Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Programa de Maestría en Computación

Desarrollo de una herramienta de visualización de redes sociales mediante un enfoque difuso

Informe final de tesis para optar por el grado de *Magister Scientiae en Computación,* con énfasis en Ciencias de la Computación

Estudiante: Natalia Marín Pérez

Profesor Asesor: Carlos González Alvarado

San José, Costa Rica

Noviembre, 2017

Abstract

A social network is a concept based on graph theory, which allows to linking between different type of objects depending on the how it is applied. This type of link, usually is binary-like, so the objects may or may not be related to each other. On the other hand, there are multiple applications where this kind of analysis is insufficient, certain objects could belong to more than one group at the same time. For example, one person that resides in certain place may need to move because of studies for 5 days a week. So, they stay at home for two days and the rest is spent outside. How can we represent this situation? Using fuzzy logic, we can provide a greater weight to the individual spending time outside (5/7) and (2/7) in its home. To resolve this, a visualization model approach was developed that is based on social network graphs, it supports the investigation of fuzzy groups in weighted directed graphs and can be adjusted to provide different levels of detail by using spatial data and fuzzy clustering. This helps find new social behavior patterns according to the dataset and the type of application. The visualization was showcased using the Costa Rican migration dataset as well as a dataset generated through a temporary internal migration survey. The case studies showed that our visualization approach helps understand better the fuzzy relationship in a social network.

Resumen

Una red social es un concepto basado en teoría de grafos, el cual permite unir diferentes tipos de objetos dependiendo de cómo sea aplicado. Este tipo de unión usualmente es de tipo binario, por lo tanto, los objetos puede que estén o no estén relacionados entre sí. Por otra parte, hay múltiples aplicaciones donde este tipo de análisis es insuficiente, algunos objetos pueden pertenecer a más de un grupo a la vez. Por ejemplo, una persona que reside en cierto lugar puede moverse debido a estudios por 5 días a la semana. Por lo tanto, se queda en la casa por dos días y el resto se encuentra fuera. ¿Cómo se podría representar este escenario? Utilizando lógica difusa, se le daría un peso mayor el tiempo que el individuo se encuentra fuera de la casa (5/7) y el resto se asigna al tiempo que se encuentra en la casa (2/7). Para resolver esto, se implementó un modelo de visualización basado en redes sociales, la cual soporta grupos difusos en grafos ponderados y puede ser ajustado para proveer diferentes niveles de detalle al utilizar datos espaciales y agrupación difusa. Esto permite encontrar nuevos patrones sociales de comportamiento de acuerdo con el conjunto de datos y al tipo de aplicación. La visualización se aplicó en el conjunto de datos de la migración en Costa Rica, la encuesta de uso de tecnología en Estados Unidos, así como datos generados en una encuesta para Costa Rica sobre migración interna temporal. Los casos de estudio mostraron que nuestra visualización ayuda a entender mejor la relación difusa en una red social.

APROBACIÓN DE LA TESIS

"Desarrollo de una herramienta de visualización de redes sociales mediante un enfoque difuso"

TRIBUNAL EXAMINADOR

Dr. Carlos González Alvarado
Profesor Asesor

Dr. José Castro Mora
Profesor Lector

MAP. Misael Castillo Brenes
Profesional Externo

Dr. Roberto Cortés Morales
Coordinador/Programa De Maestría

TEC | Tecnológico de Costa Rica
Maestría en Computación
Noviembre, 2017

Agradecimientos

Gracias al Instituto Nacional de Estadísticas y Censos por la pronta respuesta y resolución a las preguntas realizadas para poder completar esta investigación.

También me gustaría agradecer el apoyo del profesor tutor Carlos González Alvarado por sus consejos y su paciencia a lo largo del desarrollo de esta tesis.

Intel por brindar la posibilidad de recibir las clases del Instituto Tecnológico dentro del edificio, así como brindar un cierto nivel de flexibilidad para poder completar el trabajo. Así como también el apoyo de mi familia fue de gran ayuda.

Tabla de contenidos

Lista de Figuras

Capítulo 1. Introducción

En minería de datos, la cantidad de datos con la que se tiene que trabajar normalmente es muy grande y es difícil encontrar información relevante solo viendo los datos por sí solos. Existen métodos no-supervisados para hacer que altos volúmenes de datos se puedan analizar de manera más simple reduciendo su complejidad o encontrando un número menor de dimensiones para representar las estructuras relevantes. Entre estos métodos, el análisis de clústeres o agrupamientos tiene como fin agrupar los objetos en conjuntos con alta similitud entre ellos y con baja similitud entre conjuntos. Por otra parte, el resultado de la agrupación se puede utilizar para su análisis por medio de la visualización. Como parte de este proyecto, se ha utilizado el concepto de redes sociales con el fin de interpretar visualmente los resultados en el nivel del comportamiento humano.

El análisis de redes sociales es una rama del análisis de redes, que estudia el cómo un objeto (persona, producto, servicio, etc.) puede estar ligado a una conexión más grande de redes sociales. Por ejemplo, se considera en muchas ocasiones que el éxito o fracaso de las organizaciones a veces depende del patrón de comportamiento de su estructura interna. Y este análisis no resulta sencillo, porque quizás no se incluye variables lingüísticas para representar tales comportamientos. Y en esa vía, la lógica difusa nos puede ayudar para descubrir patrones y conocimientos entre objetos de un conjunto de datos dado. (INSNA, 1999).

Aunque el tema principal de este proyecto gira alrededor del comportamiento humano, los métodos para analizar redes sociales no están limitados a solo interacciones sociales, también se utilizan en redes metabólicas en biología, análisis de comunidades, análisis de consensos (decisiones políticas), para problemas técnicos en optimización de infraestructura, entre otros. (Newman, 2003). Por lo tanto, conceptos y algoritmos que se son útiles en un tipo de redes pueden asimismo aplicarse a otros tipos de redes.

En la presente investigación se pretende desarrollar un método de análisis de redes sociales basado en un enfoque difuso que permita explorar hasta qué punto aplica a nivel de las

relaciones entre objetos. Y tomaremos como validación de nuestra propuesta, por una parte, el caso de las migraciones internas de personas en nuestro país, cuyos datos por estudiar provienen de los obtenidos de la Encuesta Nacional de Hogares del INEC, en los últimos cuatro años (INEC, Catálogo Central de datos INEC, 2008). Asimismo, para validar la metodología propuesta también se estará analizando la interacción de las personas con las nuevas tecnologías en Estados Unidos, gracias a la encuesta realizada por el *Pew Research Center* (Pew Research Center, March 17-April 12,2015 - Libraries and Technology Use, 2017).

En el caso de las migraciones, los flujos de migración que ocurren dentro de los límites del territorio del país se refieren a un proceso definido como la *migración interna*. La migración interna está influenciada por regulaciones del estado, características geográficas, condiciones socioeconómicas, y factores externos, que pueden dar forma a los patrones de migración interna. (INEC, Resultados Generales, 2011)

Lo que se busca en esta investigación es aplicar al análisis de clústeres, el algoritmo fuzzy c-means y el análisis de redes sociales para desarrollar una herramienta que permita el agrupamiento y representación visual de relaciones entre objetos de manera tal que brinde más información para la toma de decisiones que las redes sociales tradicionales.

De esta manera, el objetivo general del proyecto ha sido el desarrollo de una herramienta de visualización de redes sociales utilizando un enfoque difuso, el cual será evaluado en los datos de migración de Costa Rica, así como la representación de cómo interactúan las personas con la tecnología.

Por otra parte, el proyecto se sustentó en los siguientes objetivos específicos:

- Analizar y comparar diferentes distancias con el fin de escoger cuál se adapta mejor al problema de incorporar valores de pertenencia (difusidad) a los objetos
- Desarrollar una metodología que permita el análisis de clústeres difusos en una red social
- Implementar una herramienta que represente, de forma difusa, la conexión entre objetos en diferentes particiones
- Validar la herramienta con diferentes conjuntos de datos para medir su efectividad

De acuerdo a los objetivos anteriores, en este documento se presentan las alternativas de agrupamiento, recolección y visualización de datos, así como la propuesta del proyecto en cuestión; en el capítulo 2 se analizan diferentes conceptos necesarios para el desarrollo de este proyecto respecto al análisis difuso, las redes sociales y los datos por analizar. A continuación, en el capítulo 3 se presenta la metodología la cual comprende las herramientas y el proceso utilizado para analizar, entender los datos por procesar y realizar la agrupación difusa de los mismos utilizando el algoritmo *fuzzy c-means*. En el capítulo 4 se explora con más detalle el método de visualización seleccionado para representar la red social con un enfoque difuso. En primer lugar, se utiliza el conjunto de datos sobre población y trabajo del INEC y luego se reproduce el mismo método de visualización en datos obtenidos del *Pew Research Center* los cuales muestran la relación de las personas con la tecnología en los últimos dos años. En el capítulo 5 se detallan los resultados de la metodología implementada, así como los resultados utilizando el segundo conjunto de datos. Finalmente, en el capítulo 6 se presentan las conclusiones basadas en los resultados y también se incluyen recomendaciones de trabajo futuro en el que se puede mejorar lo propuesto en esta tesis.

Capítulo 2. Marco teórico

2.1 Antecedentes

Un primer aspecto que debemos considerar tiene que ver con la técnica de minería de datos estudiada. En nuestro caso, escogimos el análisis de clústeres o agrupamientos difusos debido a que los datos generados por actividades humanas tienden a tener un comportamiento difuso, por lo que un análisis binario (blanco o negro) hace que se pierda conocimiento valioso a la hora de realizar el análisis. Además, se escogieron los diagramas de enlaces y grafos ya que son una manera efectiva para percibir las relaciones entre objetos los cuales toman en cuenta los principios de Gestalt de cierre y de continuidad, (Koffka, 1935). Cuando se usan estos diagramas, los vértices están mapeados a figuras geométricas tales como círculos o cuadrados y las relaciones entre sí normalmente se expresan ya sea por líneas rectas o curvas.

También se han realizado visualizaciones de grafos difusos utilizando el enfoque "*force-directed*" para modelar el dibujo de un grafo general como un problema de optimización numérica (Ashouri, 2016). Debido a las fuerzas de atracción (figura 1a) entre nodos adyacentes y las fuerzas de repulsión (figura 1b) entre todos los nodos, los vértices que están altamente conectados se posicionan cerca de los otros. De esta manera, las agrupaciones emergen visualmente pero no son completamente precisas. En otros estudios también se ha utilizado este método para modelo de grafos interactivo de un millón de nodos en el cual se encontró que esta metodología no escalaba bien para la cantidad de nodos dada ya que el algoritmo "*force-directed*" requiere guardar los pares de distancias de cada uno de los nodos y se tuvo que realizar ciertos ajustes como aprovechar el paralelismo para poder mejorar el desempeño para una gran cantidad de grafos. (Mi, Maoyuan, Moeti, Yong, & North, 2016). En esta tesis se aprovecha la implementación propuesta por (Dwyer, 2009) la cual es adaptada en la biblioteca de D3js que posee un enfoque "*force-directed*", pero lo combina con un esquema de restricción simple inspirado en métodos dinámicos basados en la posición.

Existen tres tipos de restricción: lineal fija, circular (para dibujar círculos en grafos dirigidos) y para que los nodos o agrupaciones, ya sean representados por círculos, rectángulos, en forma de cápsula o convexos, no se traslapen, permitiendo así un desempeño más rápido y un diseño más escalable el que está sujeto a las restricciones mencionadas.

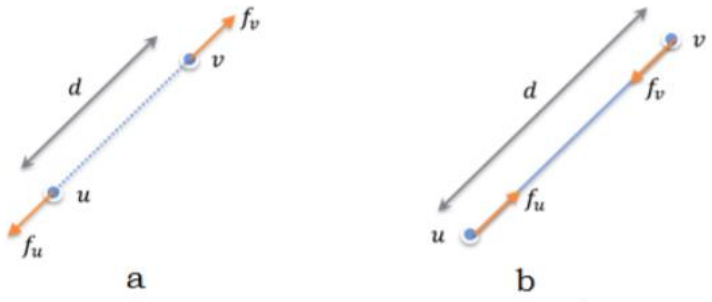

Figura 1 a. Fuerza de atracción nodo a nodo (Kamada & Kawai, 1989) b. Fuerza de repulsión nodo a nodo (Kamada & Kawai, 1989)

Aunque sí existe gran cantidad de estudios relacionados en redes sociales y cómo agrupar los grafos en una visualización, existe escasa información en donde su enfoque sea del tipo difuso. Por ejemplo, en el enfoque de Itoh (Itoh, Muelder, Ma, & Sese, 2009) se utilizan gráficos circulares para visualizar múltiples categorías a las cuales pertenece un nodo que, al tener un tamaño mínimo para permitir la diferenciación de membresía, la visualización se volvía desordenada conforme crecía el tamaño de los nodos. También, (Vehlow, Reinhardt, & Weiskopf, 2013) han trabajado en visualizaciones de comunidades difusas donde se tratan de solucionar varios de los problemas mencionados anteriormente, una visualización de grafos difusa que sea escalable y de fácil comprensión donde toman en cuenta la membresía de los nodos. Por otra parte, la implementación se enfoca en grafos no dirigidos y además no toma en cuenta el tema de espacio-tiempo en la representación, lo cual se analiza en el presente trabajo, en donde contaremos con grafos dirigidos (la migración de un punto A-B) y cómo evolucionan las agrupaciones a través del tiempo.

Una vez que se ha analizado esta problemática se procede a introducir con más detalle la propuesta de Bezdeck sobre *"Fuzzy c-means"* o algoritmo basado en clústeres difusos.

2.1.1 Cluster difuso – *"Fuzzy c-means"*:

El *fuzzy C-means* nace a partir de una modificación hecha por Bezdek a la metodología original para clústeres. Una gran parte de los algoritmos de clúster difuso se basan en la minimización de fuzzy c-means formulada por Bezdek como se muestra en la ecuación (1). (Bezdek, 1984)

$$J_m(U,V) = \sum_{k=1}^{N} \sum_{i=1}^{C} u_{ik}^{m} \|Y_k - v_i\|^2 \quad (1)$$

En este caso, $J_m(U, V)$ es la suma del error al cuadrado para el conjunto de clústeres difusos representado por la membresía de la matriz U y el conjunto de centros de clústeres asociados *V*. La expresión $\|Y_k - v_i\|^2$ representa la distancia entre los datos X_k y el centro del clúster v_i. El error al cuadrado es usado como el índice de desempeño que mide la suma (usando los pesos) de distancias entre los centros de los clústeres y los elementos en los clústeres difusos correspondientes. El número *m* se encarga de influenciar los grados de membresía, es decir, cada elemento de u_{ik} define el grado en el que cada elemento de Y_k pertenece al clúster v_i. La partición se vuelve más difusa cuando el valor *m* incrementa.

Las variables de la ecuación se explican con más detalle a continuación (Bezdek, 1984):

$$Y = \{y_1, y_2 \dots y_n\} \subset R^n = los\ datos,$$

$$c = número\ de\ clusteres\ en\ Y; 2 \leq \mathrm{c} < \mathrm{n},$$

$$m = exponente\ con\ peso; 1 \leq m < \infty,$$

$$U = partición\ difusa\ de\ Y; U \in M_{fc},$$

$$v = (v_1, v_2, \dots, v_c) = \ vectores\ de\ centros$$

$$v_i = (v_{i1}, v_{i2}, \dots, v_{in}) = centro\ de\ cluster\ i$$

$$\| \|_A = norma\ A\ Inducida\ en\ R^n$$

$$A = definida - positiva\ matriz\ de\ peso\ (n\ x\ n)$$

A continuación, se muestran los diferentes parámetros considerados en el algoritmo Fuzzy c-means:

- *Número de clústeres*: Es necesario tener una idea acerca de la estructura de los datos. Para determinar el número apropiado de clústeres es necesario contar con medidas de validez que indiquen qué tan adecuada es la partición obtenida. (Babuska, n.d.)
- *La combinación iterativa e inserción de clústeres*: Se puede, ya sea contar con una gran cantidad de clústeres y empezar a reducirla sucesivamente al combinar clústeres similares o, por el contrario, se puede empezar con un número pequeño de clústeres, para luego proceder a agregar clústeres iterativamente en las regiones donde los puntos tienen un nivel bajo de membresía.
- La principal diferencia con respecto al algoritmo de *k-means* es la introducción de un vector que expresa el porcentaje de pertenencia a un punto dado de cada uno de los clústeres. (Babuska, n.d.)

La función del clúster difuso es más lenta que el algoritmo de k-means en eficiencia, pero da mejores resultados donde los datos están incompletos.

Para el presente proyecto, se utilizó una manera diferente a la tradicional con el propósito de poder incorporar un enfoque basado redes sociales.

Normalmente los gráficos de algoritmos difusos (ver figura 2) muestran la membresía compartida como intersecciones entre las agrupaciones. Aunque se muestre que los clústeres tienen membresía compartida entre diferentes agrupaciones, no es fácil entender visualmente qué tanto es el porcentaje de la membresía al que pertenece cada elemento a dichas agrupaciones.

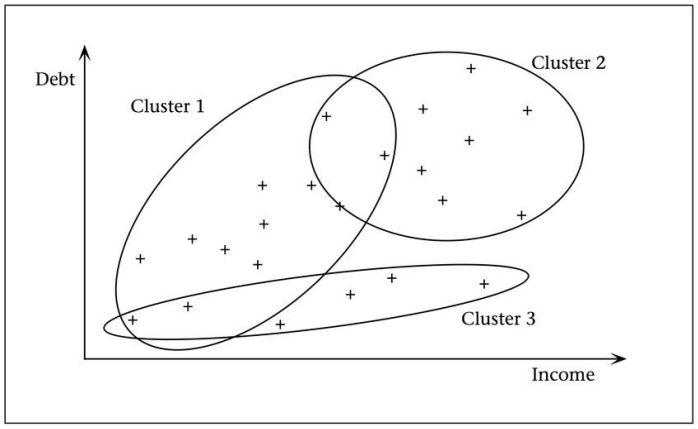

Figura 2. Ejemplo de una representación de clústeres difusos tomado de (Sippel, 2016)

Así como se ilustró anteriormente se encuentra la necesidad de buscar un tipo de visualización que permita fácilmente comprender la membresía de los elementos para facilitar el análisis de los datos. En este trabajo se estará utilizando los gráficos circulares (conocidos como *pie chart* en inglés) en donde cada color del arco del pie indica el porcentaje de membresía al que pertenece o qué tan fuerte es la pertenencia a una agrupación específica, también se estará agregando una leyenda para indicar la descripción de cada uno de los colores que representan las agrupaciones. Por ejemplo, se tiene datos sobre *papers* científicos, en los cuales se requiere saber en qué temas han participado los diferentes miembros del equipo de investigación, específicamente para sistemas operativos(SO), inteligencia artificial(IA) y minería de datos(MD). Cada uno de los miembros puede haber tenido participación en más de una de las áreas de investigación, como se puede notar en la figura 3 donde el investigador participó tanto en las áreas de Sistemas Operativos como Inteligencia artificial. Aunque en este caso participó en mayor medida en proyectos de Sistemas Operativos, es un recurso que potencialmente se podría aprovechar también en proyectos de Inteligencia artificial en un futuro.

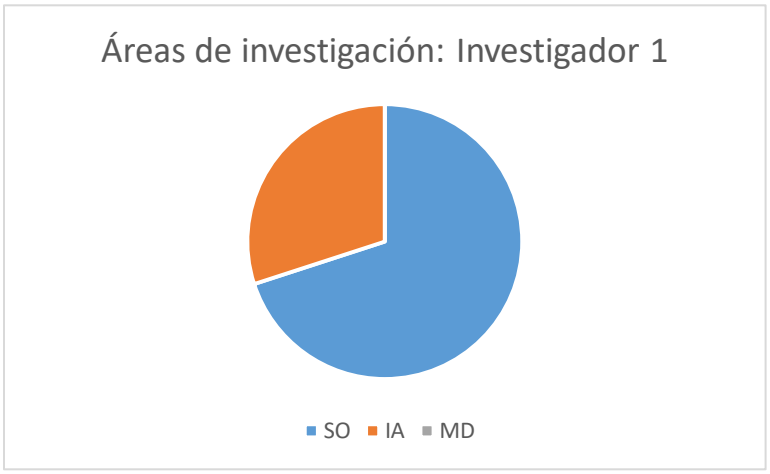

Figura 3. Ejemplo de una representación de clústeres difusos según metodología propuesta

Como se pudo notar este tipo de representación permite un análisis más intuitivo de los datos que se presentan, por otra parte, todavía se deben analizar las relaciones entre los elementos, por lo que a continuación, se estará analizando con más detalle qué son las redes sociales y cómo se han aplicado en este proyecto. Asimismo, se analiza la importancia de darle un enfoque difuso a la representación visual propuesta.

2.1.2 Análisis de Redes sociales

El enfoque de redes sociales se basa en la noción de los patrones que existen en las uniones sociales en las cuales los actores están relacionados a consecuencias importantes para esos actores. Lo que se busca con el análisis de las redes sociales es encontrar este tipo de patrones y determinar las condiciones con las cuales estos patrones se llegan a presentar y descubrir sus consecuencias. (Freeman, 2004: 2)

En este caso buscamos aplicar redes sociales con un comportamiento difuso debido a que normalmente las redes sociales van a ser utilizadas para presentar el comportamiento de las personas y como éstas son afectadas por el ambiente que las rodea. El comportamiento humano no es binario, por ejemplo, si se quisiera entender por qué las personas se mueven de un lugar a otro normalmente, es limitado si se enfoca en dos posibles respuestas. O si se quiere entender la situación en la que se encuentra la persona que migra. Se requiere entonces, contar con una comprensión más profunda de las circunstancias que han provocado esta situación.

El análisis de redes sociales se utiliza mucho en estudios sobre cómo las personas viven sus vidas, socializan, encuentran trabajos, entre otros. Pero la investigación a nivel de cómo los migrantes usan sus redes sociales para elegir el destino, encontrar un amigo, encontrar trabajo y lugar donde vivir por ejemplo es bastante limitada.

Se realizó una encuesta a 100 personas con edades entre 25 y 50 años preguntando sobre los movimientos que habían realizado en la semana anterior, así como entender la razón por la cual se movieron de una provincia a otra como se puede observar en la figura 4.

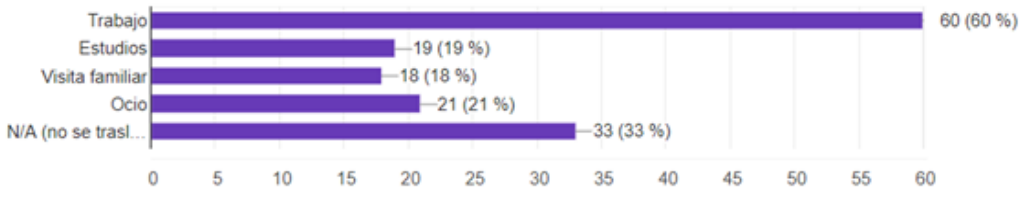

Figura 4. Razones por las cuales se realizó el movimiento de una provincia a otra

El mayor porcentaje se basó en asuntos de trabajo, la que puede ser una potencial razón por la que el individuo se mueva permanentemente hacia la provincia en la que trabaja. Por otra parte, para poder identificar este hecho de una manera más confiable se hizo necesario identificar otros datos del individuo que, por el contrario, podían evitar que se diera un movimiento permanente como capacidad económica, actual residencia de familiares y amigos, entre otros.

Para propósitos de este proyecto se empezó a experimentar con un tipo de visualización como el que se muestra en la figura 5.

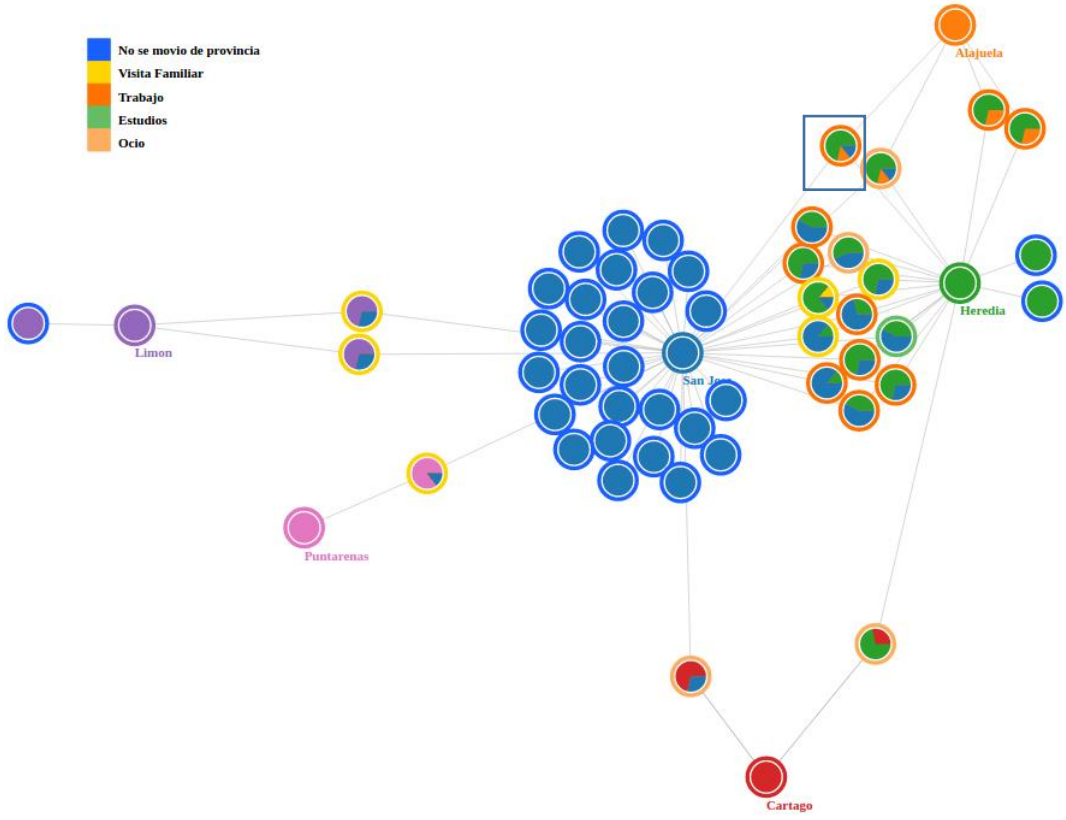

Figura 5. *Movimientos internos en Costa Rica durante una semana-según encuesta realizada en mayo 2017*

Para ilustrar los movimientos entre las diferentes provincias se le asigna un cierto grado de pertenencia a cada provincia de acuerdo con el movimiento que se realizó según la encuesta mencionada anteriormente. Por ejemplo, dado un individuo x que se encuentra en la provincia i, t_i días de la semana, pero también se traslada a la provincia j, t_j días en esa

misma semana, entonces $t_i + t_j = 7$ y los porcentajes de membresía serían $\frac{t_i}{7}$ y $\frac{t_j}{7}$ respectivamente. En este caso el color de la agrupación o el porcentaje de membresía se dan por el color de la o las provincias a que pertenece el individuo. El arco que rodea cada círculo indica la razón por la cual se dio el movimiento.

Asimismo, en la figura 5 se puede notar como existen individuos que se trasladan hasta dentro de tres provincias. Para el caso contenido en el rectángulo se puede percibir como se distribuye entre Heredia (verde), Alajuela (naranja) y San José (azul), es una persona que la mayor parte del tiempo estuvo en Heredia (lugar donde vive) y se movió a Alajuela por un periodo más largo de tiempo que San José, probablemente debido a la naturaleza de su trabajo donde tiene que trasladarse a varios puntos del país.

Ademàs, se pueden mostrar varios patrones, también se encontró que muchos de los individuos con mayor movilidad son personas entre 20 y 35 años. Es claro que, con mayor cantidad de datos, se podrían encontrar otros patrones más interesantes para entender a más profundidad los movimientos que se dan dentro del país.

Esta visualización permitió dar una idea clara de cómo representar la membresía de clúster difuso que se estará formulando en la sección 3.4. Por otra parte, en la siguiente sección se va explicar los datos que se estarán utilizando para crear la representación del clúster difuso de las migraciones internas de Costa Rica según datos obtenidos del INEC.

2.2 Datos de migración interna

Los análisis que se han hecho con redes sociales suelen ser de tipo binario. Por ejemplo, para un atributo *edad* se puede calificar a una persona en varios niveles como "*joven*", "*muy joven*", "*bastante joven*", "*viejo*", "*no tan viejo*", "*no tan joven*" en lugar de utilizar propiamente los valores de la edad (18, 22, 35, 40, etc). Más específicamente una variable lingüística es caracterizada por una quíntupla (*H, T(H), U, G,M*), en la cual *H* sería el nombre de la variable; *T(H)* es el conjunto de términos *H*; *U* es el universo de discurso; *G* es la regla sintáctica que genera los términos en *T(H)* y *M* es una regla sintáctica que asocia su significado con cada valor lingüístico *X*, donde *M(X)* denota un subconjunto difuso de *U*. El significado del valor lingüístico de *X* es caracterizado por una función de

compatibilidad c: $U \rightarrow [0,1]$ la cual asocia su compatibilidad con cada u en U. Por ejemplo, la compatibilidad de la edad 27 con joven podría ser de 0.7, mientras que 35 sería de 0.2 (L.A., 1975). Las variables lingüísticas son primordiales en la lógica difusa y serán necesarias a la hora de realizar el análisis de los datos para la implementación de la metodología, más aún, para el caso de interacciones entre personas y su comportamiento. De esta manera, al incorporar el elemento difuso se estará permitiendo una representación más precisa del comportamiento de las personas a nivel de la migración interna utilizando los conjuntos de datos de la INEC así como el caso de su interacción con las tecnologías utilizando los datos de *Pew Research Center*.

La migración representa un proceso complejo de flujos de población sobre un espacio geográfico. Existen diferentes tipos de migración interna:

- Las migraciones temporales (algunas semanas, o algunos años en el caso de trabajadores inmigrados).
- Las migraciones de ocio.
- Las migraciones pendulares (desplazamiento diario del lugar de habitación hacia el trabajo). (Alvarado, 2009)

La aplicación de la metodología será sobre las migraciones de tipo temporal. En el caso de Costa Rica para poder investigar la migración interna es necesario utilizar el censo de la población, los cuales se obtienen de INEC Costa Rica. El conjunto de datos de la encuesta nacional de hogares muestra información anual sobre las condiciones de vida de la población y sus niveles de bienestar. En este caso se recolectaron alrededor de 120.000 filas de datos con respecto a las encuestas que se realizaron en los últimos 3 años (período 2014-2016).

La manera en que se recolecta los datos de migrantes internos en el Censo del INEC es preguntando a las personas (alrededor de 120.000,00 entrevistados) sobre donde han vivido durante los últimos 5 años. Cuando la persona que responde vive en la región B en el momento de la entrevista y reporta haber vivido en la región A hace 2 años, esta persona es un migrante de la región *A* hacia la región *B*, aunque durante el periodo de 2 años podría haber vivido en alguna otra región. En la tabla 2 se muestra la matriz de migración interna

de acuerdo a la Encuesta Nacional de Hogares 2015 (INEC, INEC Censo 2015) y seguidamente la distribución de la población de acuerdo a los resultados de los censos 2000 – 2011 (INEC, Resultados Generales, 2011).

Tabla 1 Resumen de Matriz de migrantes según Censo 2011 (INEC, Resultados Generales, 2011)

	Central	***Chorotega***	***Pacífico Central***	***Brunca***	***Huetar Caribe***	***Huetar Norte***
Central	0	1 308	1 876	1 277	2 749	15 649
Chorotega	3 728	0	406	864	276	88
Pacífico Central	1 954	115	0	378	218	0
Brunca	3 583	134	378	0	992	221
Huetar Caribe	4 161	695	206	336	0	528
Huetar Norte	2 452	247	69	310	2 794	0

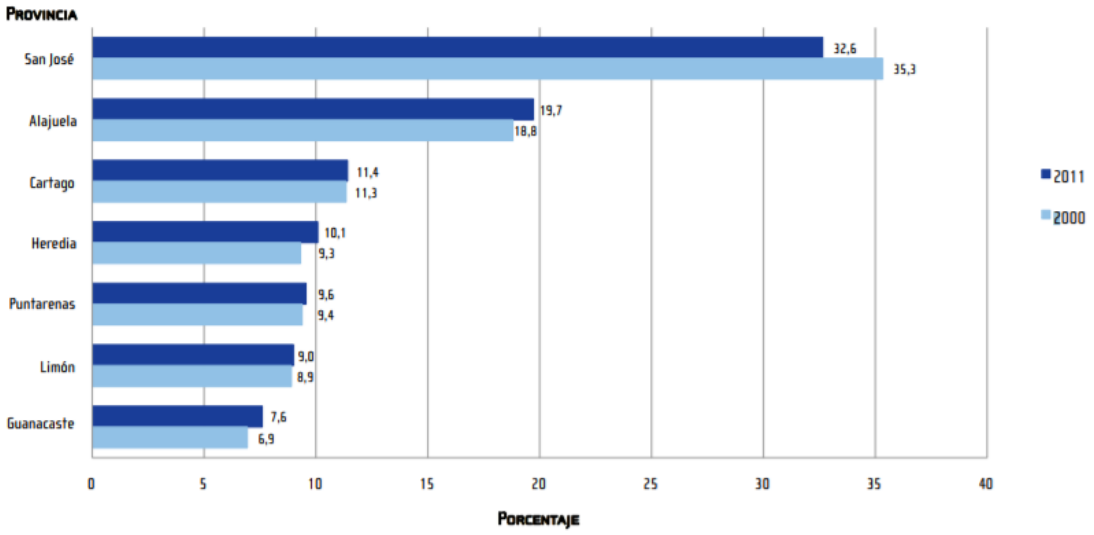

Figura 6 Distribución porcentual de la población por provincia 2000-2011 (INEC, Resultados Generales, 2011)

Los datos mencionados anteriormente serán analizados y preprocesados para tener datos relevantes que permitan la solución del problema. Una vez que los datos se procesaron (normalizar los datos, seleccionar los atributos relevantes) se aplicó la metodología usando clústeres difusos. Asimismo, se trabajó en una visualización para mostrar los datos a través del tiempo y así comprender cuáles son las zonas con mayor migración interna y sus

razones para poder establecer una correlación entre los mismos. También se tomó en cuenta encuestas que se han realizado en otros países para conocer las razones por las cuales las personas migran como se puede notar en la figura 7 (Irke, 2014). De esta manera se tiene una mejor base para la selección de los atributos de peso según los datos que se pueden obtener a través de la INEC (INEC, Catálogo Central de datos INEC, 2008).

Figura 7 Razones por las cuales las personas migran de acuerdo al reporte de resultados por el centro de Censos de Estados Unidos (Irke, 2014)

El presente trabajo asimismo se podría aplicar a otras áreas disciplinarias como la investigación geográfica, la política para tomar decisiones mejor fundamentadas, por ejemplo, para la apertura de servicios en ciertas regiones del país, mejoras a nivel de transporte e incrementar las oportunidades de empleo en las zonas donde se requiera.

Los atributos que se estarán analizando se muestran en la tabla 2. La forma en que se analizan los datos es de manera que ayuden a entender las razones por las cuales se da la migración interna en Costa Rica. En la siguiente sección se estará detallando el análisis de datos que se realizó con estos atributos.

Tabla 2 Conjunto de datos Encuesta Nacional de Hogares

Nombre de atributo	*Tipo de atributo*	*Descripción*
Región de residencia actual	Nominal	La región socioeconómica en donde se reside actualmente. 1 = Region Central 2 = Region Chorotega 3 = Region Pacifico Central 4 = Region Brunca 5 = Region Huetar Caribe 6 = Region Huetar Norte
Región de residencia hace dos años	Nominal	La región socioeconómica donde se residía hace dos años
Condición de migrante	Nominal	Condición de migrante del individuo. 0= No migrante, 1=Migrante interno, 2=Migrante externo. Para el caso de estudio actual se estará seleccionando el migrante interno
Ingreso por persona neto	Cuantitativo continuo	Ingreso por persona con deducciones
Condición de actividad	Nominal	Si la persona es retirada o si tiene empleo o no
Título	Ordinal	Último grado académico obtenido
Sexo	Nominal	1 = Masculino 2 = Femenino
Estado conyugal	Nominal	1 = En unión libre o juntado(a) 2 = Casado(a) 3 = Divorciado(a) 4 = Separado(a) 5 = Viudo(a) 6 = Soltero(a) 9 = Ignorado
Ocupación de trabajo	Nominal	1 = Directores y gerentes 2 = Profesionales científicos e intelectuales 3 = Técnicos y profesionales de nivel medio 4 = Personal de apoyo administrativo 5 = Trabajadores de los servicios y vendedores de comercios y mercados 6 = Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros 7 = Oficiales, operarios y artesanos de artes mecánicas y de otros oficios 8 = Operadores de instalaciones y máquinas y ensambladores, 9 = Ocupaciones elementales 10 = No bien especificadas

Los datos anteriores fueron analizados en la herramienta de JMP para identificar las variables que brindaban información interesante a nivel del tema de migración en Costa Rica. También se encontró que ciertas variables como el nivel de pobreza no aportaban información interesante para el análisis por lo cual es necesario realizar una limpieza de

los datos, eligiendo aquellas columnas que brindan información interesante a nivel de las personas que migran.

Como se muestran en las figuras 8, 9 y 10, las distribuciones se mantienen bastante similares a través de los años para el conjunto de datos seleccionado. La mayor cantidad de los entrevistados que se movieron de región cuentan con al menos la secundaria completada, cuentan con un trabajo; por otra parte, al menos un 60% (en todos los años estudiados) cuenta con un salario menor a 300.000 colones.

También se creó otra columna que se llama "*mantiene_familia*" la cual se crea de acuerdo con si la persona mantiene o no a la familia y la capacidad económica que tiene para apoyar a la familia económicamente. Al menos 40% en todos los años mantienen una familia con un salario menor a 500.000 colones.

De esta manera, tomando en cuenta estas distribuciones y las relaciones entre las variables se asignó a cada una de ellas un peso que más adelante será necesario para calcular las distancias respectivas para su posterior análisis difuso.

Figura 8 Análisis de distribuciones conjunto de datos migraciones en Costa Rica 2014. Datos obtenidos a partir del Catálogo Central de datos de la INEC (INEC, Catálogo Central de datos INEC, 2008)

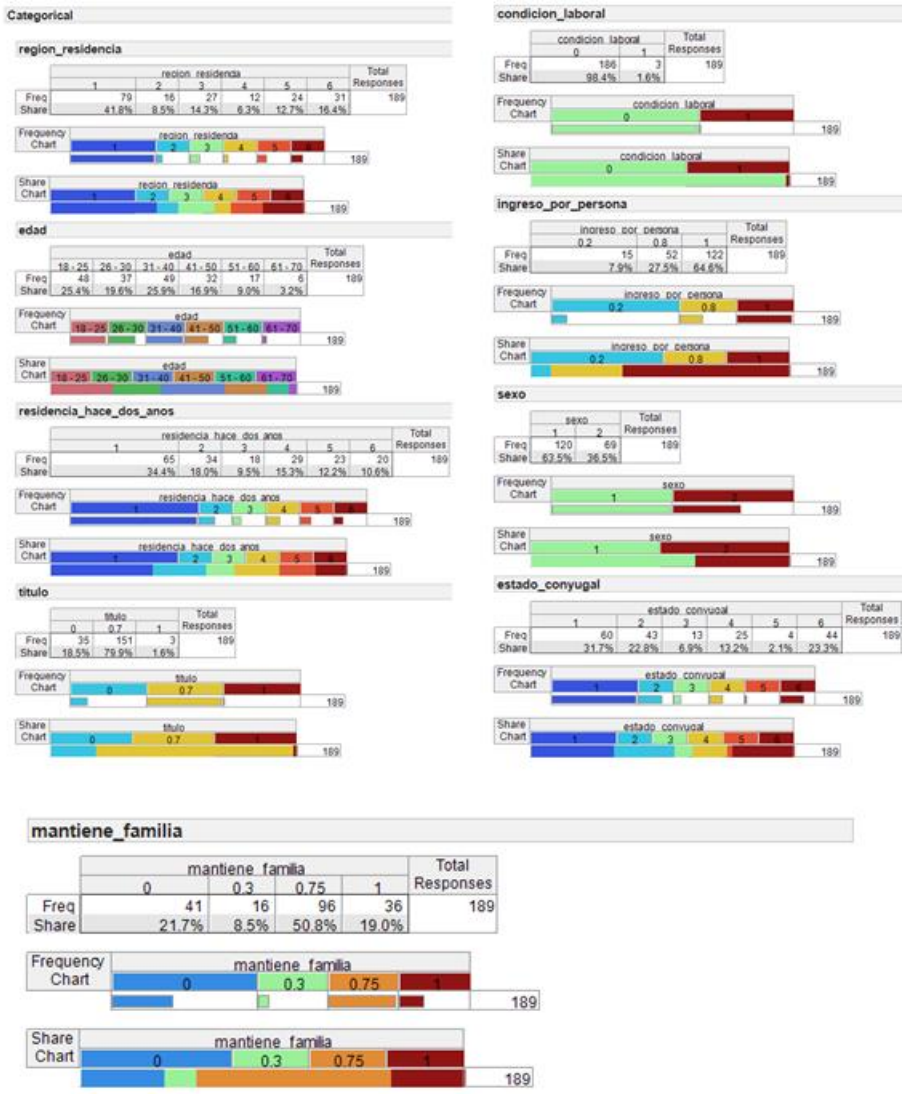

Figura 9 Análisis de distribuciones conjunto de datos migraciones en Costa Rica 2015. Datos obtenidos a partir del Catálogo Central de datos de la INEC (INEC, Catálogo Central de datos INEC, 2008)

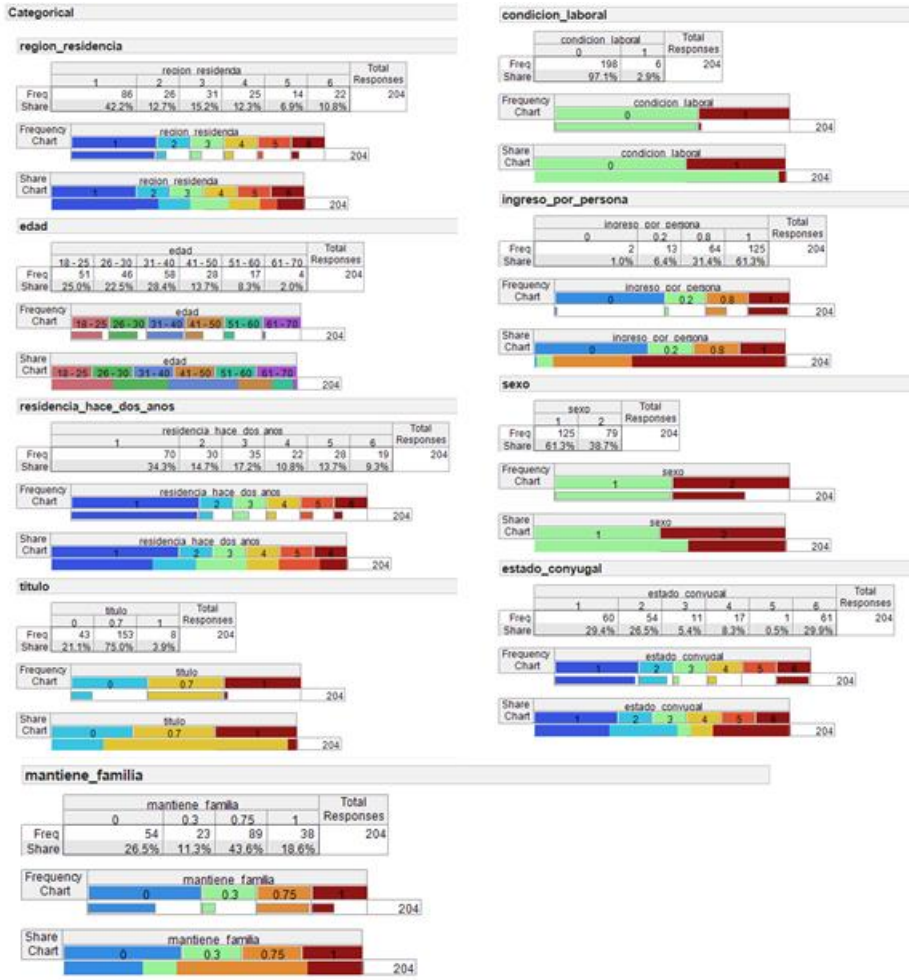

Figura 10 Análisis de distribuciones conjunto de datos migraciones en Costa Rica 2016. Datos obtenidos a partir del Catálogo Central de datos de la INEC (INEC, Catálogo Central de datos INEC, 2008)

Una vez que se contó con una mayor comprensión de los datos y la relación entre sus variables, se presenta a continuación el proceso que se siguió para analizar y procesar los datos.

Capítulo 3. Propuesta de una visualización de redes sociales mediante un enfoque difuso

3.1 Proceso general

En la figura 11 se muestra el proceso general que se ha cubierto en este proyecto desde el proceso de datos hasta lo necesario para generar la visualización. Se describe con más detalle cada uno de los elementos en las siguientes subsecciones.

1. *Análisis y modelado de datos*: Se hace un análisis profundo de los datos, realizando un preprocesamiento de los mismos (realizar la limpieza respectiva, seleccionar atributos, discretizar y calcular matriz distancia). Además, entender la manera en que se comportan los datos para su agrupación.
2. *Clasificación* utilizando algoritmo de clúster difuso
3. *Visualización* con redes sociales como base
4. *Validación de resultados*: Se analiza la visualización obtenida y se detectan patrones comparación estadística de los resultados con respecto a la salida esperada según la propuesta inicial según la hipótesis.

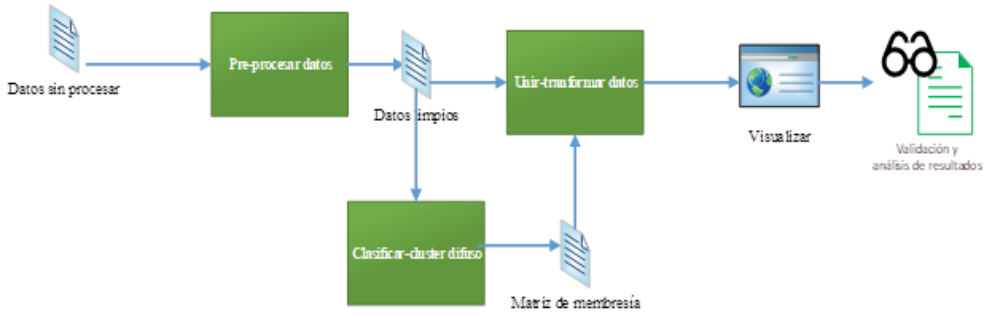

Figura 11. Proceso de minería de datos de alto nivel

3.2 Limpieza de los datos

Los archivos originalmente vienen en formato tipo SPSS (*Statistical Package for Social Sciences*) y se tuvo que convertir este archivo a CSV (archivo separado por comas) para facilitar el preprocesamiento de los datos. Inicialmente se filtraron los datos de acuerdo con una serie de reglas:

- Los datos por analizar se definieron solo de individuos con condición migratoria.
- Además de tener condición migratoria se debe cumplir que solo se haya dado de manera interna y no internacionalmente.
- Solo se tomaron en cuenta los datos donde la edad sea de 18 años o más (con capacidad de laborar)

Luego, se procedió a importar la biblioteca de pandas en Python:

```
import pandas as pd
```

Una vez filtrado los datos se remueven las filas con valores vacíos y se obtiene una muestra aleatoria del 30% de los datos. En pandas se logra obtener la muestra de la siguiente manera:

```
migrantes = pd.DataFrame(data_migrantes)

migrantes.sample(frac=0.2)
```

Se le asigna pesos a categorías que pueden ser posibles razones por las cuales una persona migra de una región a otra.

Se aplica *binning* (agrupamiento por intervalos) para discretizar *el ingreso por persona* y se utilizan los siguientes intervalos:

- *Ingreso bajo* (peso asignado: 1): Salario menor a 300.000 colones
- *Ingreso medio bajo* (peso asignado: 0.8): Salario entre 300.0000 y menos de 1.500.0000 colones

- *Salario medio alto* (peso asignado: 0.3): Salario entre 1500.000 menos de 3.000.000 colones
- *Salario alto* (peso asignado: 0): Salario mayor a 3.000.000 colones

Para crear el agrupamiento (bins) en pandas se definen los rangos en un arreglo de la variable *bins* como se muestra en el código. De la misma manera se crean para la columna de edad.

```
bins = [0, 300000, 700000, 1500000, 3000000]
group_names = ['Ingreso_Bajo', 'Ingreso_Medio_Bajo', 'Ingreso_Medio_Alto', 'Ingreso_Alto']
ingreso_discreto = pd.cut(df_migrantes['itpn'], bins, labels=group_names)
df_migrantes['ingreso_discreto'] = pd.cut(df_migrantes['itpn'], bins, labels=group_names)
```

Para crear las columnas con los pesos respectivas se realiza de una manera similar, en el arreglo de grupo de nombres se puede notar que entre más cercano es el valor a 1, quiere decir que la condición socioeconómica más baja es más fuerte para cuando se utilice la variable para la clasificación del clúster difuso.

```
bins = [0, 300000, 700000, 1500000, 3000000]
group_names = ['1', '0.8', '0.2', '0']
peso_ingreso_discreto = pd.cut(df_migrantes['itpn'], bins, labels=group_names)
df_migrantes['peso_ingreso'] = pd.cut(df_migrantes['itpn'], bins, labels=group_names)
```

De la misma manera se aplica esta técnica para la columna de *edad*

- Joven Adulto: entre 18 y 24 años
- Adulto: entre 24 y 60 años
- Ciudadano de oro: más de 60 años

Se limita la categorización de la educación en 3 categorías:

- Sin educación (peso asignado: 1)
- Primaria completada: (peso asignado 0.9)
- Secundaria completada (peso asignado: 0.7)
- Universidad completada (peso asignado: 0)

En el caso de la educación se define valores de manera ordinal y se agrega un peso para cada uno de los posibles valores los cuales van a influir en el cálculo del clúster difuso, es decir, para cuando se aplique el algoritmo de Gower y así determinar las distancias.

```
df_migrantes['titulo'] =
    np.where(df_migrantes['NivInst'] == 0, 'sin_educacion',
    np.where(df_migrantes['NivInst'] < 5, 'primaria_completada',
    np.where(df_migrantes['NivInst'] < 7, 'secundaria_completada',
    np.where(df_migrantes['NivInst']   <   9,   'universidad_completada',
'Desconocido'))))
df_migrantes['peso_titulo'] =
    np.where(df_migrantes['NivInst'] == 0, 1,
    np.where(df_migrantes['NivInst'] < 5, 0.9,
    np.where(df_migrantes['NivInst'] < 7, 0.7,
    np.where(df_migrantes['NivInst'] < 9, 0, 0))))
```

Se limita la categorización del estado de ocupación en 3 categorías:

- Fuera de la fuerza laboral (peso asignado: 1)
- Retirado (peso asignado: 0.7)
- Ocupado (peso asignado: 0)

La columna *mantiene_familia* revisa la columna A23 (Mantiene económicamente el hogar) y revisa la capacidad económica de la persona para asignar un valor entre 1 y 0 (Si mantiene a la familiar y el salario es menor a 200.000 el valor *mantiene_familia* es igual a 1; si el salario es mayor a 800.000 el valor es igual a 0; de lo contrario se le asigna un valor entre 0 y 1)

```
   df_migrantes['mantiene_familia'] =
```

```
np.where( (df_migrantes['A23'] == 1) &
(df_migrantes['itpn'] < 200000), 1,
np.where((df_migrantes['A23'] == 1) &
(df_migrantes['itpn'] < 500000), 0.75,
np.where((df_migrantes['A23'] == 1) &
(df_migrantes['itpn'] < 800000 ), 0.3, 0)))
```

Una vez aplicada la limpieza se guardan los datos en un archivo separado por comas, en la tabla 3 se presenta una parte de los datos de ejemplo que se obtienen.

Tabla 3. Datos migratorios para análisis difuso

Región Residencia	***Edad***	***Residencia hace dos años***	***Titulo***	***Condición laboral***	***Ingreso por persona***
2	Adulto	1	Universidad_completa	Ocupado	Ingreso_Medio_Bajo
4	Adulto	1	Secundaria_completa	Ocupado	Ingreso_Bajo
2	Adulto	5	Secundaria_completa	Ocupado	Ingreso_Bajo
3	joven_adulto	4	Secundaria_completa	Ocupado	Ingreso_Bajo
6	Adulto	1	Universidad_completa	Ocupado	Ingreso_Medio_Alto
3	Adulto	1	Universidad_completa	Ocupado	Ingreso_Bajo
5	ciudadano_oro	1	Secundaria_completa	Retirado	Ingreso_Bajo

Con este preprocesamiento de los datos, es posible proceder con el cálculo de la matriz de distancias que será utilizada en el análisis de clústeres difuso. Como se detalla en la siguiente sección.

3.3. Distancia de Gower

Se analizaron diferentes opciones de distancias, entre ellas, se encuentra la distancia de *Ahmad and Dey* la cual es calculada mediante el cálculo de la coocurrencia de los valores del atributo (para el cual se calcule la distancia) con los valores de otros atributos, pero por otra parte solo se calcula entre dos valores categóricos. (Dey, 2007). Asimismo, se encontró esa misma limitación en otras distancias existentes para solo datos continuos como es el caso de la distancia euclidiana y Manhattan. Así, se encontró que la distancia de Gower es

más apropiada para calcular la matriz de distancias para datos mixtos multi-variables, además, permite utilizar pesos para darle importancia a ciertas variables para el cálculo de la distancia. (Hoven, 2015)

Como se mencionó anteriormente y al contar con datos mixtos, Gower se convirtió en la distancia que mejor se adapta a los datos de migración interna.

La *distancia de Gower* compara dos casos *i* y *j,* y se define según la ecuación (2) según (Hoven, 2015)

$$S_{ij} = \frac{\sum_k w_{ijk} \times S_{ijk}}{\sum_k w_{ij}} \quad (2)$$ (Gower, Dec., 1971)

Donde: s_{ijk} denota la contribución dada por la variable *k*-ésima y

w_{ijk} utiliza 1 o cero dependiendo de si la comparación es válida para la variable *k*-ésima; 0 si la comparación no es válida o si se especifican pesos variables.

El concepto de la distancia de Gower se calcula de acuerdo con el tipo de cada variable, se selecciona la distancia más adecuada de acuerdo con el tipo utilizado y el resultado es una matriz de distancias con valores de entre 0 y 1 según (Gower, Dec., 1971).

Las distancias que se utilizan en Gower son las siguientes:

- Para valores cuantitativos (ordinales y continuos): se utiliza la distancia de Manhattan con una modificación como se puede notar en la fórmula (3) donde se divide entre el rango de valores de la *k*-ésima variable (r_k), que sería el rango de la población o de una muestra.

  $$S_{ijk} = \frac{1 - |X_{ik} - X_{jk}|}{r_k} \quad (3)$$

  Los valores de S_{ijk} oscilan entre 1 cuando $X_{ik} = X_{jk}$ y 0 para los valores que se encuentran en los lados opuestos $X_{max} - X_{min}$

- Valores cualitativos (nominales): Se define $S_{ijk} = 1$ si los dos casos *i* y *j* son iguales o 0 en el caso de que sean diferentes.

Para la generación de matriz de disimilitudes se utiliza los conceptos anteriores y la biblioteca *pdist* la cual se modificó para que acepte datos mixtos. En la figura 12 se muestra el pseudocódigo de la implementación de la distancia de Gower.

```
FUNCTION    distancia_gower(vector_i,    vector_j,    rangos_datos_mixtos,
tipo_de_datos, peso):


    //rangos_datos_mixtos= vector que se obtiene al calcular Xmax - Xmin
    suma_objeto_ij=0.0
    suma_peso_ij=0.0
    for columna in columnas:
        objeto_ij=0.0
        peso_ij=0.0
        IF tipo_de_datos[columna] es numero:
            objeto_ij=valor_absoluto(vector_i[columna]-
vector_j[columna])/(rangos_datos_mixtos[columna])
            peso_ij=peso[columna]
        ELSE:
            suma_objeto_ij=(1,0)[vector_i[columna]==vector_j[columna]]
            peso_ij=(peso[columna])
        ENDIF
        suma_objeto_ij+= (peso_ij*objeto_ij)
        suma_peso_ij+=peso_ij
    ENDFOR
    RETURN suma_objeto_ij/suma_peso_ij
ENDFUNCTION
```

Figura 12 Pseudocódigo de Distancia de Gower

Para verificar que el algoritmo de la distancia de Gower haya sido exitoso y que los datos tengan sentido, se revisa los valores más similares y disimilares de la matriz. Para el caso actual se obtuvieron los siguientes pares:

Par de elementos más similares:

Tabla 4 Par de elementos más similares

Título	*Condición laboral*	*Ingreso por persona*	*Estado conyugal*	*Calidad vivienda*	*Mantiene familia*
0.7	0	1	1	0	1
0.7	0	1	1	0	0.75

Par de elementos más disimilares:

Tabla 5 Par de elementos más disimilares

Título	*Condición laboral*	*Ingreso por persona*	*Estado conyugal*	*Calidad vivienda*	*Mantiene familia*
1	0	1	1	0.9	1
0	0	0.2	6	0	0

Con la matriz de distancias de Gower se tiene una de las entradas necesarias para poder calcular el clúster difuso utilizando el algoritmo *fuzzy c-means.*

3.4. Algoritmo de *fuzzy c-means*[1]

En la sección anterior se habló de la distancia que se utilizó en los datos de migración, la distancia de Gower, la matriz que se generó se utilizó como entrada para el cálculo de membresías de cada uno de los clústeres utilizando el algoritmo de *fuzzy c-means.*

La figura 13 muestra la implementación general del algoritmo difuso.

[1] *El algoritmo implementado para fuzzy c-means permite utilizar distancias personalizadas o las que se encuentran actualmente dentro de la biblioteca de Scipy (Python) o en R (se probaron ambas opciones).*

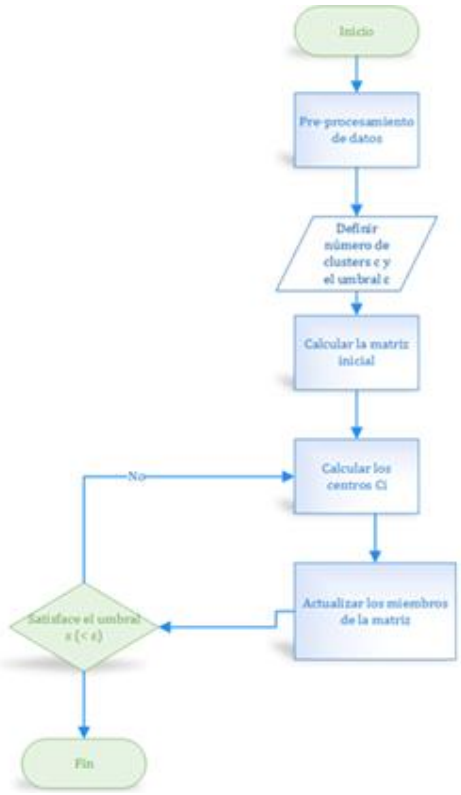

Figura 13 Proceso de difusidad

Utilizando el diagrama anterior como base entonces, se procedió a implementar el algoritmo difuso de la siguiente manera.

Definición de número de clústeres: La cantidad de clústeres se define al número 4 ya que fue el que brindaba mejor información de los datos. El umbral ε será un valor entre [0,1], cuanto más bajo sea el valor de ε se obtendrán mejores resultados, pero, se sacrificaría el desempeño debido a que la cantidad de iteraciones aumentaría.

El número de clústeres óptimo se eligió de acuerdo con el análisis de la "s*ilueta*" después de correr el algoritmo varias veces con diferentes variables analizando los resultados que brindaba el algoritmo. El análisis de silueta mide qué también está agrupada una observación y estima el promedio de distancia entre las agrupaciones. Los gráficos 20 y 21 representan la silueta de los clústeres para $k = 3$ y para $k = 4$ respectivamente. El gráfico muestra que tan cercano está cada uno de los puntos del clúster a sus clústeres *vecinos*. Para cada observación i se calcula el ancho de la silueta s_i de la siguiente manera según (Malika Charrad, 2014):

1. Por cada observación *i* se calcula el promedio de la disimilitud (*a*) entre *i* y los otros puntos del clúster al cual pertenece la observación.
2. Para los otros clústeres *C*, a los cuales *i* no pertenece se calcula el promedio de disimilitud entre la observación y el clúster *C* que se define como $b_i = min_C\, d(i, C)$ donde el valor de b_i se denota como la disimilitud entre *i* y su clúster "*vecino*", es decir, el clúster más cercano al que no pertenece o pertenece solo en cierto porcentaje (cuando s_i tiene un valor cercano a 0)
3. El ancho de la silueta de la observación *i* entonces se define como:
   $S_i = (b_i - a_i)/max(a_i, b_i)$.

En el lenguaje *R* se obtiene la silueta del clúster al proveer los resultados del cálculo de la distancia de Gower y el resultado del clúster difuso como entrada.

```
plot( sillhouete(fuzzy_cluster_result, gower_dist))
```

Volviendo a la figura 14 se muestra cómo para $k = 3$ se tiene un error de agrupación para el año 2016 en el clúster número 2, ya que el valor de S_i fue negativo. Además, analizando los datos como se había verificado al revisar las distribuciones, debería de haber un grupo que sea visiblemente más bajo, en este caso los grupos de personas con capacidad socioeconómica alta y medio alta. Por lo que no se estaba obteniendo los resultados requeridos. Por otra parte, en la figura 15 se puede notar cómo las siluetas se distribuyen mejor, para los casos en los que los elementos están más cercanos a cero se refiere a las observaciones que pueden pertenecer a dos o más clústeres y en los casos en los que se encuentran más cerca del valor de 1, indica que están correctamente agrupados en el clúster indicado.

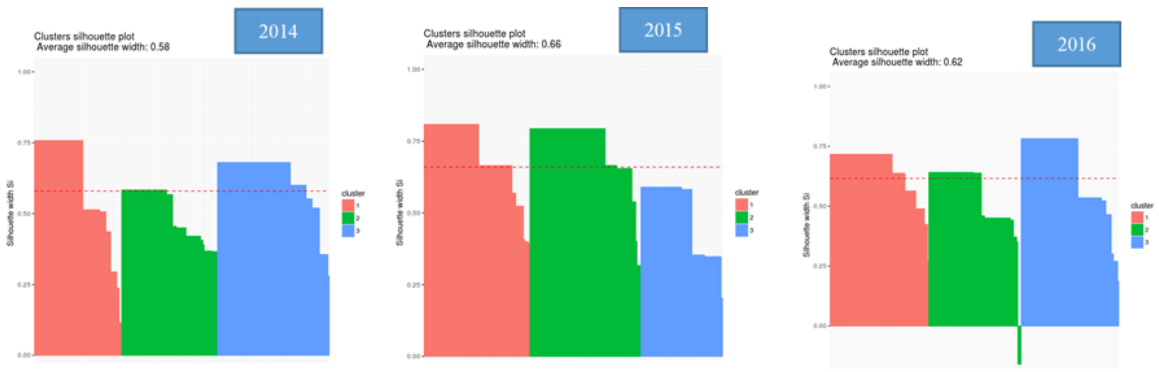

Figura 14. Gráfico de siluetas, k=3

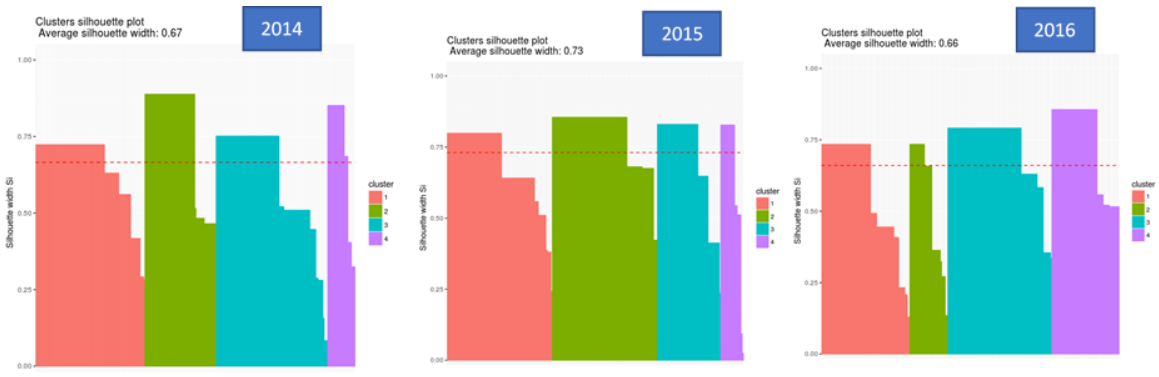

Figura 15. Gráfico de siluetas, k=4

Una vez se indica los valores de entrada iniciales se procede a realizar el cálculo de la matriz inicial con valores aleatorios. A continuación, se muestra la manera en que se calculó la matriz inicial.

Cálculo de la matriz inicial: Se debe inicializar la matriz $U=[u_{ij}]$, se calcula utilizando la ecuación (4):

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \quad (4)$$

Cálculo de centros C_i: Calcular los vectores de centros utilizando la ecuación (5).

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (5)$$

La implementación que se mostrará a continuación se logra utilizando la biblioteca de *numpy* (*np*). Numpy permite manipular fácilmente arreglos multidimensionales y a las dimensiones se les llaman *axis*. A continuación, se explica la forma en que se implementó el algoritmo. Con el método de *np.power* se calcula el valor exponencial del arreglo de membresías al grado de difusidad definido por el usuario, el cual se utilizará luego en el cálculo de centros, con *np.dot* el cual permite multiplicar matrices ($u_{ij}^{m} \cdot x_i$) y dividir entre la sumatoria de membresías, ecuación (5).

```
def fuzzifier(self, memberships):
    return np.power(memberships, self.m) #m es el grado de difusidad

def calculate_centers(self, conjunto_datos):
    return np.dot(self.fuzzifier(self.memberships).T, conjunto_datos) / \
           np.sum(self.fuzzifier(self.memberships).T,axis=1)[...,
np.newaxis]
```

Actualizar miembros de matriz: Utilizando nuevamente la ecuación (4) se recalculan la membresía al clúster y aplicando los conceptos que se explicaron anteriormente.

```
def calculate_memberships(self, x):
    distance_gower_result = self.distances(x)
    return np.sum(np.power(
        np.divide(distance_gower_result[:,:,np.newaxis],
distance_gower_result[:, np.newaxis, :]),
        2 / (self.m - 1)), axis=2) ** -1
```

El pseudocódigo del proceso general del cálculo de clúster difuso se demuestra en la figura 16.

```
FUNCTION fuzzy_cluster(conjunto_datos, numero_clusters, matriz_distancias,
numero_maximo_iteraciones):
    Inicializar(matriz_membresia, centroides)
    //se inicializa con valores aleatorios
    i=0
    WHILE i < numero_maximo_iteraciones:
        calcular nuevos centroides utilizando ecuación (5)
        nueva_matrix=matriz_membresia
        calcular nueva matriz utilizando ecuación (4) y resultados de matriz
de distancia obtenida de Gower (2)
        i=i+1
        IF Maximo(matriz_membresia - nueva_matrix) < epsilon:
            //Terminar proceso
        ENDIF
    ENDWHILE
ENDFUNCTION
```

Figura 16 Pseudocódigo para la implementación de fuzzy c-means

Al finalizar el proceso del algoritmo difuso se devuelven dos resultados: 1) una matriz de membresía lo que contiene el grado al cual cada una de las observaciones pertenece a un clúster dado y 2) el coeficiente de la partición de Dunn el cual mide qué tan cerca es la solución difusa a la solución de *k*-means; si el valor es bajo indica que es más difuso, y si es cercano a 1 significa que está más cerca de la solución de *k-means*. Para calcularlo se toma la suma de todos los coeficientes de las membresías (*m*) al cuadrado dividida por el número de observaciones (*N*). Por lo tanto, la fórmula se define de la siguiente manera: (Statistical, 2017)

$$F(U) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} m_{ik}^{2}$$

En este caso se eligió un *m* ligeramente menor a 2 lo que disminuye el nivel de difusidad, por otro lado, esta variable se puede probar con diferentes valores hasta encontrar uno que se ajuste mejor a los datos que se están procesando. Se puede notar al observar las relaciones de los clústeres en las figuras 17 y la figura 18 un comportamiento más difuso cuando *m* es igual 2 (y el coeficiente de Dunn es más bajo) en contraste a *m* igual 1.7. Para efectos de este estudio se llegó a tener una mejor clasificación con *m* = 1.7.

Tabla 6. Resultados Fuzzy Clustering usando m=1.7

M	**1.7**
Objetivo	23.57
Iteraciones	39
N	271
Coeficiente de Dunn	0.5120022

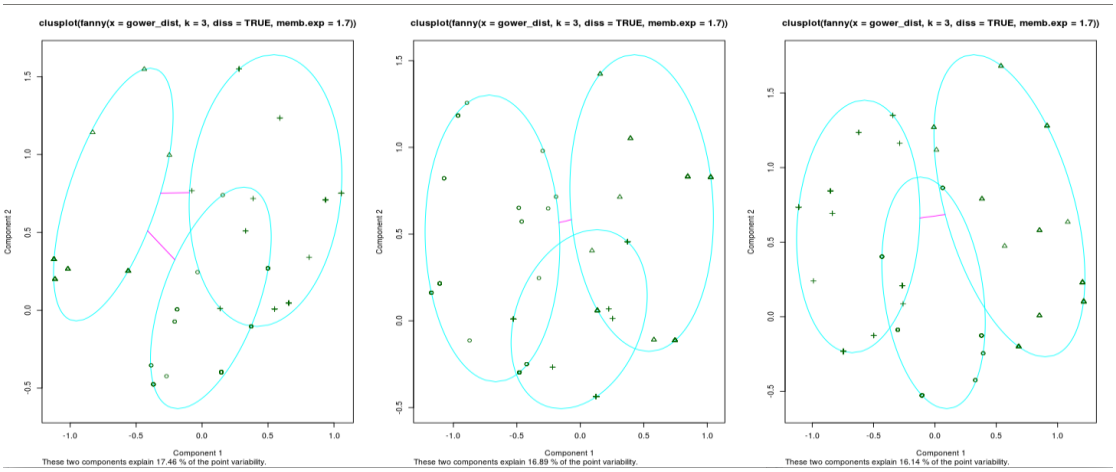

Figura 17 Fuzzy Cluster utilizando m = 1.7

Tabla 7 Resultados Fuzzy Clustering usando m=2

M	**2**
Objetivo	17.83
Iteraciones	40
N	271
Dunn Coefficient	0.4013675

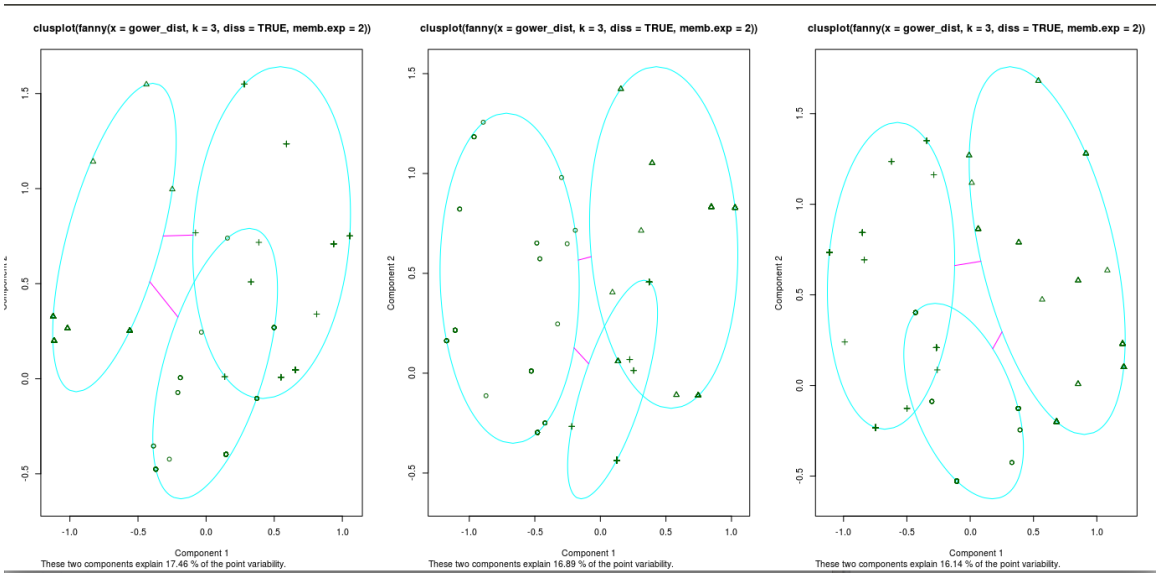

Figura 18 Fuzzy Cluster utilizando m = 2

Finalizando este paso se obtiene la agrupación de cada uno de los elementos de manera difusa, es decir, cada uno tiene un nivel de membresía asignado por cada uno de los cuatro clústeres que fueron definidos. En la próxima sección se procede a visualizar los clústeres en una red social con enfoque difuso.

Capítulo 4. Implementación de la propuesta

4.1. Visualización de red social

4.1.1 Redes sociales jerárquicas

Se define el *grafo* $G = (V, E)$ como el conjunto de vértices y aristas $E \subseteq V \times V$, donde cada arista $e_j \in E$ se le asigna un peso(w): $w_{e_j} \in \mathbb{R}$. El peso en este caso sería la población que emigra hacia otras regiones del país.

Las redes sociales con enfoque difuso que se exploran en este proyecto también poseen formato jerárquico. La razón por la cual se hizo de esta manera es debido a que se buscó la mejor forma de agrupar los datos, en este caso, de acuerdo con el lugar dónde reside cada uno de los individuos.

En la implementación actual se muestra la agrupación por región, pero también se podría incluir otros niveles (desde el nivel más alto hasta el nivel más bajo) según la disponibilidad de los datos, por ejemplo, Provincia → Cantón → Distrito → Individuo. A continuación, se muestra el documento JSON que se generó al unir y procesar los datos de membresía obtenidos como producto del algoritmo difuso, así como los datos de regiones e individuos obtenidos a partir de las estadísticas de (INEC, Catálogo Central de datos INEC, 2008), estos últimos permiten brindar información adicional que ayudará en el análisis de la visualización.


```
{
    "nodes": [{"id": "Region Central",
            "inmigracion_total": 71,
            "emigracion_total": 99,
        "children": [
        {
            "id": "element_214",
            "titulo": "secundaria_completada",
            "ingreso_por_persona": "medio_bajo",
             "condicion_laboral" : "Ocupado",
```

```
                "membership": [ {
                    "percent": 0.06953217343210001
                }, {
                    "percent": 0.88512507829
                }, {
                    "percent": 0.045342748278
                }]
            },
            }],
    "links": [{
            "source": "Region Chorotega",
            "length": 500,
            "target": "Region Central",
            "width": 16
        }, {
            "source": "Region Pacifico Central",
            "length": 500,
            "target": "Region Central",
            "width": 13
}
```


Como se puede notar el archivo de JSON tiene una estructura anidada. En este caso se utilizó los archivos de CSV como la entrada para poder generarlo, se transformó a un archivo de JSON anidado utilizando la solución que se presenta a continuación.

En el código a continuación se hace la unión de los datos, en este caso, sería el resultado de los datos de migrantes que se realizó después de la limpieza y transformación junto con los datos de la membresía del clúster. Con la función *concat* de pandas se logra hacer la unión de los dos conjuntos de datos con las columnas requeridas para ser convertidos a un JSON anidado.

```
df_data = pd.read_csv( datos_migrantes,index_col=False, header=0,
                       usecols=['region_residencia',
                                'residencia_hace_dos_anos',
                                'condicion_laboral',
                                'ingreso_por_persona',
                                'titulo',
                                'mantiene_familia'])
df_memb = pd.read_csv( datos_membresia,index_col=False, header=0,
                       usecols=["Cluster1","Cluster2",
                                "Cluster3", "Cluster4"] )
df_output = pd.concat([df_data, df_memb], axis=1)
```

Una vez que los resultados están unidos se empiezan a leer los datos del csv para adaptarlos de una manera que sea simple para después convertirlos a un documento JSON anidado.

Esto nos brindaría la información de los nodos, pero por otra parte todavía faltarían los enlaces. Para el caso de los individuos dentro de las regiones ya los enlaces se crean debido a la relación jerárquica. Por otra parte, también es necesario agregar las relaciones a nivel de los movimientos migratorios entre las regiones. Como se muestra en el código se crea una tabla contingencia con la función *crosstab* que ofrece la biblioteca de pandas para revisar las frecuencias entre la región de residencia actual y la región de residencia hace dos años. Con esta tabla se obtiene además de la región fuente y la región del destino, el ancho (*width*) que va a tener el enlace entre ambas regiones.

```
df = pd.crosstab(links['region_residencia'],
links['residencia_hace_dos_anos'])
print df
link_array = []
for  column in df:
    for index, row in df.iterrows():
        if row[column] != 0:
            link_array.append({"source": region_names[int(index)-1],
"target": region_names[column-1], "length":500, "width": row[column]})
```

4.1.2 D3.js y la visualización

Por otra parte, existía el problema de que la visualización aún era difícil de comprender debido a la manera en que se mostraban los grafos, no había un orden y no se mostraba de una manera limpia y fácil de entender al usuario. Por lo que se encontró otra manera de desplegar los datos de manera que las agrupaciones no se intersecaran entre sí.

Se utilizó D3js que es una biblioteca de JavaScript que brinda las herramientas necesarias para crear visualizaciones personalizadas, D3 además enfatiza el énfasis en estándares web que permite que funcione en navegadores modernos combinado poderosos componentes de visualización. Esta biblioteca viene a resolver los problemas de agrupación a través de un método llamado *force layout* el cual utiliza una simulación física.

Con simulación física nos referimos a la manera en que permite D3js dar a cada uno de los grupos de grafos la capacidad de repelerse entre sí para evitar que se traslapen al configurar el parámetro *charge* a un valor negativo. También brinda otras características útiles como las de mantener la visualización en el centro de la página a través de un parámetro llamado *Gravity*.

Para crear la visualización se generan los nodos que van a ser pie charts. Como se muestra en la figura 19, los pies se crean calculando cada uno de los arcos de acuerdo con los porcentajes definidos en el archivo JSON. Se genera el nodo utilizando la biblioteca de Javascript D3Js, para crear y rellenar cada uno de los arcos y se calcula la circunferencia del semi-círculo: $C = 2\pi\frac{r}{2}$ y con el porcentaje se calcula el segmento específico que se quiere dibujar.

```
function drawPieChart(nodeElement, percentages, options) {
    var radius = getOptionOrDefault('radius', options);
    var halfRadius = radius / 2;
    var halfCircumference = 2 * Math.PI * halfRadius;
    var classo = getOptionOrDefault('classo', options);
    var percentToDraw = 0;
    var color_arc =0;
    var range = ["#1A5FFF","#FFD500","#FF7100"];
    var cluster = ["cluster1", "cluster2", "cluster3"];
    for (var p in percentages) {
        percentToDraw += percentages[p].percent;
        nodeElement.insert('circle', '#parent-pie + *')
            .attr("r", halfRadius)
            .attr("fill", 'transparent')
            .attr("class", classo)
            .style('stroke', range[color_arc])
            .style('stroke-width', radius)
            .style('stroke-dasharray',
                    halfCircumference * percentToDraw
                    + ' '
                    + halfCircumference);

            color_arc++;
    }
}
```

Figura 19 Función para crear los nodos que serán parte de la red social con enfoque difuso

Los enlaces se generan de acuerdo con el archivo de JSON procesado de acuerdo con los datos obtenidos de la encuesta acerca de los lugares destino u origen de los individuos.

Se agrupan de acuerdo con las regiones de Costa Rica, como se puede observar en la figura 20, donde se podría definir una jerarquía aún más grande que la que se menciona.

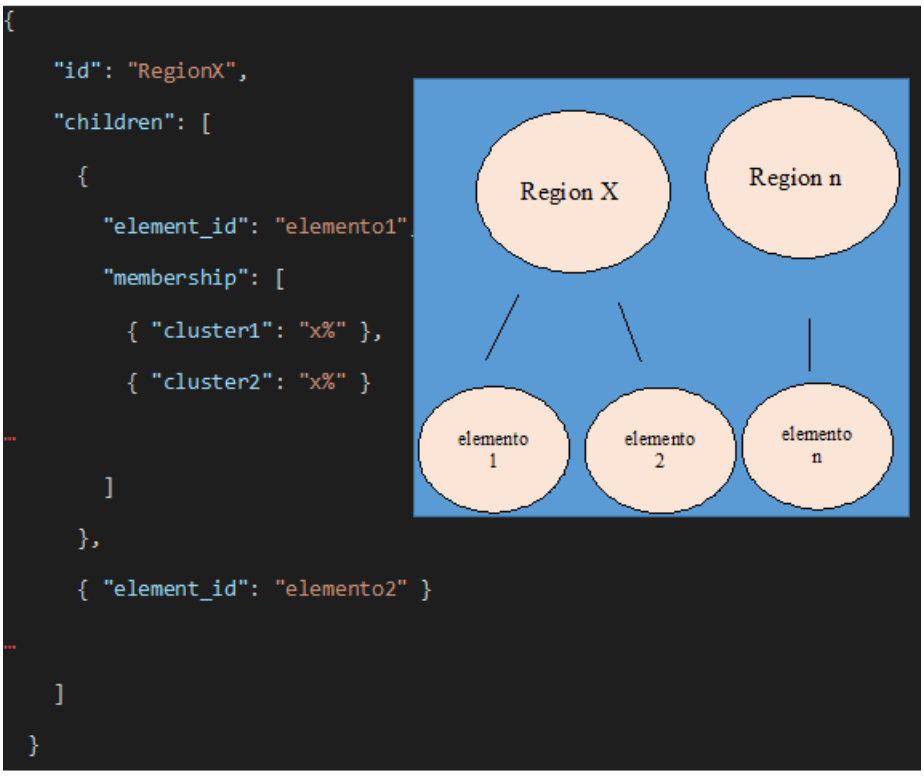

Figura 20 Representación de grafos red social de acuerdo con el JSON creado

Para identificar de qué región proviene cada elemento se agrega un arco del color de la región correspondiente como se indica en la figura 21. La flecha roja indica el nodo que provee la información del nombre de la región y el color por el cual se identifica. Las flechas verdes en el lado izquierdo indican de cuál región provienen esos dos elementos. En este caso, de acuerdo con esta visualización, existen dos individuos que provenían de la región Huetar Norte y que migraron a la región Chorotega.

Los individuos (nodos hijos - hojas) no poseen etiqueta y tienen un tamaño menor que los nodos padres para poder tener un mejor contraste; las regiones (nodos padres) tienen una etiqueta que los identifica, también cambia el tamaño del nodo padre de acuerdo al resultado de la migración neta, entre mayor sea el valor el radio del nodo también será afectado provocando el crecimiento del nodo.

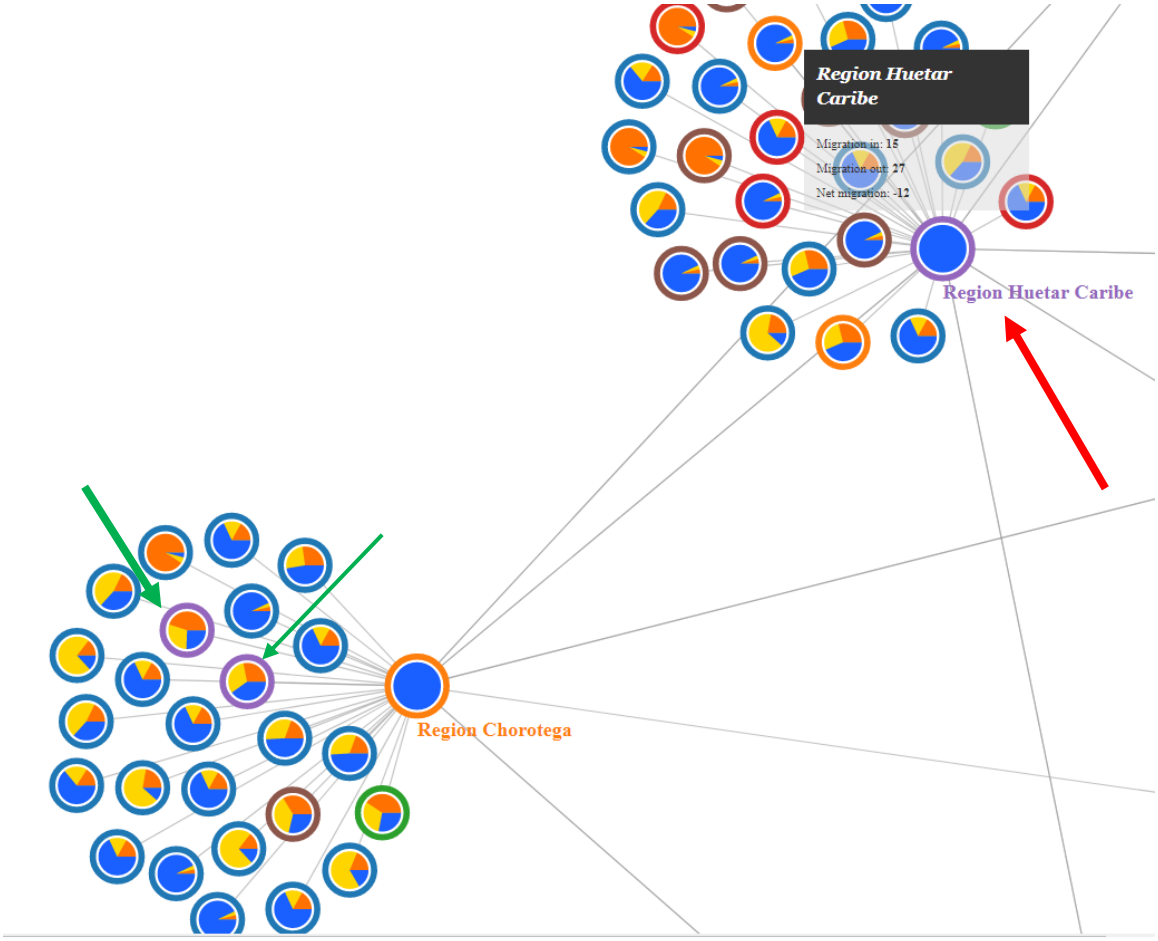

Figura 21 Relaciones entre grafos

Además, para aprovechar la naturaleza del *force layout*, en lugar se agregó a los nodos de región una mejor manera de representar las distintas agrupaciones, donde se le brinda mayor fuerza de atracción a aquellos nodos que tienen un porcentaje de membresía más alto para así también mostrar una visualización más ordenada y fácil de comprender. Por ejemplo, en el caso de la figura 22, se puede notar cómo los colores morados más claros están más cerca del nodo de región que tiene un porcentaje mayor de este tipo de agrupación, mientras que los morados más oscuros se encuentran más lejos del nodo.

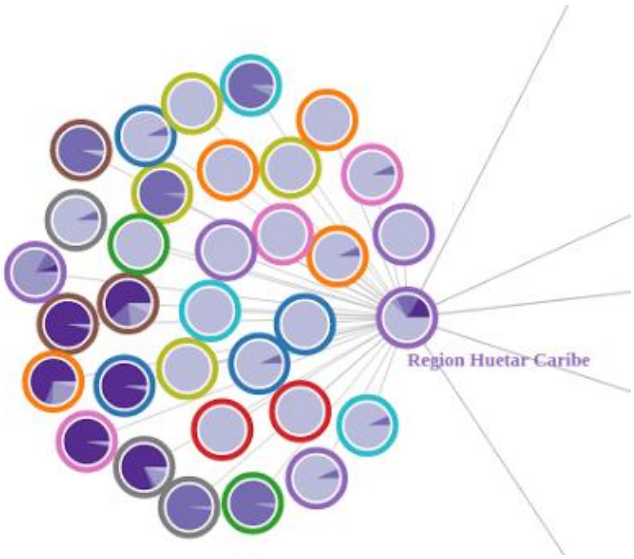

Figura 22 Fuerza de atracción a nodo padre de acuerdo con el clúster al que pertenecen los nodos hijo

Además, en la figura 23 se puede notar las diferencias en el análisis de un algoritmo tipo k-means (binario) con respecto al resultado del algoritmo difuso, para el elemento 133, se puede notar que es una persona que está en secundaria y que no trabaja, aunque el algoritmo difuso lo clasifica como una persona con buena posición económica esto se da debido a que es una persona no asalariada que aún depende de los padres, aunque también pertenece a la agrupación socioeconómica medio baja, ya que el nivel de educación es bajo de acuerdo a la edad del individuo (la muestra se realizó solo para edades mayores a 18 años), en el caso del algoritmo k-means simplemente se agrupa como una persona de nivel socioeconómico bajo, lo cual provocaría un tipo de análisis innecesario (de ser el caso que se esté analizando personas con necesidades económicas).

Para el elemento 178, aunque sea una persona de educación superior como el título lo indica el individuo reciba un salario medio bajo el cual necesita para mantener a su familia, por lo que el algoritmo difuso indica una partición considerablemente alta para el nivel socioeconómico medio bajo y bajo. Por otra parte, en el análisis binario se ignoran estos hechos y se clasifica como una persona de alto nivel económico.

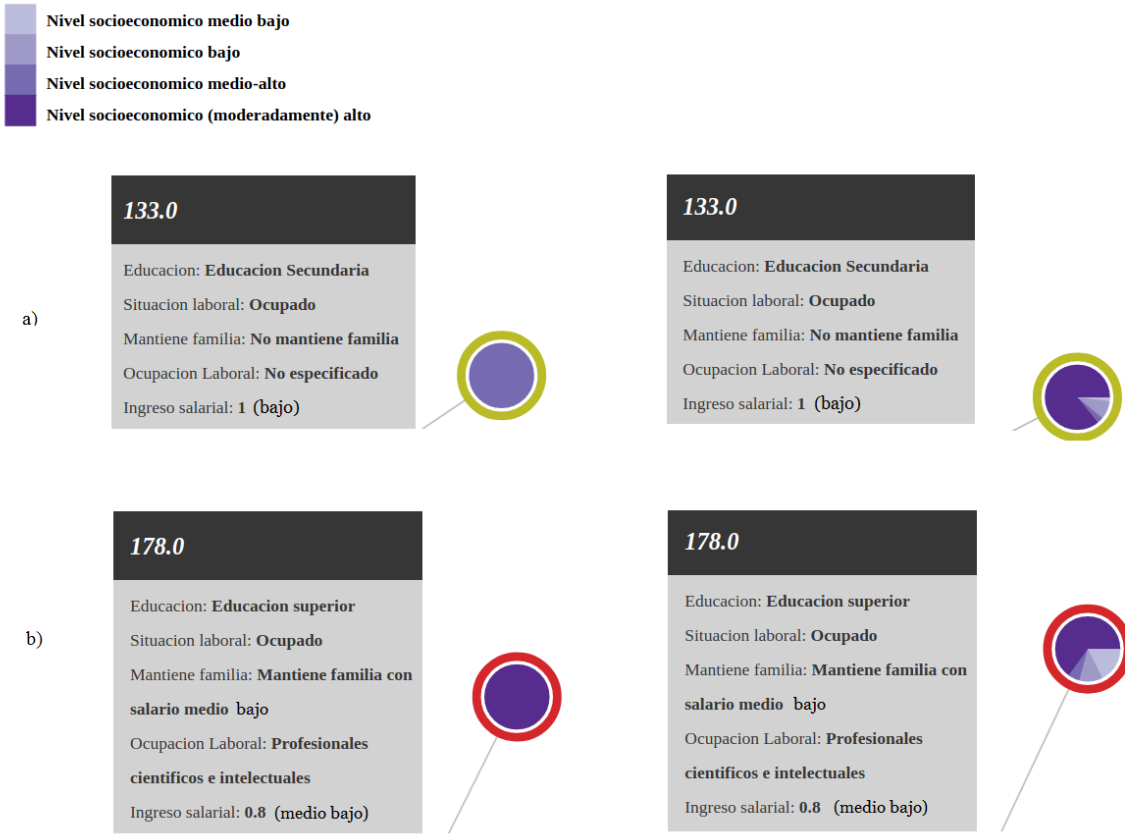

Figura 23 Comparación de análisis de agrupación utilizando algoritmo k-means con respecto al algoritmo de fuzzy c-means

Una vez definido el tipo de visualización se procede a validar la misma con las agrupaciones definidas por el clúster difuso y analizar los patrones que se pueden observar, así como información interesante que no se podría haber encontrado en un análisis de agrupación *hard* o binaria como se analizó anteriormente.

Capítulo 5. Validación de la propuesta

Caso 1: Migraciones en Costa Rica según catálogo de datos de la INEC

Se analizan los resultados comenzando desde el nivel más alto, las regiones cambian de tamaño de acuerdo con la migración neta (data por la diferencia de cantidad de individuos que inmigran y los que emigran a una región). Si el valor es muy pequeño se mantiene un radio con valor 20, de lo contrario se asigna un radio mayor que denote la migración neta para esa región.

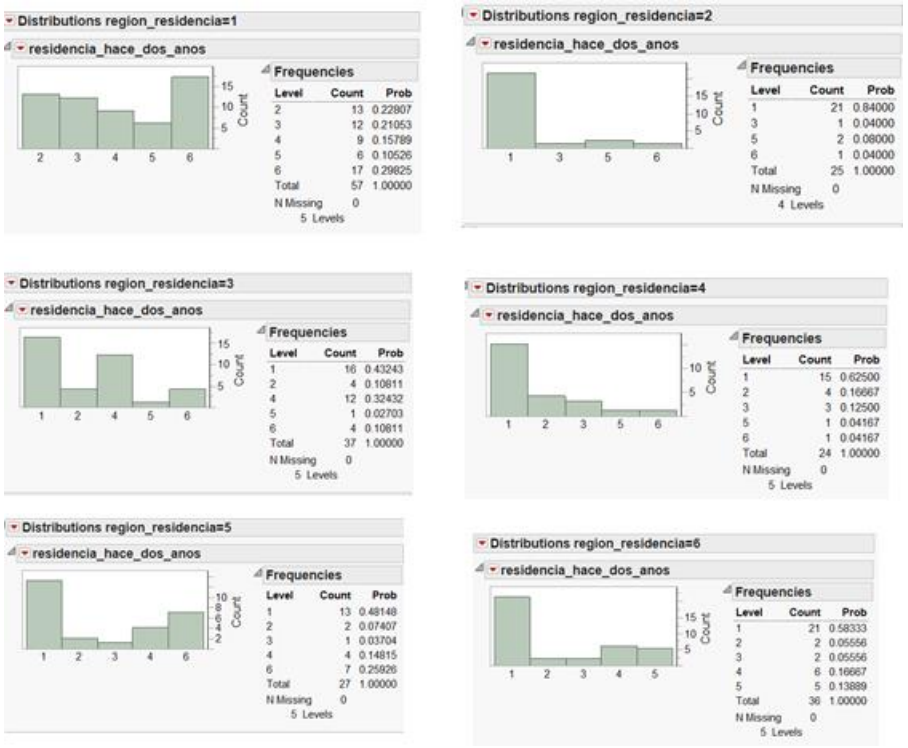

Figura 24. Distribución de migración de región inicial a región destino (2014)

La región central se caracteriza por proveer oportunidades de empleo estable y una buena educación por lo que puede influir en la necesidad de buscar un lugar donde se brinden mejores condiciones de vida. En las distribuciones de la figura 24 se muestran las regiones (donde 1 = Region Central, 2 = Region Chorotega, 3 = Region Pacifico Central, 4 = Region Brunca,5 = Region Huetar Caribe y 6 = Region Huetar Norte) como la Región Central tiene

una alta migración de todas las otras regiones donde la Region Huetar Norte es la que tiene la concentración más alta de emigraciones hacia esta región. Este comportamiento se puede notar también en la visualización de la figura 25 gracias a los arcos que se encuentran alrededor de los nodos hijos. También se puede notar en los enlaces de color rojo (cuando se da click a la "región actual" se muestran los lugares o regiones a los que migraron los individuos de esa región), a partir de la región Pacífico Central donde se migró las otras 5 regiones: Región Chorotega, Región Central, Región Huetar Caribe, Región Brunca y Región Huetar Norte, también de acuerdo con el grosor del enlace se puede notar que la mayor cantidad de personas migraron hacia la Región Central principalmente, con una diferencia bastante alta con respecto a la segunda región con mayor cantidad de personas (Brunca).

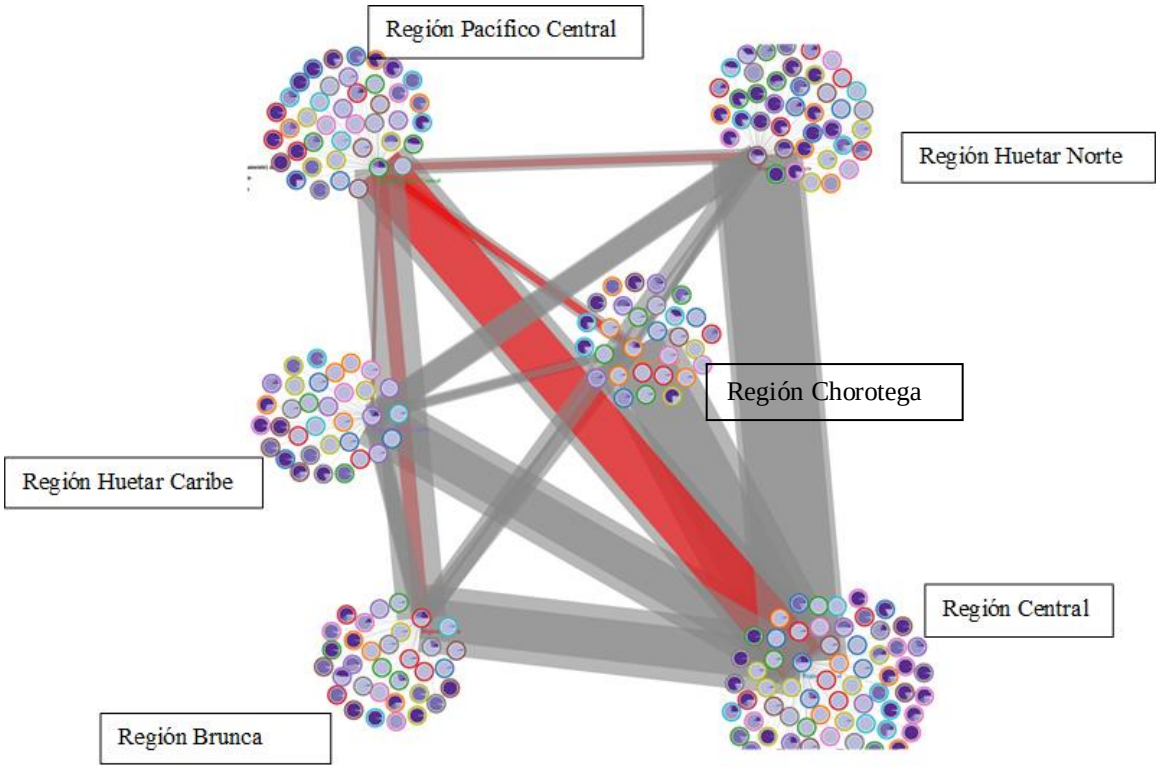

Figura 25 Migraciones a partir de la región Pacífico Central

Para los clústeres se cumplen las siguientes condiciones:

- Clúster - Capacidad socioeconómica baja: Mantiene a la familia, tiene salario mínimo, tiene educación baja (secundaria o menor), tiene baja calidad de vivienda, desempleado agrega más peso para este clúster.

- Clúster - capacidad socioeconómica (moderadamente) alto: Posee educación superior (Bachillerato universitario o mejor), no mantiene a la familia, tiene trabajo, Salario medio alto o alto, calidad de vivienda óptima.
- Clúster - capacidad socioeconómica medio baja: Tiene educación medio baja (no menor a educación secundaria), mantiene a la familia, salario medio bajo o medio, tiene trabajo.
- Clúster - capacidad socioeconómica medio alta: Tiene educación superior o media-técnica, no mantiene a una familia, salario medio o medio bajo, tiene trabajo

Se sabe que no todas las personas van a encajar en los cuatro clústeres. Por lo que se hace necesario aplicar el análisis difuso para esos casos. En la figura 26 se muestran los colores utilizados para representar la pertenencia a cada una de las agrupaciones que se describieron anteriormente. En la figura 27 se pueden observar 4 elementos (individuos) diferentes con los detalles que influyeron a su clasificación. El elemento 130 de la figura 27 muestra un comportamiento tipo k-means donde solo se pertenece a una sola agrupación en este caso a la capacidad socioeconómico-baja, un individuo con escasos recursos que además tiene que hacerse cargo de su familia con un nivel de educación baja, para este caso no se especificó el puesto.

Para los otros elementos se van a notar un comportamiento más difuso, como lo es el caso del elemento 109 se tiene a una persona con rasgos principalmente de una capacidad socioeconómica medio baja (manteniendo a la familia con un salario medio bajo) pero con un porcentaje, aunque bajo de un nivel socioeconómico alto al tener educación superior (nivel universitario). Para el elemento 114 al contrario se tiene un individuo con características predominantemente de una capacidad socioeconómica medio alta (salario bajo, pero no mantiene familia) con características de un nivel alto, en este caso nuevamente un nivel universitario de educación. Y por otra parte tenemos el elemento 106 donde se tienen características de personas con un nivel económico muy bajo, mantiene a la familia y además con salario mínimo, pero tiene educación superior lo cual corresponde a una característica de un nivel socioeconómico alto.

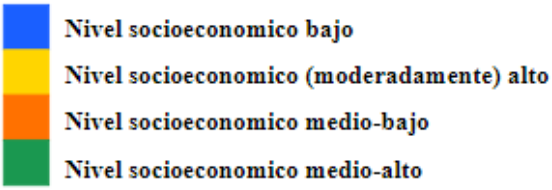

Figura 26 Etiquetas de membresía a clústeres según color

Figura 27. Análisis difuso de elementos estudiados como parte de una migración en Costa Rica

Ahora bien a nivel del tiempo para este caso a través de los años 2014, 2015 y 2016 se puede observar como en la figura 28 para el año 2014 las regiones Brunca y Chorotega tienen una migración neta baja a comparación de la Región Central que fue relativamente alta (con 86 personas inmigrando, 57 emigrando, con una migración neta de 29, para las otras regiones con un número menor a 10 o negativo – más emigrantes que inmigrantes) y con un nivel socioeconómico medio bajo y bajo en su gran mayoría, por otra parte en la figura 28 y 29 se nota cómo aunque predominantemente se encuentre un nivel socioeconómico medio bajo y bajo aún, tiene más individuos con capacidades socioeconómicas medio-altas que en el 2014 donde se tuvo una inmigración mayor, a diferencia de 2015 y 2016 con una migración neta de -14 y -16 respectivamente.

También para el año 2015 en la figura 29 se puede notar como al contrario de la Región Central para las regiones Brunca y Chorotega se tuvo una migración neta más alta en 2015 que en 2014 y volvió a bajar en 2016, por otra parte, no hubo un cambio fuerte a nivel socioeconómico bajo y medio bajo a través de los años.

En el caso de las regiones Pacífico central y Brunca (aplicable para los 3 años, figuras 28, 29 y 30) se encontró que, aunque se encuentran varios individuos con nivel socioeconómico medio-alto, tienen características que van desde tener un salario medio, pero no mantiene familia, o con un salario bajo con educación superior y sin mantener familia. Mucho de esto se puede deber a que a pesar de que existen profesionales calificados, estos no están trabajando en la carrera por la cual estudiaron debido a la región donde viven actualmente.

Por lo que sería interesante revisar esos casos o encontrar más información acerca no solo del trabajo que tienen actualmente. Sino que también la carrera que estudiaron originalmente. También qué tipo de oportunidades de empleo se ofrecen en la zona.

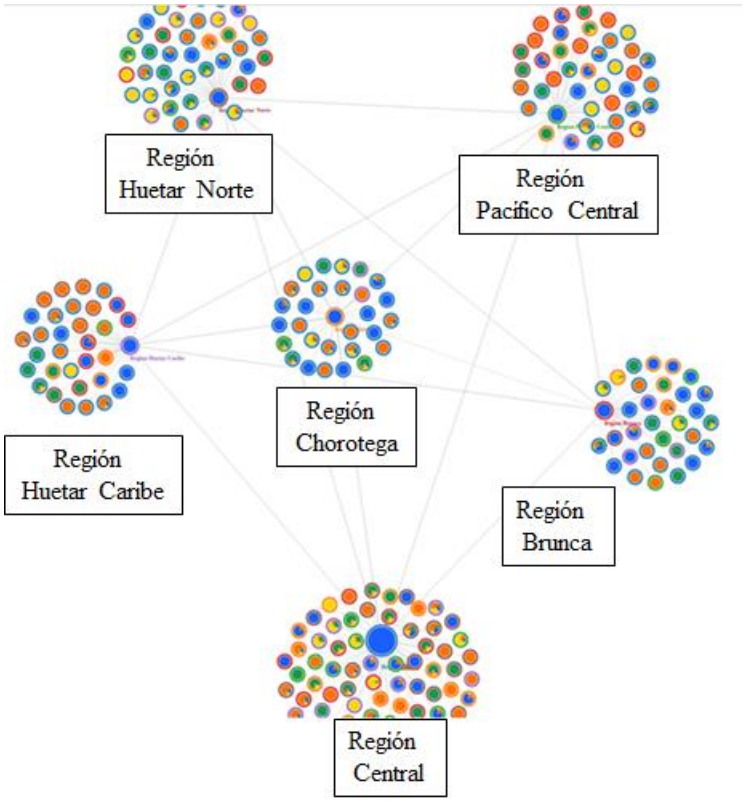

Figura 28 Representación con enfoque difuso de migración interna en Costa Rica 2014

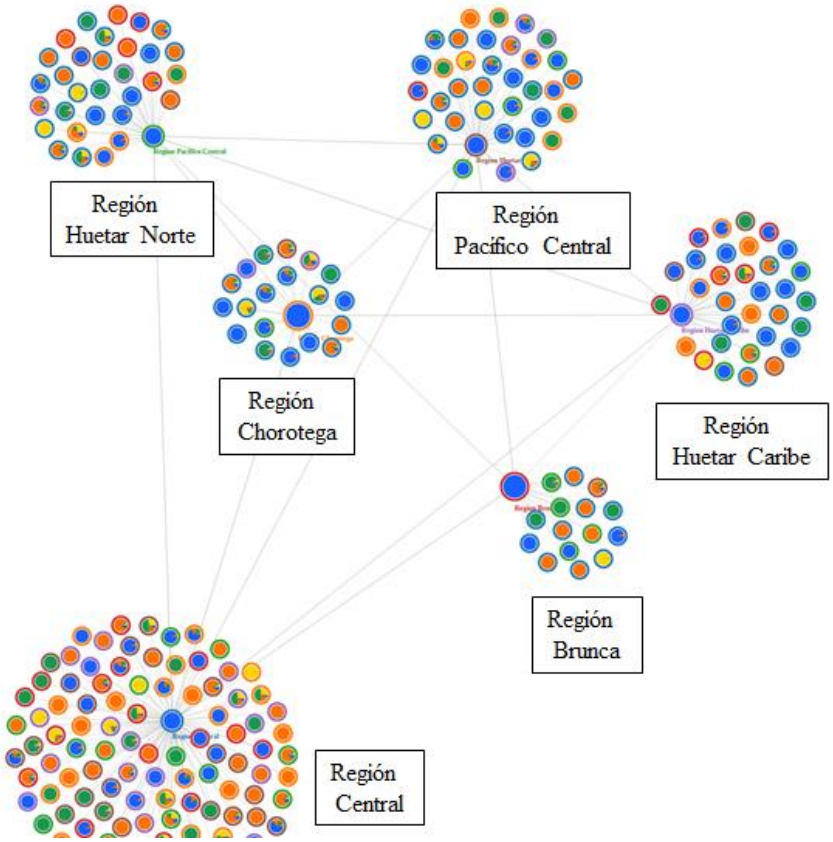

Figura 29 Representación con enfoque difuso de migración interna en Costa Rica 2015

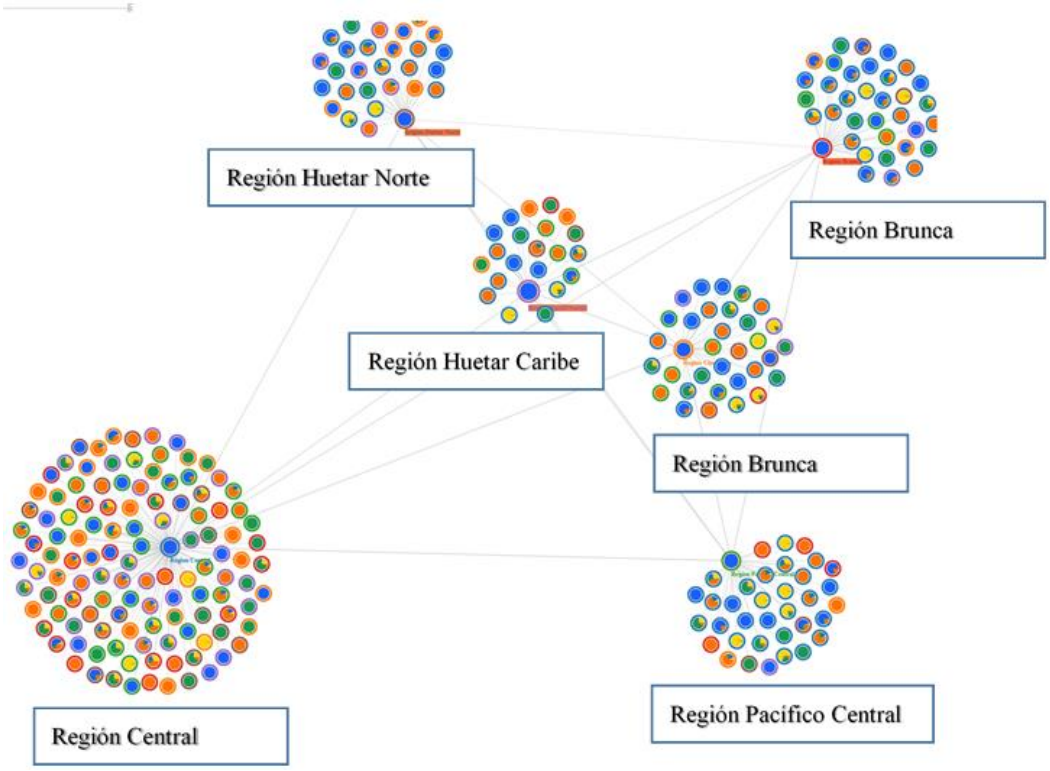

Figura 30 Representación con enfoque difuso de migración interna en Costa Rica 2016

Aunque los resultados se mantienen, se revisó que la paleta de colores podría causar problemas para personas con ceguera de color, por lo cual se hizo cambios para que los colores fueran más amigables para estas personas (Tableau, 2016). Por lo que se cambió a un color con gradientes morados como se puede notar en la figura 31.

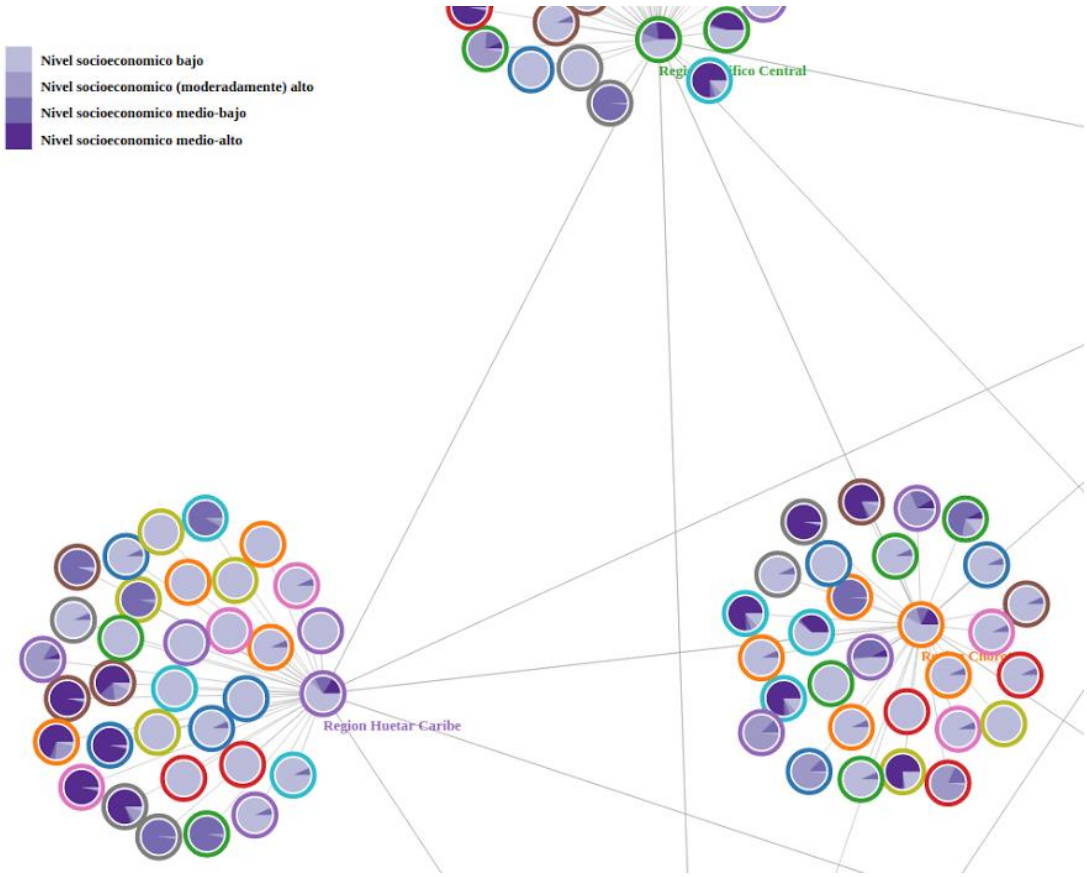

Figura 31 Visualización amigable para personas con problemas de ceguera de color

Otra mejora que se realizó a la visualización también presente en la figura 31 es el hecho de indicar un nodo con la membresía general de las agrupaciones definidas a las que pertenecen los individuos de una región. Donde se puede fácilmente ver cuál la situación que predomina en una región específica, en el caso de la Región Huetar Caribe se puede denotar un muy bajo nivel socioeconómico.

Caso 2: Encuesta sobre uso de Tecnologías en Estados Unidos de acuerdo a datos obtenidos a través del Pew Research Center

El segundo caso se enfoca en el análisis de una encuesta realizada por *Pew Research Center* en los años 2015 y 2016, la cual contiene preguntas sobre la interacción de las personas con el uso de varios dispositivos tecnológicos. (Pew Research Center, March 17-April 12,2015 - Libraries and Technology Use, 2017). Se estarán analizando el comportamiento

de los usuarios a través del tiempo con respecto al uso de estas tecnologías utilizando la metodología propuesta en este trabajo.

Así como se describió para el preprocesamiento de los datos de la migración, las variables se convirtieron a discretas para poder agrupar y analizar los datos de manera adecuada en sus respectivas categorías, por ejemplo, el uso de internet en la casa o el tipo de teléfono que poseen. Por otra parte, también se tienen datos que se refieren a la frecuencia en que utilizan ciertos dispositivos o plataformas tecnológicos, por lo cual se requiere normalizarlos. A través de la biblioteca de pandas se logra al aplicar la fórmula en cada una de las columnas que representan la frecuencia.

```
df_preproc = df_data[freq_array].apply(lambda x: (x - x.min()) / (x.max()
- x.min()))
```

$$Normalizar(e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

Una vez los datos fueron pre-procesados con la muestra seleccionada se utiliza el análisis de componentes principales, la cual enfatiza variabilidad y ayuda a reducir los datos a los componentes más básicos. Esto permite entender mejor la relación entre las variables y a elegir las variables más relevantes (de entre las 41 variables) para análisis de clustering. Utilizando el lenguaje *R* se puede ejecutar el código siguiente para obtener el resultado que se muestra en la figura 32 donde las variables que se encuentran en amarillo-rojo son las que proveen una mayor contribución, de las 41 variables se identificaron 9 variables para el análisis de clustering difuso.

```
df <- survey_data[cols]
res.pca <- PCA(df,  graph = FALSE)
fviz_pca_var(res.pca, col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)
```

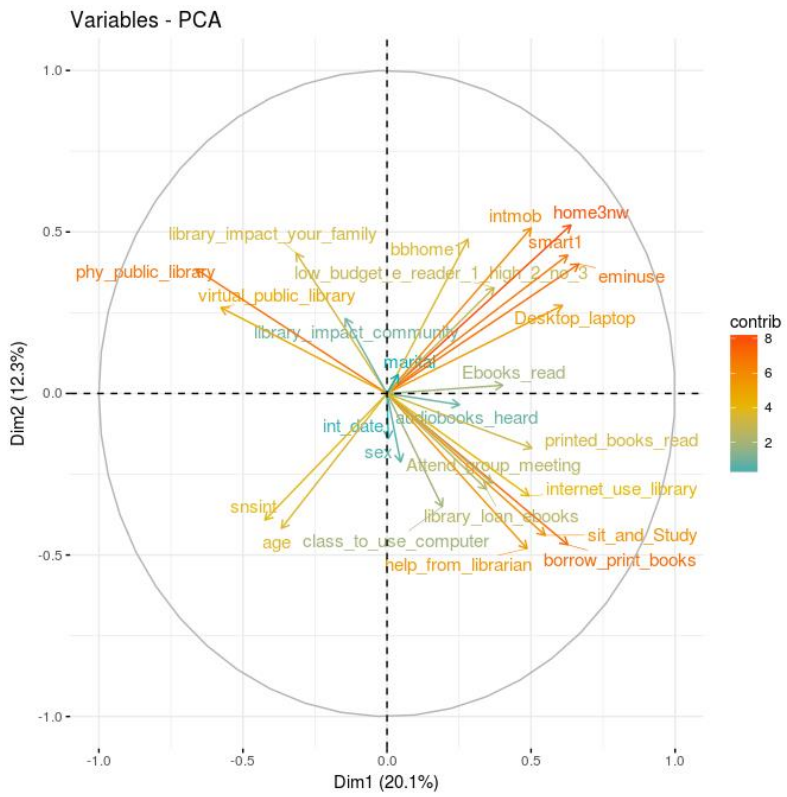

Figura 32 *Análisis de variables utilizando el Análisis de componentes principales*

Como parte de los resultados del análisis se procedió a seleccionar las variables que tuvieran un nivel de contribución al menos mayor a 4 y que también estuviesen relacionados con el tema de uso de la tecnología. Se presentan en la tabla los atributos seleccionados:

Tabla 8 Atributos seleccionados para la prueba de la metodología en datos brindados por (Pew Research Center, March 17-April 12,2015 - Libraries and Technology Use, 2017)

Atributo	**Descripción**
Uso de internet (general)	Indica si la persona utiliza internet o correo electrónico al menos ocasionalmente
Uso de internet en dispositivo móvil	Indica si la persona utiliza internet o correo electrónico al menos ocasionalmente a través de un dispositivo móvil
Uso de internet en el lugar de residencia	Indica si la persona utiliza internet o correo electrónico al menos ocasionalmente en el lugar de residencia
Nivel de conexión a internet	Nivel de conexión de internet en el lugar de residencia: Velocidad alta, baja, no tiene o solo accede a través del teléfono móvil

Tenencia de teléfono inteligente	Si la persona posee un teléfono inteligente
Tenencia de computadora (PC) o laptop	Si la persona posee una computadora (PC) o laptop
Frecuencia en que utiliza las redes sociales	Frecuencia en que utiliza las redes sociales: Facebook, Pinterest, Instagram y Twitter

Una vez seleccionados los atributos se procede a analizar si son óptimos para poder aplicar el algoritmo de clustering utilizando el método de la silueta se revisa que no existan valores negativos que podrían representar un error en el proceso. En las figuras 33 y 34 se muestran los resultados obtenidos del proceso para los años 2015 y 2016 respectivamente.

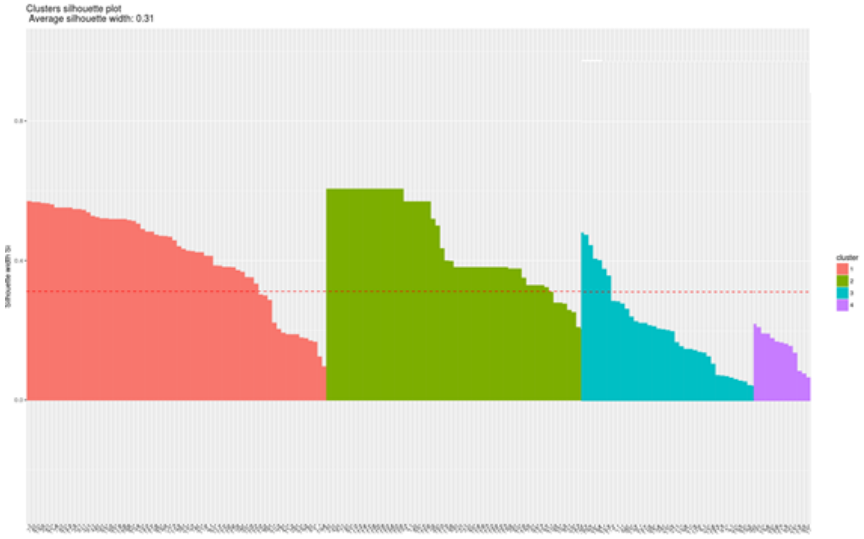

Figura 33 Silueta Clústeres para datos de encuesta por Pew Research Center, Año 2015

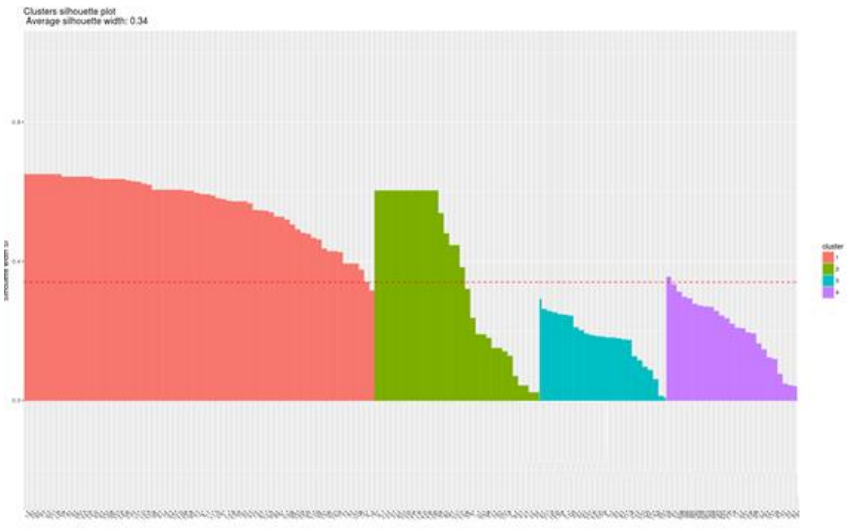

Figura 34 Silueta Clústeres para datos de encuesta por Pew Research Center, Año 2016

Finalmente, una vez analizados y pre-procesados los datos, se procede a crear la visualización para entender el comportamiento de los individuos a través del tiempo y el espacio. En este caso se utilizan los arcos alrededor de los grafos para identificar el grupo de edad al que pertenece cada individuo como se describe en la figura 35. En la figura 36 se muestra la clasificación de los 6 clústeres.

Figura 35 Color de arcos para la visualización de resultados de encuesta obtenida de Pew Research Center

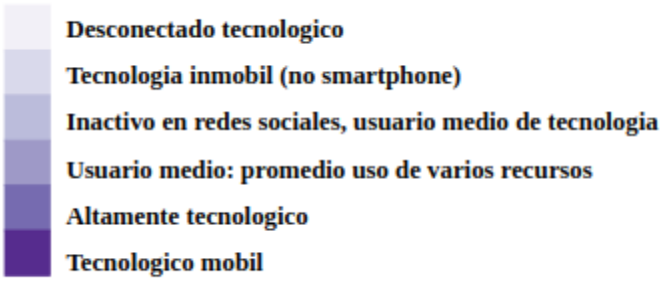

Figura 36 Clasificación de clúster para los resultados de la encuesta obtenida por Pew Research Center

En la figura 37 se puede observar como en el caso de las personas adultas de Arizona se tiene una transición entre el 2015 y 2016 donde pasan de ser personas en su mayoría con una predominancia a estar poco envueltos en lo que es consumo de servicios tecnológicos (redes sociales, internet, teléfonos inteligentes, entre otros) hasta que en el año 2016 se empieza a notar un cambio donde se empiezan a notar personas que utilizan más la tecnología de una manera bastante notable, aunque aún se encuentren casos donde varios adultos aún tienen poca exposición, pero se ve un potencial incremento, igualmente se podría observar los datos más específicos si se quiere encontrar cuales son las plataformas tecnológicas que ahora están utilizando los individuos que realizaron el cambio.

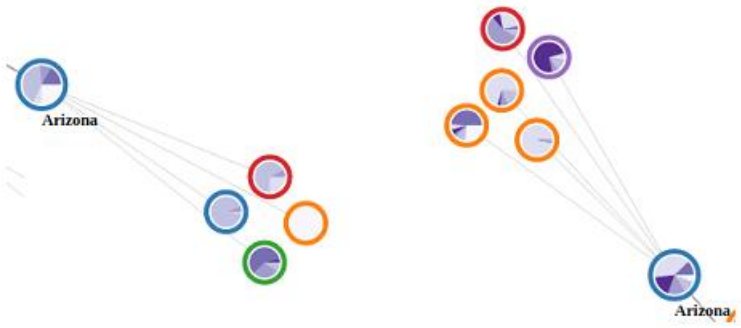

Figura 37 Análisis de Arizona, resultados 2015 y 2016 respectivamente

Se puede analizar además cada uno de los estados para identificar algún nicho donde se podría explorar más formas de realizar publicidad por ejemplo para personas oriundas del lugar de interés (una posible aplicación para segmentación de mercado). Como se puede notar en la figura 38 donde el elemento 141 muestra un joven sin hijos que, aunque tiene una alta exposición tecnológico también tiene ciertos recursos que no utiliza tanto a comparado con otros individuos. Al revisar los datos se puede notar que varios jóvenes con este patrón normalmente no tienen Tablet o desktop/laptop, por lo que, en este caso, empresas que vendan este tipo de productos se podrían beneficiar al empezar a apuntar a este tipo de usuarios.

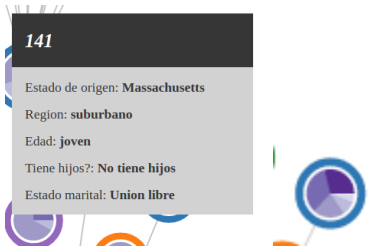

Figura 38 Análisis de Arizona, resultados 2015 y 2016 respectivamente

En las figuras 39 y 40 se muestra una vista general de la visualización para los años 2015 y 2016, en la cual se puede notar que en varios estados para los ciudadanos de oro se ve una transición de personas que tenían una exposición media o baja a la tecnología a tener una adopción considerablemente alta para 2016. En el caso de personas de mediana edad se ve un comportamiento muy variable entre los diferentes estados. También se puede notar que con esta implementación se hace más simple comparar los diferentes estados entre

ambos años, a nivel general por Estado se nota una mejora a nivel de consumo tecnológico de los individuos incluso como se indicó anteriormente para personas mayores de edad.

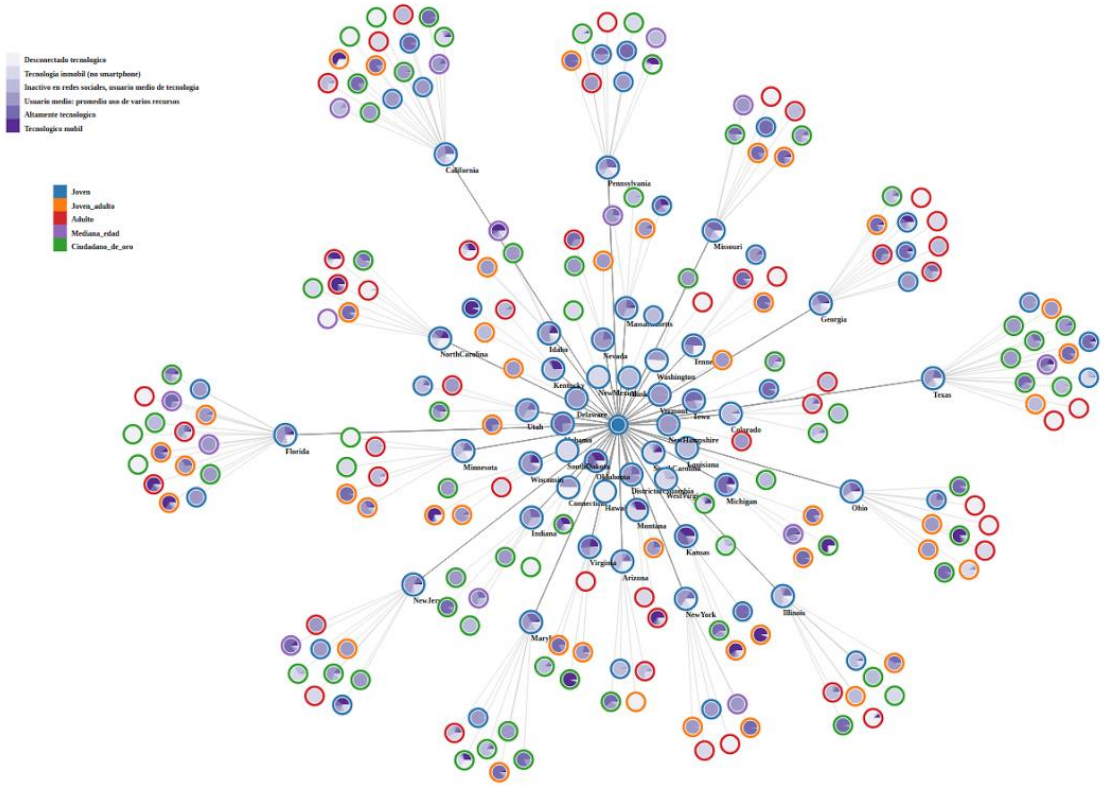

Figura 39 Resultados visualización Estados Unidos - Encuesta Pew Research Center, 2015

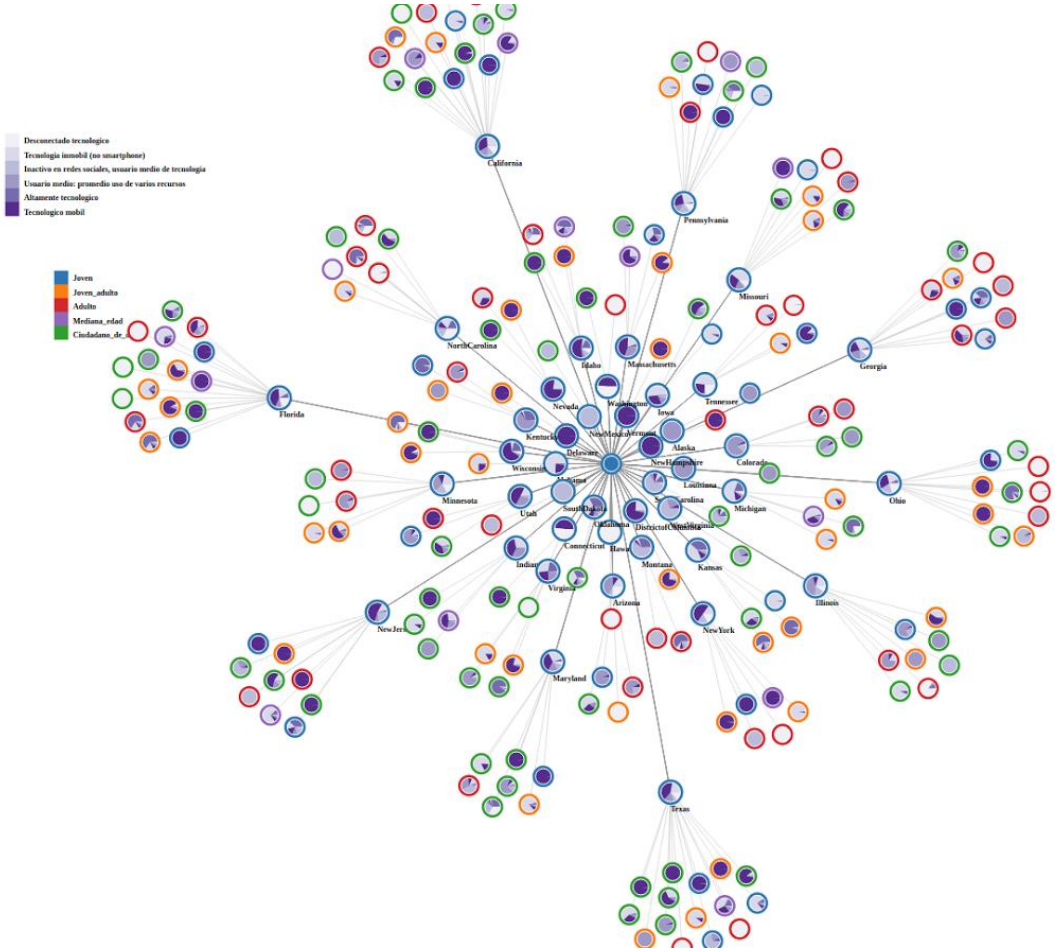

Figura 40 Resultados visualización Estados Unidos - Encuesta Pew Research Center, 2016

Las redes sociales con enfoque difuso tienen un papel fundamental para representar el comportamiento de los seres humanos en diferentes contextos, ya sea a nivel de tecnología, social, económicamente donde sus relaciones se podrían traslapar. La ventaja de la visualización presentada en este trabajo es que toma en cuenta la membresía de cada uno de los nodos. Las visualizaciones que se han visto en otros trabajos no incluyen información la información difusa sobre los nodos, por lo tanto, el código de colores indica el grado de difusidad donde los gráficos circulares se utilizan para investigar la distribución exacta de la membresía.

Además, se incorpora el punto de vista geográfico para poder realizar una comparación de acuerdo con el lugar donde se reside al agrupar los nodos de esta manera, así como su evolución a través del tiempo con el análisis a través de los años.

Capítulo 6 Conclusiones y Recomendaciones

6.1 Conclusiones

La metodología propuesta fue implementada utilizando una serie de librerías de programación y nos permite una visualización de datos que involucra la evolución en el tiempo de datos representados con un enfoque difuso para una red social. Esta red social comprende comportamientos humanos donde al utilizar la agrupación difusa se asegura contar con la representación más adecuada, manteniendo conocimiento valioso a la hora de realizar el análisis, en comparación con una agrupación de tipo binaria en donde no se podría reflejar la realidad de los datos adecuadamente ya que no se puede representar cuándo un elemento pertenece a más de una agrupación.

Se utilizaron los grafos en la red social para expresar las relaciones entre los migrantes y las regiones a las que se movían de una región a otra. Se aprovechó también la naturaleza de los grafos para poder expresar la membresía a la que pertenecían cada uno de los nodos, así como utilizar el algoritmo *force-directed* para aprovechar la fuerza de atracción hacia el nodo padre de acuerdo con el porcentaje de membresía predominante en la visualización, así como los enlaces para denotar la densidad de la población que migra de un punto *A* a un punto *B*.

La paleta de colores utilizada permite contar con un entendimiento más claro de los grupos a los que pertenecen los nodos de acuerdo con su intensidad y además, es una paleta de colores la cual ha sido validada y cumple las recomendaciones de (Tableau, 2016) para personas con ceguera de color.

Esta visualización permitió analizar dinámicamente en una red social el nivel de las personas migrantes en Costa Rica y se lograron encontrar cómo las diferentes situaciones de los individuos van más allá de solo sus propias condiciones socioeconómicas. En efecto, también su situación geográfica puede influir en cómo se agrupan o comparan para poder clasificarlos de manera difusa, lo que permite analizar casos que pasan desapercibidos. Por

ejemplo, el tener una alta educación no asegura necesariamente que se va a obtener un empleo que cubra las necesidades básicas. Son casos que permiten hacer las preguntas correctas para poder solucionar los problemas correctos.

Como un segundo caso para validar la metodología propuesta se utilizó el conjunto de datos para el uso de tecnologías y bibliotecas en Estados Unidos a través de la encuesta realizada por el Pew Research Center la cual brindó información sobre cómo se están comportando los usuarios a nivel del uso de tecnologías, principalmente para diferentes grupos de edad a través del espacio-tiempo dado. Con el análisis difuso se pudo observar con mayor precisión el nivel de exposición a la tecnología donde se pudieron encontrar patrones donde aunque se tenían personas con alto uso de las tecnologías tenían deficiencias en casos de uso específicos los cuales se pudieron analizar gracias a la posibilidad de poder indicar la membresía a las agrupaciones definidas. Esto determina la posibilidad de replicar esta metodología para datos de esta naturaleza donde se analizan comunidades y comportamientos de personas que se requieran analizar a través del tiempo por medio de un enfoque difuso.

6.2 Recomendaciones

Esta implementación es un prototipo que necesita ser refinado y mejorado con otras fuentes de datos que permiten encontrar más patrones interesantes. También es importante resaltar el hecho de que la herramienta puede brindar información más interesante teniendo datos como rutas de usuarios e información específica que refleje el comportamiento de los individuos y así aprovechar el concepto de difusidad.

Integrar esta visualización con alguna herramienta existente permitiría un uso más simple para usuarios en el futuro, además de agregarle características para mejorar la experiencia del usuario como la posibilidad de importar los archivos a visualizar o conectarse a una base datos como MongoDB.

La metodología fue creada de manera que pudiera ser fácilmente reutilizada. Por otra parte, se puede adaptar el algoritmo dependiendo del tipo de agrupación difusa que se quiera realizar.

La visualización propuesta se puede mejorar agregando una mejor transición de la evolución y los cambios de los datos en el tiempo de una manera más fluida y dinámica.

Otro caso en el que se podría aplicar esta metodología sería en analizar con más detalle relaciones de migrantes, por ejemplo, indicar cuándo las personas migran debido a que tienen familiares en algún lugar en específico, esto se podría extender más allá de solo Costa Rica, sino que también a través de las migraciones externas. Asimismo para el caso de la interacción con diferentes recursos tecnológicos se podría encontrar patrones de cómo estos pueden influir en el movimiento de las personas a través del tiempo, por ejemplo el cómo las nuevas tecnologías le permite a las personas acceder a varios servicios en línea lo cual reduce la necesidad de viajar a las instituciones o agencias para realizar trámites, igualmente analizar que tanto ha evolucionado este tipo de interacciones gracias a los avances tecnológicos.

Anexos

La visualización de este trabajo se subió a la plataforma de aplicaciones de Google Cloud por lo cual puede ser accedida a través del enlace: https://fuzzyanalysis-1492273892476.appspot.com

La implementación y los datos utilizados del presente trabajo están incluidos en el repositorio de versionamiento de Github, el cual se puede acceder en este enlace: https://github.com/exolain/Redes-Sociales-con-Enfoque-Difuso

Referencias

Alvarado, G. (10 de 09 de 2009). *Migraciones Internas en Costa Rica: Una Aproximación Regional al problema*. Obtenido de La geografía visión ciudadana: http://lageografiavisionciudadana.blogspot.com/2009/09/migraciones-internas-en-costa-rica-una.html?m=1

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

Ashouri, M. (2016). Graphs Drawing through Fuzzy Clustering. *CoRR*, 1603.07011.

Babuska, R. (s.f.). Recuperado el Mayo de 2016, de Fuzzy clustering lecture: http://homes.di.unimi.it/~valenti/SlideCorsi/Bioinformatica05/Fuzzy-Clustering-lecture-Babuska.pdf

Bezdek, J. C. (1984). FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM. *Computers & Geosciences Vol. 10, No. 2-3*, pp. 191-203.

Cárdenas-Montes, M. (05 de 2017). Obtenido de Ciemat: http://wwwae.ciemat.es/~cardenas/docs/lessons/MedidasdeDistancia.pdf

Dennett, A. (2011). A new area classification for understanding internal migration in Britain. *Population Trends nr 145 Autumn 2011 - London: Office for National Statistics*.

Dey, A. A. (2007). A method to compute distance betweentwo categorical values of same attribute in unsupervised learn-ing for categorical data set. *Pattern Recognition Letters, vol. 28*(no 1), pp. 110–118.

Dwyer, T. (2009). Scalable, Versatile and Simple Constrained Graph Layout. *Eurographics/ IEEE-VGTC Symposium on Visualization 2009*, Volume 28.

Gloor, P. A. (2003). Temporal Visualization and Analysis of Social Networks. *IEEE*.

Gower, J. C. (Dec., 1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics, Vol. 27, No. 4.*, 857-871. Obtenido de A General Coefficient of Similarity and Some of Its Properties.

Guo, G. a.-S. (2014). ETAF: An Extended Trust Antecedents Framework for Trust Prediction. *Proceedings of the 2014 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 540-547.

Hoven, J. v. (05 de 2015). Obtenido de Clustering with optimised weights for Gower's metric: http://www.math.vu.nl/~sbhulai/papers/thesis-vandenhoven.pdf

INEC. (2008). *Catálogo Central de datos INEC*. Obtenido de http://sistemas.inec.cr/pad4/index.php/catalog

INEC. (2011). Resultados Generales. *X Censo Nacional de Población y VI de Vivienda 2011*.

INEC. (s.f.). *INEC Censo 2015*. Obtenido de http://www.inec.go.cr/anda4/index.php/catalog/153/datafile/F1

INSNA. (1999). *What is Social Network Analysis?* Recuperado el 9 de 7 de 2016, de https://www.insna.org/what_is_sna.html

Irke, D. (June de 2014). *Reason for Moving: 2012 to 2013*. Obtenido de U.S. Census Bureau: https://www.census.gov/prod/2014pubs/p20-574.pdf

Itoh, T., Muelder, C., Ma, K.-L., & Sese, J. (2009). A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing. *Visualization Symposium, 2009. PacificVis '09. IEEE Pacific*.

Izakian, H. W. (Vol. 21, No. 5, 2013). Clustering Spatiotemporal Data: An Augmented. *IEEE Transactions on Fuzzy*, pp. 855-868.

Kamada, T., & Kawai, S. (1989). An Algorithm for Drawing General Undirected Graph. *Information Processing Letters*, vol.31, pp. 7-15.

Koffka, K. (1935). *Principles of gestalt psychology*. Lund Humphries, London: Psychology Press.

L.A., Z. (1975). The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I . *Information Sciences 8*, 199-249.

Leung, Y. (August 1992). Visualization of fuzzy scenes and probability fields. *In Proceedings 5th*.

Lin, D. (1998). An information-theoretic definition of similarity. *In ICML '98: Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

Malika Charrad, N. G. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 1-36.

Mi, P., Maoyuan, S., Moeti, M., Yong, C., & North, C. (2016). Interactive Graph Layout of a Million Nodes. *Informatics*, 3,23.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM REVIEW*, 167--256.

NPM. (s.f.). *NPM*. Recuperado el March de 2017, de https://www.npmjs.com/package/csvtojson

Pedrycz, M. L. (2009). The fuzzy C-means algorithm withfuzzy P-mode prototypes for clustering objects having mixedfeatures. *Fuzzy Sets and Systems, vol.160*(24), pp.3590–3600.

Pew Research Center, W. D. (20 de 07 de 2017). *March 17-April 12,2015 - Libraries and Technology Use.* Obtenido de http://pewinternet.org/datasets/

Pew Research Center, W. D. (20 de 07 de 2017). *March 7-April 4,2016 - Libraries and Technology Use.* Obtenido de http://pewinternet.org/datasets/

Shanbhag, P. (2005). Temporal Visualization of Planning Polygons for Efficient Partitioning. *IEEE Symposium on Information Visualization 2005*.

Shyam Boriah, V. C. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. *SIAM International Conference on Data Mining*, 243-254.

Sippel, S. (2016). Domain-specific recommendation based on deep understanding of text. *10.13140/RG.2.1.2958.6809*.

Statistical, N. (04 de 2017). Obtenido de NCSS: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Fuzzy_Clustering.pdf

Tableau. (April de 2016). *5 tips on designing colorblind-friendly visualizations*. Obtenido de Tableau: https://www.tableau.com/about/blog/2016/4/examining-data-viz-rules-dont-use-red-green-together-53463

Vehlow, C., Reinhardt, T., & Weiskopf, D. (2013). Visualizing Fuzzy Overlapping Communities in Networks. *IEEE Transactions on Visualization and Computer Graphics*, volume 19, pp.2486-2495.

Yingdi Guo, K. L. (2015). *A New Spatial Fuzzy C-Means for Spatial Clustering.* Obtenido de http://www.wseas.org/multimedia/journals/computers/2015/a745705-775.pdf