y Estadística (INEGI) organismo encargado en México de proveer y analizar estadísticas sociodemográficas, económicas, censos, etc. De la base de datos del INEGI se obtuvo la información dela cantidad de habitantes, la ubicación geográfica, expresada en términos de latitud y longitud y el código postal de 2358 municipios del país.

Como en México no se cuenta con información de los flujos de viajes al trabajo se aplicó el modelo de radiación (Simini 2012) para estimar éstos flujos de viajes entre los 2358 municipios.

Se utilizó también la base de datos del Directorio de empresas registradas en la Secretaría de Economía, la cual en México, es el organismo público federal responsable de formular y conducir las políticas de industria, comercio exterior, interior, abasto y precios del país.

El Directorio de Empresas que proporciona este organismo especifica el nombre de cada empresa, su ubicación geográfica, expresada en términos de código postal, la cantidad de trabajadores empleados en ella y la actividad económica que realiza cada empresa. A partir de éste Directorio se identificaron y clasificaron las empresas de acuerdo a las actividades que realizan. Específicamente se identificaron las empresas con actividades industriales de manufactura. La clasificación fue realizada en base a la Metodología del ISTAT, que se muestra en la Tabla 15.

5.4. Metodología

El modelo de radiación propuesto por Simini (2012) es un proceso estocástico que captura las decisiones de movilidad local y que ayuda a derivar los flujos de viaje al trabajo, y que dado su naturaleza libre de parámetros puede ser aplicado en áreas donde no existen medidas de movilidad previa (Simini 2012). Para aplicar este modelo es necesario contar con la cantidad de habitantes de cada municipio y su ubicación geográfica.

Una vez realizado el proceso de radiación se obtiene la matriz estimada origen-destino de viajes al trabajo entre los 2358 municipios de los que se compone la muestra.

Para llevar a cabo el proceso de formación morfológica de los sistemas locales de trabajo se aplicó el método de clúster jerárquico. Se midió la similitud de los destinos entre dos municipios en función de la correlación que presentaban los orígenes y los destinos de las personas.

Se utilizó el método del vecino más cercano. La regionalización produjo 312 Sistemas Locales de Trabajo, (SLT). La cantidad óptima de los Sistemas Locales de Trabajo fue inferida de acuerdo a la medida del grado de pertenencia de un objeto a su clúster, con base en el promedio de la distancia entre este objeto y todos los objetos del clúster al que pertenece, comparada con la misma medida calculada para el siguiente clúster más cercano.

Una vez obtenidos los 312 SLT se utilizó la metodología del ISTAT (1996), a fin de verificar la composición laboral de cada uno de ellos. En la base datos utilizada se especifica la cantidad de trabajadores en cada empresas y las actividades de cada empresa. Una vez formados los SLT se identificaron las industrias que los constituían, sus zonas postales y la actividad a la que se dedica cada una de las empresas agrupadas en los 312, con lo que se determinó la composición de cada SLT.

Se utilizaron los coeficientes de concentración anidados propuestos por la Metodología del ISTAT que reflejan las características socio económicas de los 312 SLT y permitieron identificar cuáles de ellos presentaron características de distrito industrial.

Los índices utilizados muestran el grado de aglomeración de empresas con actividades directamente relacionadas con la manufactura de un proceso productivo en particular.

Tabla 15. Agregados sectoriales manufactureros originales del ISTAT (1997)

Agregado	**Descripción**
Industria de la alimentación, bebidas y tabaco (1).	Industria de productos alimenticios y bebidas. Industria del tabaco.
Industria textil y de la confección (2)	Industria textil. Industria de la confección y de la peletería
Industrias del cuero y del calzado (3)	Preparación curtido y acabado del cuero; fabricación de artículos de marroquinería y viaje; artículos de guarnicionería talabartería y zapatería. Industria de la madera y del corcho, excepto muebles; cestería y espartería.
Papel, edición y artes gráficas (4)	Edición. Artes gráficas y actividades de los servicios relacionados con las mismas.
Productos para la casa	Fabricación de otros productos minerales no Metálicos. Fabricación de muebles. Fabricación de artículos de joyería, orfebrería, platería y artículos similares. Fabricación de instrumentos musicales.
Metalurgia	Metalurgia. Coquerías, refino de petróleo y tratamiento de combustibles nucleares.

Tabla 15. Agregados sectoriales manufactureros originales del ISTAT (1997). *Continuación*

Agregado	Descripción
Industria Mecánica	Industria de la construcción de maquinaria y equipo mecánico. Fabricación de máquinas de oficina y equipos Informáticos. Fabricación de maquinaria y material Eléctrico. Fabricación de material electrónico; fabricación de equipo y aparatos de radio, televisión y comunicaciones. Fabricación de equipo e instrumentos médico-quirúrgicos, de precisión, óptica y relojería.
Material de transporte	Fabricación de vehículos de motor, remolques y semirremolques. Fabricación de otro material de transporte. Industria del papel.
Otras industrias manufactureras	Fabricación de artículos de deporte Fabricación de juegos y juguetes Otras industrias manufactureras diversas Reproducción de soportes grabados Fundición de metales Fabricación de productos metálicos, excepto maquinaria y equipo

5.5. Resultados

De los 312 SLT que se obtuvieron al agrupar los 2358 municipio en la regionalización inicial, solamente 36 tienen características de distrito industrial.

En un inicio se clasificaron 100 SLT especializados en manufactura. De éstos 100 solo se identificaron 44 SLT como empresas pequeñas y medianas.

A continuación se identificó la industria preponderante en cada SLT y se verificó que las Pymes fueran las principales empleadores en el SLT. Al final de proceso se identificaron 36 Distritos Industriales (SLT), en los que además las Pymes son los mayores productoras en su especialidad.

Ya que no se dispone en México de un censo industrial, se utilizó la base provista por la Secretaría de Economía para clasificar a todas las empresas registradas en ella. En esta base de datos se especifica el número de trabajadores, con los datos agrupados en rangos de empleados. Por lo que se utilizó el punto medio de cada rango para aproximar la cantidad de trabajadores en la empresa. Para las empresas con más de 250 trabajadores sea aproximó a 500 trabajadores.

Los 36 distritos industriales se extienden por todo el país como se muestra en la Figura 26.

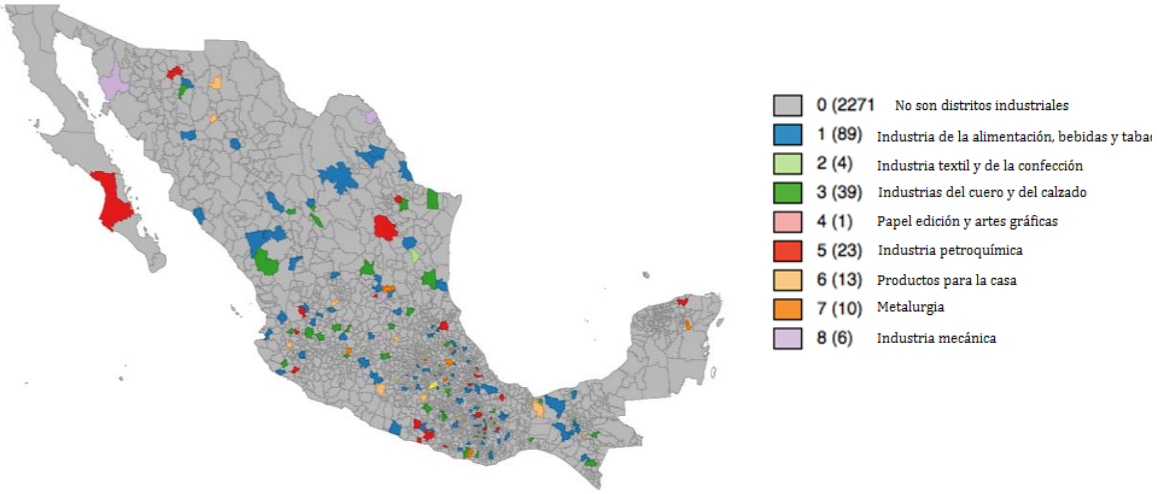

Figura 26. Distritos Industriales de Manufactura en México

5.6. Conclusiones.

Los beneficios de la aglomeración son numerosos y ampliamente documentados. A partir de la rica evidencia empírica que muestra que las empresas no compiten ni aprenden de manera aislada, y que el contexto en que operan es importante para su desempeño, se ha producido una revalorización del papel que juega el entorno regional. Esto es importante porque una parte sustancial de la competitividad de las empresas se genera al exterior de las mismas, en las relaciones que éstas logran establecer con su entorno y, en particular, con otras empresas. Además, las pequeñas empresas ubicadas en distritos industriales cuentan con mayores capacidades para superar algunas de las restricciones que enfrentan, como la falta de

habilidades especializadas o las dificultades de acceso a la información, al mercado, al crédito y/o a los servicios externos

Resulta sorprendente que en los agregados manufactureros estudiados solo 36 tengan características de Distritos Industriales, lo que supone no se están aprovechando las ventajas que ofrecen las aglomeraciones productivas de las Pymes.

Se debería planificar desarrollo regional en base a distritos industriales y canalizar recursos del FONDO PYME, instrumento gubernamental que busca apoyar a las Micros, Pequeñas y Medianas Empresas con el propósito de promover el desarrollo económico nacional a través del otorgamiento de apoyo, hasta que se alcanzaran su maduración.

Es importante resaltar la importancia del modelo de radiación utilizado para estimar los flujos de viajes ante la ausencia de datos en México y la conveniencia de utilizar los coeficientes de concentración anidados propuestos por la Metodología del ISTAT que permiten identificar de los clusters formados cuales presentan características de distrito industrial.

6. Conclusiones

En la presente Tesis Doctoral se realizaron tres ensayos referentes a la movilidad laboral y su impacto en los mercados laborales y sus formas de agrupación:

Se realizó un análisis de cluster jerárquico, método de vinculación de promedio intra-grupos, con lo que a partir de la matriz de viajes origen-destino, se pudieron formar grupos o regiones homogéneas por género, por industria y por actividad con base a la cantidad de personas que comparten destinos en su viaje al trabajo y lugares de residencia.

Con el objetivo de medir y comparar la movilidad y los procesos de articulación y expansión de los mercados laborales locales (MLL), se analizaron cuatro variables para los años 2001-2011: autonomía de la oferta, autonomía de la demanda, empleo y cantidad de trabajadores, lo que permitió identificar las características de los mercados laborales por sector.

Utilizando el enfoque de la regresión geográficamente ponderada se analizó el grado de impacto que tiene sobre el crecimiento del empleo el capital humano de las personas que habitan en cada municipio. El capital humano fue capturado por la proporción de personas de cada municipio que cuentan con las características expresadas en las variables independientes seleccionadas. Se encontraron efectos diferenciales en las variables explicativas en el crecimiento del empleo a lo largo del espacio.

Para México se estimaron los flujos de viajes con el modelo de radiación y se realizó un análisis de cluster jerárquico para determinar los Sistemas Locales de Trabajo, (SLT), para posteriormente mediante los coeficientes de concentración anidados propuestos por la Metodología del ISTAT identificar cuáles de ellos presentaron características de distrito industrial.

Se puede concluir en cuanto a la metodología utilizada:

1. Los flujos de viaje al trabajo son un buen criterio de agrupación de los municipios.
2. El análisis de conglomerados es una técnica eficiente para formar los grupos.
3. El uso de índices como la autonomía de la oferta y la autonomía de la demanda permiten caracterizar a los MLL.
4. La regresión geométricamente ponderada permite detectar variaciones espaciales que el modelo de regresión

5. El modelo de radiación es una herramienta útil para estimar flujos en ausencia de datos.
6. Obtener los Sistemas Locales de Trabajo mediante un análisis clúster parece adecuado.
7. El uso de los coeficientes de concentración anidados de la metodología del ISTAT son de gran utilidad para la identificación de distritos industriales, además de su facilidad de uso.

Se puede concluir desde el enfoque de regionalización:

1. Es importante entender el territorio a partir del funcionamiento de la economía y, en forma particular, a partir del comportamiento espacial del empleo.

2. Investigaciones referentes a la movilidad de los trabajadores son de la mayor importancia porque permiten realizar estudios de corte socioeconómico y del comportamiento de los mercados de trabajo, con especial atención al desempleo.

3. Investigaciones referentes a la movilidad de los trabajadores permiten llevar a cabo estudios comparativos del desempeño económico y pronósticos regionales, evaluar la competitividad y las disparidades entre regiones, e identificar territorios frágiles que requieren apoyos especiales.

4. Investigaciones referentes a la movilidad de los trabajadores proporcionan instrumentos adecuados para la planificación y desarrollo regional.

A la luz de la investigación realizada en la presente Tesis Doctoral, se proponen algunas posibles líneas de investigación:

1. Es importante reconocer que la cercanía entre las empresas localizadas al interior de un distrito industrial tiene ventajas sustanciales, pero también presenta dificultades que pueden ser no despreciables como puede ser la saturación de mercados y la sobre especialización. Investigar sobre las ventajas y desventajas presentes en los distritos industriales y caracterizar la variante “Mexicana” del distrito industrial natural sería importante para el desarrollo propio de las Pymes en el país.

2. Además es muy posible que en México encontremos que de los distritos industriales identificados en esta Tesis son forzados por una decisión en las esferas

gubernamentales de establecer una cadena productiva en una región que se considera propicia para el desarrollo local y regional, algunas veces por mera conveniencia política, por lo que investigar el desarrollo de distritos naturales contra la evolución de distritos forzados puede mostrar deficiencias o bondades importantes.

3. Investigar la evolución de los distritos industriales conformados por la industria maquiladora que fue motor económico de la zona norte del país.

4. Investigar y determinar con los viajes al trabajo los MLL para México con la metodología utilizada en esta Tesis Doctoral.

5. La zona metropolitana de la ciudad de México es el área metropolitana formada por el Distrito Federal y 60 municipios aglomerados uno de ellos en el Estado de Hidalgo, los restantes del Estado de México. Según los resultados del censo elaborado por el INEGI en el año 2010 esta zona contaba con una población de alrededor de 20 millones de habitantes (tan solo en el Distrito Federal son 8 851 080 habitantes). Según datos de la ONU en el año 2012, es la tercera aglomeración humana más poblada del mundo. En esta área metropolitana por políticas gubernamentales se fueron construyendo viviendas unifamiliares cada vez más alejadas del núcleo urbano, formándose una ciudad-región por lo que sería de interés investigar la organización espacial del mercado de la vivienda y su heterogeneidad con el fin de coadyuvar en la política habitacional.

Bibliografía

ABRAMSON, I. (1982). On bandwidth variation in kernel estimates - a square root law-. Annals of Statistics.(9), 168-176.

ACEMOGLU, D., & ANGRIST, J. (2001). How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws. 15, 9-74.

AGGARWAL, C., WOLF, J., YU, P., & PROCOPIUS, C. a. (1999). Fast algorithms for projected clustering. In proceeding of the 1999 ACM SIGMOD international conference on management of data, 61-72. Philadelphia: ACM Press.

AKAIKE, H. (1974). A new look at the statistical model identification. IEEE Transaction on Automatic Control, 19(6), 716-723.

ALI, K., PARTRIDGE, M., & OLFERT, M. (2007). Can geographically weighted regressions improve regional analysis and policy making? International Regional Science Review, 30(3), 300-329.

AMARA, N., & LANDRY, R. O. (2005). Milieux Innovateurs: Determinants and Policy Implications. European Planning Studies, 13(6), 930-965.

AMIN, A., & THRIFT, N. (1992). Neo-Marshallian nodes in global rietworks. International Journal of Urban and Regional Research, 16, 571-587.

ANABLE, J. (2005). “Complacent Car Addicts” or “Aspiring Enviromentalists”? Identifying travel behaviour segments using attitude theory. Transport Policy, 65-78.

ANDERBERG, M. (1973). Cluster analysis for applications. New York: Academic Press.

ATKINSON, P., & KUBY, M. (2011). The Geography of advance transit- oriented development in metropolitan Phoenix, Arizona 2000-2007. Journal of Transport Geography, 19, 189-99.

BACAO, F., LOBO, V., & and PAINHO, M. (2005). Applying genetic algorithms to zone design. Soft computing, 9(5), 341-348.

BANFIELD, C. (1976). Statistical algorithms: Algorithm AS102: Ultrametric distances for a single linkage dendrogram. Applied Statistics, 25(3), 313-315.

BAPTISTA, R., & SWANN, P. (1998). Do firms in clusters innovate more? Research Policy, 27, 525-540.

BASS, F., TIGER, D., & and LONSDALE, R. (1968). Market Segmentation: Group Versus Individual Behavior. Journal of Marketing Research., V, 264-70.

BEANE, T., & and ENNIS, D. (1987). Market Segmentation: A review. European Journal of Marketing, 21, 20-42.

BECKMAN, J., & GOULIAS, K. (2008). Imigration, residential location, car ownership, and commuting behavior: a multivariate latent class analysis from California. Transportation , 35(5), 655-671.

BELL, G. (2005). Clusters, networks, and firm innovativeness. Strategic Management Journal, 26, 287-295.

BELLANDI, M. (1989). The industrial district in Marshall. In E. G. Bamford (Ed.), In Small firms and industrial districts in Italy (pp. 136-152). London: Routledge.

BERDEGUE, J., JARA, B., FUENTEALBA, R., TOHA, J., MODREGO, F., & SCHEJTMAN, A. A. (2011). Territorios Funcionales en Chile. Documento de trabajo. Programa dinámicas territoriales rurales. Documento de trabajo. Programa dinámicas territoriales rurales- Rimisp-centro latinoamericano para el desarrollo rural.

BLOOM, J. (2005). Market Segmentation. A Neural Network Application. Annals of Tourism Research, 32(1), 93-111.

BOBISUD, H., & and BOBISUD, L. (1972). A metric for clasification. (Taxon, Ed.) 21, 607-613.

BOERMANS M. A. ROELFSEMA H. and ZHANG, Y. (2011). Regional determinants of FDI in China: a factor-based approach. Journal of Chinese Economic and Business Studies, 9(1), 23-42.

BOIX, R., & and VENERI, P. (2010). Retrieved from Metropolitan areas in Spain and Italy.http://ddd.uab.cat/pub/worpap/2010/hdl_2072_87965/wpierm0901.pdf. (I.d.Barcelona, Producer)

BOIX, R., & VENERI, P. (2010). Metropolitan Areas in Spain and Italy. Working paper, Institut d'Estudis Regionals i Metropolitans de Barcelona.

BOIX., R., & and GALLETTO, V. (2004). Identificación de sistemas locales de trabajo y distritos industriales en España, Dirección General de Política de la Pequeña y Mediana Empresa. MITYC.

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika(71), 353-360.

BRUNSDON, C., FOTHERINGHAM, A., & and CHARLTON, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Analysis(28), 281-298.

BRUNSDON, C., FOTHERINGHAM, A., & AND CHARLTON, M. (1998). Spatial nonstationarity and autoregressive models. Environment and Planning A, 30, 9957-73.

BULL, A., PITT, M., & SZARKA, J. (1991). Small firms and industrial districts, structural explonations of small firm viability in three countries. Entrepreneurship and Regional Development, 3, 83-99.

CAREY, G. (1966). The regional interpretation of Manhattan population and housing patterns through factor analysis. Geographical Review, 551-569.

CARMICHAEL, C. (1978). Local labour market analysis: its importance and a possible aproach. Geoforum, 9, 127-148.

CASADO, D., BERNABEU, L. M., & REVUELTA, F. (2010). Los mercados locales de trabajo españoles. Una aplicación del nuevo procedimiento británico. La ciudad metropolitana en España: Procesos urbanos en los inicios del siglo XXI. Madrid: Thompson-Civitas.

CASADO-DÍAZ, J. (2000). Local labour market areas in Spain: A case study. Regional Studies, 934(9), 843-856.

CATTELL, R. (1965). Factor Analysis and its role in Research. Biometrics, 21(2), 405-435.

CHARLTON, M., FOTHERINGHAM, A., & BRUNSDON, C. (1997). The geography of relationships: an investigation of spatial nonstationarity. In C. D. Bocquet-Appel J-P. (Ed.), Spatial analysis of biodemographic data (pp. 23-47). Montrouge: John Libbey Eurotext.

CHAUDHURI, H. (2005). Understanding the Interrelationship Between Regional Differences and Material Aspiration in the Context of Indian Diversity: Results of an Exploratory Study. Asia Pacific Journal of Marketing and Logistics, 17(4), 3-14.

CLASIFICACIÓN NACIONAL DE ACTIVIDADES DE ESPAÑA. (1993). CNAE. www.ine.es

CLASIFICACIÓN NACIONAL DE OCUPACIONES DE ESPAÑA. (1994). CNO. www.ine.es

CLEVELAND, W. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association(74), 829-836.

CLIFF, A., & HAGGETT, P. (1970). On the effciency of alternative aggregations in region-building problems. Enviroment and Planning, 16(3), 285-294.

COLGAN, C., & BAKER, C. (n.d.). A Framework for Assessing Cluster Development. Economic Development Quaterly, 17(4), 352-366.

CONOVER, W., & IMAN, R. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. The American Statistician, 35(3), 124-129.

CONSTANTINE, A., & GOWER, J. (1978). Graphical representantion of asymmetric matrices. Applied Statistics, 27, 297-304.

CONSTANTINESCU, P. (1966). The classification of a set of elements with respect to a set of properties. The computer Journal, 8(4), 352-357.

COOMBES, M., GREEN, A., & OPENSHAW, S. (1986). An efficient algorithm to Generate Official Statistical Reporting Areas: The case of the 1984 Travel-to-Work Areas Revision in Britain. Journal of the Operational Research Society, 27(10), 943-953.

COOMBES, M., GREEN, A., & OWEN, D. (1988). Substantive issues in the definition of " localities": Evidence from sub-group Local Labor Market Areas in the West Midlands. Regional Studies, 22(4), 303-318.

CÖRVERS, F., HENSEN., M., & BONGAERTS, D. (2009). Delimitation and Coherence of Functional and Administrative Regions. Regional Studies, 43(1), 19-31.

COVA, T., & CHURCH, R. (2000). Contiguity Constraints for Single-Region Site Search Problems. Geographical Analysis, 32(4), 306-329.

CRAVEN, P., & WAHBA, G. (1979). Smoothing noisy data with spline functions. Numerische Mathematik.(31), 377-403.

CUI, G., & LIU, Q. (2000). Regional Market Segments of China: Opportunities and Barriers in a big emerging market. Journal of Consumer Marketing, 17(1), 55-72.

DECAROLIS, D., & DEEDS, D. (1999). The Impact of Stocks and Flows of Organizational Knowledge on Firm Performance: An Empirical Investigation of the Biotechnology Industry. Strategic Management Journal, 20(10), 953-968.

DÍAZ DE RADA, V. (2002). Técnicas de Anális Multivariante para investigación Social y Comercial. Ejemplos Prácticos Utilizando SPSS. Ra-MA.

DIEHR., G. (1971). Clustering with finite number of potential centroids: Upper and lower bounding algorithms. Second Annual Meeting of the Classification Society.

DINIZ, C., & RAZAVI, M. (1994). Emergence of new industrial districts in Brazil: Sao Jose dos Campos and Campinas cases. Brazil: Universidad Federal de Minas Gerais, CEDEPLAR.

DUQUE, J., RAMOS, R., & SURINACH, J. (2006). Supervised Regionalization Methonds: A survey. International Regional Science Review, 30(3), 195-220.

ECKEY, H., KOSFELD, R., & TÜRK, M. (2007). Spatial Economic Analysis. Regional Convergence in Germany: Geographically Weighted Regression Approach, 2(1).

EVERITT, B. (1993). Cluster analysis. (3rd ed.). New York, Toronto: Halsted Press.

FELDMAN, M., FRANCIS, J., & BERCOVITZ, J. (2005). Creating a Cluster while Bulding a Firm: Entrepreneurs and the Formation of Industrial Custers. Regional Studies, 39(1), 129-141.

FERIA, M. (2008). Un Ensayo Metodológico de Definición de las Áreas Metropolitanas en España a Partir de la Variable Residencia-Trabajo. Investigaciones Geográfica, 46, 49-68.

FESER, E., & BERGMAN, E. (2000). National Industry Cluster Templates: A Framework for Applied Regional Cluster Analysis. Regional Studies, 34(1), 1-19.

FINEBERG, D., GILMORE, R., KRANTZ, J., LIANES, M., MILLER, R., MANN, U., et al. (1993). The biopharmaceutical industry in New Jersey: Prescriptions forregional economic development. Rutgers University.

FISHER, M. (1980). Regional Taxonomy: A comparison of some hierarchic and no-hierarchic strategies. Regional Science and Urban Economics, 10, 503-537.

FLOREK, K., LUKASZEWICZ, J., STEINHAUS, H., & and SUBRZYCKI, S. (1951). Sur la liaison et la division des points d'un ensemble fini. Colloquium Mathematicum, 2, 282-285.

FOTHERINGHAM, A. S., BRUNDSON, C., & CHARLTON, M. (2002). Geographically Weithed Regression. John Wiley & Sons Ltd.

FOTHERINGHAM, A., BRUNSDON, C., & CHARLTON, M. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environment and Planning A(30), 1905-1927.

FOTHERINGHAM, A., CHARLTON, M., & BRUNSDON, C. (1996). The geography of parameter space: an investigation into spatial nonstationarity. International Journal of GIS(10), 605-27.

FOTHERINGHAM, A., CHARLTON, M., & BRUNSDON, C. (1997). Two techniques for exploring nonstationarity in geographical data. Geographical Systems(4), 59-82.

FUNKE, M., & NIEBUHR, A. (2005). Regional Geographic Research and Development Spillovers and Economic Growth: Evidence from West Germany. Regional Studies, 39, 143-153.

GARFINKEL, R., & NEMHAUSER, G. (1970). Optimal political districting by implicit enumeration techniques. Management Science, 16(8), B495-B508.

GLASMEIER, A. (1991). Technological discontinuities and flexible production networks: The case of Switzerland and the world watch industry. Research Policy, 20, 469-485.

GODDARD, J. (1970). Functional Regions within the City Centre: A Study by Factor Analysis of Taxi Flows in Central London. Transactions of the Institute of British Geographers(49), 161-182.

GOLDSTEIN, H. (1987). Multilevel models in educational and social research. London: Oxford University Press.

GOODMAN, E. (1989). Introduction: The political economy of the small firm in Italy. In E. G. Bamford (Ed.), In small firms and industrial districts in Italy (pp. 1-3). London: Routledge.

GOODMAN, E. (1989). Introduction: The Political Economy of the small firm in Italy. In small firms and industrial districts in Italy. (Routledge, Ed.) London: E. Goodman and J. Bamford.

GORDON, A. (1987). A review of hierarchical classification. Journal of the Royal Statistical Society. Series A (General), 150(2), 119-137.

GORDON, A. (1996). A survey of constrained classification. Computational Statitics & Data Analysis, 21(1), 17-29.

GOVIND, R. N., & SUN, W. (2014). Geographically Varying Effects of Weather on Tabacco Consumption: Implication for Health Marketing Initiatives. Health Marketing Quarterly, 31, 46-64.

GOWER, J. (1967). A comparison of some methods of cluster analysis. Biometrics, 23(4), 623-637.

GOWER, J., & ROSS, G. (1969). Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 18(1), 54-64.

GREEN, A., COOMBES, M., & OWEN, D. (1986). Gender-specific Local Labour Market Areas in England and Wales. Geoforum, 17(3), 339-351.

GROSE, D., & BRUNSDON, C. H. (n.d.). Introduction to Geographically Weighted Regression (GWR) and to Grid Enabled GWR. Lancaster University.

GUASTELLA, G., & VAN OORT, F. (2015). Regional Heterogeneity and Interregional Research Spillovers in European Innovation: Modelling and Policy Implications. Regional Studies.

HALEY, R. (1968). Benefit Segmentation: A Decision-oriented Research Tool. Journal of Marketing, 32, 30-35.

HANSEN, H., & WINTHER, L. (2015). Employment growth, human capital and educational levels: uneven urban and regional development in Denmark 2002-2012. Geografisk Tidsskrift-Danish Journal of Geography.

HARRISON, B. (1992). Industrial Districts: Old Wine in new bottles;. Regional Studies, 26, 469-483.

HARRISON, B. K., & GANT, J. (1996). Innovative Firm Behavior and Local Milieu: Exploring the Intersection of Agglomeration, Firm Effects, and Technological Change. Economic Geography, 72(3), 233-258.

HARTIGAN, J. (1967). Representation of similiraty matrices by trees. Journal of the American Statical Association, 62(320), 1140-1158.

HASTI, E. T., & TIBSHIRANI, R. (1990). Generalized additive models. London: Chapman & Hall.

HAUSER, R. (1970). Context and consex: a cautionary tale. American Journal of Sociology., 75, 645-64.

HAWKINS, D., & CONEY, K. (2001). The Influence of Geographic Subcultures in the United States. Nature of Geographic Subcultures, 713-717.

HEMMASI, M. (1980). The Identification of Functional Regions Based on Lifetime Migration Data: A Case Study of Iran. Economic Geography, 56(3), 223-233.

HESS, S., & SAMUELS, S. (1971). Experiences with a sales districting model-criteria and implementation. Management Science, 18(14), 41-54.

HESS, S., WEAVER, J., SIEGFELD, H., WHELAND, J., & ZITLAU, P. (1965). Nonpartisan political redistricting by computer. Operations Research, 13(6), 998-1006.

HILL, E., & BRENNAN, J. (2000). A Methodology for Identifying the Drivers of Industrial Clusters: The Foundation of Regional Competitive Advantage. Economic Development Quarterly, 14(1), 65-96.

HOAGLIN, D., & WELSCH, R. (1978). The hat matrix in regression and ANOVA. The American Statistician(32), 17-22.

HOERL, A., & KENNARD, R. (1970a). Ridge regression: biased estimation for non-orthogonal problems. Technometrics(12), 55-67.

HOERL, A., & KENNARD, R. (1970b). Ridge regression: applications to non-orthogonal problems. Technometrics(12), 69-82.

HOFSTEDE, F., STEENKAMP, J. E., & WEDEL, M. (1999). International market segmentation based on consumer-product relations. Journal of Marketing Research, 36(1), 1-17.

HOGNI, K., & LARS, W. (2015). Employment growth, human capital and educational levels: uneven urban and regional development in Denmark 2002-2012. Geografisk Tidskrift.

HOLSMAN, A. (1980). Higher-order factor analysis and its application to transport networks. The Professional Geographer, 32(2), 192-198.

HOLSMAN, A. (1980). Higher-Order Factor Analysis and its application to transport networks. The Professional Geographer, 32(2), 192-198.

HONGMIAN, G., & WHEELER, J. O. (2002). The Location and Suburbanization of Business and Professional Services in the Atlanta Metropolitan Area. Growth and Change, 33(3), 341-369.

HOOVER, E. (1937). Spacial price discrimination. Review of Economic Studies, 4, 182-191.

HORN, M. (1995). Solution techniques for large regional partitioning problems. Geographical Analysis, 27(3), 230-248.

HUANG, B., WU, B., & BARRY, M. (2010). International Journal of Geographical Information Science. 24(3), 383-401.

ILLERIS, S., & PEDERSON, P. (1968). http://img.kb.dk/tidsskriftdk/pdf/gto/gto_0067-PDF/gto_0067_97642.pdf. Obtenido de Central Places and Functional Regions in Denmark Factor Analysis of Telephone Traffic Geografiks Tidsskrift.

INSTITUTO NACIONAL DE ESTADÍSTICA. Censo 2001. INE. www.ine.es/censo2001/

INSTITUTO NACIONAL DE ESTADÍSTICA. Censo 2011. INE. www.ine.es/censo2011/

ISTAT. (1997). I sistemi locali del lavoro 1991. Istituto Poligrafico e Zecca dello Stato, Roma.

JAIN, A., & DUBES, R. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall.

JAIN, A., MURTY, M., & FLYNN, P. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264-323.

JARDINE, N. (1971). A new approach to pattern recognition. Nature, 234, 526-528.

JOFRE-MONSENY, J. (2009). The scope of agglomeration economies: Evidence from Catalonia. Papers in Regional Science, 88(3), 575-590.

JOHNSON, R., & WICHERN, D. (1998). Applied Multivariate Statistical Analysis. Englewood Cliffs, NJ: Prentice-Hall.

JOHNSON, S. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

JOHNSTON, R. (1968). Choice in classification: The subjectivity of objective methods. Annals of the Association of American Geographers, 58(3), 575-589.

KAHLE, L. (1986). The Nine Nations of North America and the Value Basis of Geographic Segmentation. Journal of Marketing, 50, 37-47.

KAISER, H. (1966). An objective method for establishing legislative districts. Midwest Journal of Political Science, 10(2), 200-213.

KAUFMAN, L., & ROUSSEEUW, P. (1990). Finding Groups in Data : An introduction to Cluster Analysis. (W. S. Statistics, Ed.) New York: John Wiley & Sons, Inc.

KIRKPATRICK, S., GELATT, C., & VECCHI, M. (1983). Optimization by Simulated Annealing. Science, New Series, 220(4598), 671-680.

KOSFEL, R., ECKEY, H., & DREGER, C. (2006). Regional productivity and income convergence in the unified Germany. Regional Studies, 40(7), 755-767.

KOSTFELD, R., & LAURIDSEN, J. (2012). Identifying Clusters, within R&D Intensive Industries Using Local Spatial Methods MAGKS. ERSA European Regional Science Association, (p. 232).

KOUTSIAS, N., MARTÍNEZ-FERNÁNDEZ, J., & ALLGÖWER, B. (2010). Do Factors Causing Wilfires Vary in Space? Evidence from Geographically Weighted Regression. GIScience & Remote Sensing, 47(2), 221-240.

KRUSKAL, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society, 7(1), 48-50.

LANCE, G., & WILLIAMS, W. (1967a). A general theory of classificatory sorting strategies I. Hierchical systems. The Computer Journal, 9(4), 373-380.

LANKFORD, P. (1969). Regionalization: Theory and Alternative Algorithms. Geographical Analysis, 1(2), 196-212.

LEGENDRE, L., & LEGENDRE, P. (1983). Numerical Ecology.

LLOYD, C., & SHUTTLEWORTH, I. (2005). Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. Environment and Planning, 37(1), 81-103.

LO, C. (2008). Population Estimation Using Geographically Weighted Regression. GIScience & Remote Sensing, 45(2), 131-148.

LOADER, C. (1999). Local regression and likelihood. New York: Springer.

LOFTSGAARDEN D, O., & QUESENBERRY, C. (1965). A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics(36), 1049-1051.

LOPEZ-BAZO, E., & MOTELLÓN, E. (n.d.). Human capital and Regional Wage Gaps. Regional Studies, 46(10), 1347-1365.

MAO, J., & JAIN, A. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). IEEE Transactions on Neural Networks, 7(1), 16-29.

MARKUSEN, A. (1991). The Military Industrial divide: Cold war transformation of the economy and the rise of new industrial complexes. Environment and Planning D: Society and Space, 9, 391-416.

MARKUSEN, A. (1996). Sticky Places in Slippery Space: A Typology of Industrial Districts. Economic Geography, 72(3), 293-313.

MARSHALL, A. (1890). Principles of economic. Edition. London: Macmillan and Co., Ltd. Pub. Date.

MARTIN, D., NOLAN, A., & TRANMER., M. (2001). The application of zone-design methodology in the 2001 UK Census. Enviroment and Planning A, 33(11), 1949-1962.

MARTIN, R., & SUNLEY, P. (2003). Deconstruction Clusters: Chaotic Concept or Policy Panacea? Journal of Economic Geography, 3(1), 5-35.

MASSER, I., & BROWN, P. (1975). Hierarchical Aggregation Procedures for Interaction Data. Environment and Planning A, 7(5), 509-523.

MATHUR, V. (1999). Human Capital - Based Strategy for Regional Economic Development. Economic Development Quarterly, 13(3), 203-216.

McMORRIS, F., MERONK, D., & NEUMANN, D. (1983). A view of some consensus methods for trees. In J. In Felsenstein (Ed.). Berlin: Springer-Verlag.

MCQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley Symposium on mathematical statistics and probability, 1(14), 281-297.

McQUITTY, L. (1957). Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. Educational and Psychological Measurement, 17, 207-222.

MEADOWS, M., & DIBB, S. (1998). Assessing the implementation of market segmentation in retail financial services. International Journal of Service Industry Management, 9(3), 266-285.

MILLER, C., TUCKER, A., & ZEMLIN, R. (1960). Integer programming formulation of traveling salesmen problems. Journal of the ACM (JACM), 7(4), 326-329.

MILLIGAN, G., & COOPER, M. (1988). A study of standardization of variables in cluster analysis. Journal of Classification, 5, 181-204.

MITCHELL, W., & WATTS, M. (2009). Identifying Functional Regions in Australia using Hierarchical Aggregations Techniques. Geographical Research, 48(1), 24-41.

MITCHELSON, R., & WHEELER, J. (1994). The flow of information in a global economy: the role of the American urban system in 1990. Annals of the Association of American Geographers, 84(1), 87-107.

MITTAL, V., KAMAKURA, W., & GOVIND, R. (2004). Geographic Patterns in Customer Service and Satisfaction: An Empirical Investigation. Journal of Marketing, 68(3), 48-62.

MOLINA, M. (2001). European Industrial Districts: Influence of Geographic Concentration on Performance of the Firm. Journal of International Management, 7, 277-294.

MONGAY, C. (2005). Quimiometría. Universitat de Valencia.

MORRISON, D. (1967). Measurement problems in cluster analysis. Management Science (Series B, Managerial), 13(12), B775-B780.

MULHERN, F., & WILLIAMS, J. (1994). A comparative analysis of shopping behavoir in Hispanic and non-Hispanic market areas. Journal of Retailing, 70(3), 231-251.

MURTAGH, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The Computer Journal, 26(4), 354-359.

MURTAGH, F. (1984b). Counting dendrograms: A survey. Discrete Applied Mathematics, 7(2), 191-199.

MURTAGH, F. (1985). A survey of algorithms for contiguity-constrained clustering and related problems. Computer Journal, 28(1), 82-88.

NAKAYA, T. (2002). Local spatial interaction modelling based on the geographically weighted regression approach. (K. Modelling geographical systems: statistical and computational appications. Dordrecht, Ed.) Thomas, R.; Boots, B.; Okabe A.

NORONHA, V., & GOODCHILD, M. (1992). Modeling interregional interaction: Implications for Defining Functional Regions. Annals of the Association of American Geographers, 82(1), 86-102.

OPENSHAW, S. (1977b). Optimal zoning systems for spatial interaction models. Environment and Planning A, 9(2), 169-184.

OPENSHAW, S., ALVANIDES, S., & WHALLEY, S. (1998). www.geog.leeds.ac.uk/papers/98-9/. Recuperado el 19 de 06 de 2006, de Some further experiments with designing output areas for the 2001 UK census.

OVERALL, J., & KLETT, C. (1972). Applied Multivariate Analysis. (Series in Psychology ed.). New York: McGraw- Hill.

OWEN, D., & GREEN, A. (2000). Estimating commuting flows for minority ethnic groups in England and Wales. Journal of Ethnic and Migration Studies, 26(4), 581-608.

PALM, R. (2002). International Telephone Calls: Global and Regional Patterns. Urban Geographic, 23(8), 750-770.

PARTRIDGE, M., RICKMAN, D., ALI, K., & OLFERT, M. (2008). The Geographic Diversity of U.S. Nonmetropolitan Growth Dynamics: A Geographically Weighted Regression Approach. Land Economics, 84(2), 241-266.

PIORE, M., & SABEL, C. (1984). The second industrial divide: Possibilities for prosperity. New York.

PLANE, D., & HEINS, F. (2003). Age articulation of U.S. inter-metropolitan migration flows. The Annals of Regional Science, 37, 107-130.

PORTER, M. (1998). Clusters and the New Economics of Competition. Harvard Business Review, 76(6), 77-90.

PORTER, M. (2000). Location Competition and Economic Development: Local Clusters in a Global Economy. Economic Development Quarterly, 2(1), 15-34.

RABELLOTTI, R. (1995). Is There an Industrial District Model? Footwear Districts in Italy and Mexico Compared. World Development, 23(1), 29-41.

RABELLOTTI, R. (1997). External Economies and Cooperation in Industrial Districts: A Comparison of Italy and Mexico. London: McMillan.

ROHLF, F. (1974). Algorithm 81: Dendogram plot. The Computer Journal, 17(1), 89-91.

ROHLF, F. (1982). Single link clustering algorithms. Handbook of Statistics, 2, 267-284.

ROSENBERG, B. (1973). A survey of stochastic parameter regression. Annals of Economic and Social Measurement, 1, 381-97.

ROSS, G. (1969). Algorithm AS 15: Single linkage cluster analysis. Applied Statistics, 18(1), 106-110.

ROSSITER, D. J. (1981). Program GROUP-the identification of all possible solutions to a constituency-delimitation problem. Environment and Planning A, 13(2), 231-238.

ROUWENDAL, J. (2004). Commuting Cost and Commuting Behavior: Some Aspects of the Economic Analysis of Home-Work Distance. NECTAR's Cluster 4. Alicante.

SALOM, J., & CASADO D., J. (2007). Movilidad Cotidiana y Mercados Locales de Trabajo en la Comunidad Valenciana. Boletín de la A.G.E.(44), 5-28.

SALTON, G., & McGILL, M. (1983). Introduction to Modern Information Retrieval. New York, Tokio: McGraw-Hill.

SARABIA, F. (2012). Using values and shopping styles to identify fashion apparel segments. International Journal of Retail & Distribution Management, 40(3), 180-199.

SARABIA-SANCHEZ, F., VIGARAY, M., & HOTA, M. (2012). Using values and shopping styles to identify fashion apparel segments. International Journal of Retail & Distribution Management, 40(3), 180-199.

SAXENIAN, A. (1994). Regional Networks: Industrial adaptation in Silicon Valley and Route 128. Harvard University Press.

SCHMITZ, H. (1999). Global Competition and Local Cooperation: Success and Failure in the Sinos Valley, Brazil. World Development, 27(9), 1627-1650.

SCHWARTZ, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-4.

SCOTT, A. (1986). New industrial space. London: Pion.

SFORZI, D. (1989). The Geography of industrial district in Italy. In E. a. Bamford (Ed.), In small firms and industrial districts in Italy (pp. 153-173). London: Routledge.

SFORZI, F. (1990). The quantitative importance of Marshallian industrial districts in the italing economy. (F. Pyke, G. Becattin, & W. Sengenberger, Eds.) Geneva: Industrial Districts and inter-firm co-operation in Italy.

SFORZI, F., & LORENZINI, F. (2002). I distretti industriali a VVAA L'esperienza Italiana dei Distretti Industriali. Istituto per la Promozione Industriale (IPI).

SHAVER, J., & FLYER, F. (2000). Aglomeration Economies, Firm Heterogeneity, and Foreign Direct Investment in the United States. Strategic Management Journal, 21, 1175-1193.

SHAVER, M. (2000). Aglomeration Economies, Firms Heterogeneity, and Foreign Direct Investment in the U.S. Strategic Management Journal , 21, 1175-1193.

SIBSON, R. (1973). SLINK: An optimally efficient algorithm for the single link cluster method. The Computer Journal, 16(1), 30-34.

SIMINI, F., GONZÁLEZ, M., MARITAN, A., & BARABÁSI, A. (2012). A Universal Model for Mobility and Migration Patterns. Nature, 484, 96:100.

SIMPKIN, L. (2008). Achieving market segmentation from B2B sectorisation. Journal of Business & Industrial Marketing, 23(7), 464-474.

SNEATH, P. (1957). The Application of Computers to taxonomy. Journal of General Microbiology, 17, 201-226.

SOKAL, R. a. (n.d.). The comparison of dendrograms by objective methods Tax. Taxon, 33-40.

SOKAL, R., & MICHENER, C. (1958). A Statistical Method for Evaluating Systematic Relationships. The University Science Bulletin, 38(22), 1409-1438.

SOKAL, R., & ROHLF, F. (1962). The comparison of dendograms by objective methods. Taxon, 11, 33-40.

SPJOTVOL, E. (1977). Random coefficients regression models: A review. Mathematische Operationsforschung und Statistik(8), 69-93.

STORPER, M., & WALKER, R. (1989). The capitalist imperative: Territory, technology and industrial growth. New York: Basil Blackwell.

STUART, T., & SORENSON, O. (2003). The geography of opportunity: Spatial heterogeneity in founding rates and the performance of biotechnology firms. Research Policy, 32, 229-253.

TAYLOR, P. (1973). Some implications of spatial organization of elections. Transactions of the Institute of British Geographers, 60, 121-136.

TÖDTLING, F., & WANZENBÖCK, H. (2003). Regional Differences in Structural Characteristics of Start-ups. Entrepeneurship & Regional Development, 15, 352-370.

VAN RIJSBERGEN, C. (1970). Algorithms 52: A fast hierarchical clustering algorithm. The Computer Journal, 13(3), 324-326.

VIDAL, D. (2003). Técnicas de Análisis Multivariante para la Investigación Social y Comercial, ejemplos prácticos utilizando SPSS versión 11. Revista Internacional de Sociología(35), 228-230.

VILADECANS-MARSAL, E. (2004). Agglomeration economies and industrial location: City-level evidence. Journal of Economic Geography, 4(5), 565-582.

WAND, M., & JONES, M. (1995). Kernel smoothing. London: Chapman & Hall.

WARD Jr., J. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

WARD Jr., J., & HOOK, M. (1963). Application of hierarchical grouping procedure to a problem of grouping profiles. Educational and Psychological Measurement,, 23(1), 69-81.

WEAVER, J., & HESS, S. (1963). A procedure for nonpartisan distriting-development of computer technics. Yale Law Journal, 73(2), 288-308.

WHEELER, J., & MITCHELSON, R. (1989). Information Flows among Major Metropolitan Areas in the United States. Annals of the Association of American Geographers, 79(4), 523-543.

WILLIAMS, J. (1995). Political redistricting: A review. Papers in Regional Science, 74(1), 13-40.

WISE, S., HAINING, R., & MA, J. (2001). Providing spatial statistical data analysis functionallity for the GIS user: The SAGE project. International Journal of Geographical Information Science, 15(3), 239-254.

WISHART, D. (1969). An algorithm for hierarchical classifications. (Note 256). Biometrics, 25(1), 165-170.

WISHART, D. (2002). k-means clustering with ourtlier detection, mixed variables and missing values (Vol. Exploratory Data in Empirical Research.). (M. a. Schwaiger, Ed.) New York: Springer.

XIAO, Y., & DUNHAM, M. (2001). Interactive clustering for transaction data. Lecture notes in Computer Science, 2114, 121-130.

XU, H., WANG, H., & LI, C. (2002). Fuzzi tabu search method for the clustering problem. International conference on machine learning and cybernetics, 2, 876-880. Beijing.

ZHANG, B., & SRIHARI, S. (2003). Properties of Binary Vector Dissimilarity Measures. Retrieved from Technical report, CEDAR, Departament of Computer Science & Engineering, University of Buffalo, the State University of New York: http://www.cedar.buffalo.edu/papers/publications,html.

ZOLTNERS, A., & SINHA, P. (1983). Sales territory alignment - a review and model. Management Science, 29(11), 1237-1256.