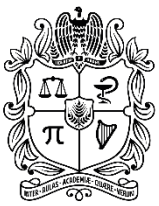


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Evaluación de la calidad de la información geográfica voluntaria mediante un enfoque de análisis multivariado - caso de estudio malla vial Bogotá-Colombia.

Luis Armando Niño Beltrán

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Agronomía
Bogotá, Colombia
2019

Evaluación de la calidad de la información geográfica voluntaria mediante un enfoque de análisis multivariado - caso de estudio malla vial Bogotá-Colombia.

Luis Armando Niño Beltrán

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Geomática

Director:

Ph.D. en Estadística. Aquiles Enrique Darghan.

Codirectora:

M.Sc. en Geomática, Libia Denise Cangrejo.

Línea de Investigación:

Sistemas de Información Geográfica y procesamiento de datos.

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Agronomía

Bogotá, Colombia

2019

A mi madre,

Que siempre me ha ayudado, que siempre se esforzó para que nunca me faltara nada.

Resumen

El aumento en la producción de información geográfica voluntaria (VGI) ha venido creciendo considerablemente y se han realizado diversos estudios al respecto. Sin embargo, el desconocimiento de la calidad de la información generada en forma voluntaria y participativa, plantea retos y cuestionamientos sobre el uso de este tipo de información. Este trabajo tiene como objetivo realizar un análisis de la calidad de la información geográfica voluntaria (VGI) desde una perspectiva multivariada. Para ello se compararán los datos colectados a través de la plataforma Open Street Map (OSM) para el año 2017, respecto a la malla vial de Bogotá-Colombia, proveniente esta de una fuente oficial (Infraestructura de datos Espaciales de Bogotá catastro Distrital IDECA).
En la revisión efectuada para el caso colombiano no se identificaron estudios relacionados con el tema; en consecuencia, se evaluó la calidad VGI de la malla vial de Bogotá bajo un enfoque multivariado, usando las medidas de completitud, exactitud posicional y exactitud temática.

Esta evaluación se realizó por medio de un proceso semiautomático que usa un buffer móvil y el centroide de las vías para realizar las comparaciones correspondientes entre dos fuentes de datos. Los resultados encontrados revelan que el método empleado permitió comparar hasta el 85% de los datos, además se calculó que la malla OSM (Open Street Maps) tiene una completitud del 85.42%, sobre toda el área de Bogotá. Una exactitud posicional de 3.98 m y en general una calidad VGI deficiente, pues el porcentaje de error encontrado fue de aproximadamente 60.3%. Se concluyó que los datos VGI gozan de una completitud aceptable, una exactitud posicional óptima y una exactitud temática deficiente.

Palabras clave: VGI, Calidad, Completitud, exactitud, red vial, Expresiones regulares, Multivariado.

Abstract

The increase in the production of voluntary geographic information (VGI) has been growing considerably and many studies have focused on studying this phenomenon. The problem to use this type of information it's related specifically in the ignorance of its quality and as in Colombia there are still no studies related to the subject. In this paper, the VGI quality of the Bogotá road network was evaluated using completeness, positional accuracy and thematic accuracy. This evaluation was made through a semi-automatic process that uses a mobile buffer and the centroid of the lines to perform the corresponding comparisons between two data sources. The results revealed that the method used could match 85% of the data, and it was calculated that OSM has a completeness of 85.42%, over the entire area of Bogotá. A positional accuracy of 3.97 m and generally, a poor VGI quality : Because the percentage of mistake found was over 60.3%, It was concluded that VGI data enjoy acceptable completeness, optimal positional accuracy and poor thematic accuracy..

Keywords: VGI, Quality, Completeness, Accuracy, Road network, Regular expresión, Multivariated.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

1.1 Contexto

Con la aparición de nuevas y mejoradas formas de adquirir, compartir y actualizar información por medio de plataformas y dispositivos online (WEB 2.0) se ha incrementado la cantidad de contribuidores que pueden crear, almacenar y editar datos geográficos(Hudson-Smith, Batty, Crooks, & Milton, 2009) facilitando la obtención de información a bajo costo y de manera más rápida que los métodos tradicionales, sin embargo, estas ventajas están acompañadas de problemas relacionados con la heterogeneidad de atributivos textuales y espaciales, duplicidad y ausencia de eleméntenos, errores topológicos entre otros, lo cual impacta directamente su calidad (Esmaili, Naseri, & Esmaili, 2013; Longley, Goodchi, Ld, Maguire, & Rhind, 2011). La dinámica relacionada a la recolección y codificación de datos geográficos es conocida bajo el término "*Volunteered Geographic Information*" (VGI) (Janelle, D.G. and Goodchild, 2011), donde los datos son aportados por ciudadanos que actúan como sensores sobre el mundo que los rodea, generando bajo el conocimiento local, la información geográfica (Michael F. Goodchild, 2007).

Uno de los proyectos VGI más exitosos que actualmente se encuentran vigentes es OpenStreetMaps[1] , donde los datos geográficos colectados provienen de múltiples usuarios y fuentes, haciendo que estos se encuentren acompañados de un alto grado de heterogeneidad(Michael F Goodchild, 1992). Para minimizar esta alta variabilidad en cuanto a su calidad, OSM ha creado un modelo de producción publicado en Wikipedia (Haklay & Weber, 2008), con el fin de indicar a los usuarios la forma de agregar y editar

[1] https://www.openstreetmap.org

los datos geográficos y así estandarizar la codificación de estos. Sin embargo, debido al creciente uso de los datos VGI, y a la falta de confiabilidad, muchos investigadores se han enfocado en estudiar su calidad y usabilidad (Cidália C. Fonte, Bastin, See, Foody, & Lupia, 2015; Haklay, 2010; Mooney, Corcoran, & Winstanley, 2010a), con el fin de dar el uso más apropiado para este tipo de datos. Todos estos estudios han permitido ahondar en la clasificación de los datos VGI dada la naturaleza por la cual fueron colectados, si estos vienen dados por usuarios que están enfocados en actividades de edición de mapas (Naturaleza Implícita)Antoniou et al. (2010) como lo es el etiquetado de tributos o bien, datos que tienen como fin describir la localización especifica de un lugar (Taylor, 2014, p. 44).

Diversos autores como Goodchild et al. (2012) han discutido alternativas para evaluar la calidad de los datos VGI y entre ellas se destacan el uso de grupos de usuarios para validar las ediciones realizadas (crowd-sourcing) así como también el uso de documentación para controlar la su calidad. Investigadores como (Elwood, Goodchild, & Sui, 2012; Foody et al., 2013; Jonietz & Zipf, 2016; See et al., 2013) han creado completos marcos de trabajo donde se evalúa y controla la calidad de estos tipos de datos.

Existen una serie de métodos cuantitativos y cualitativos para medir la calidad, entre ellos se tienen las **medidas** e **indicadores** VGI (Antoniou & Skopeliti, 2015; Kresse & Danko, 2011). Los principios y directrices de las medidas de calidad están dados por la International Organization for Estandarization (por sus siglas en inglés, ISO), en su versión más reciente (ISO 19157:2013, 2013) donde se han definido los siguientes elementos de calidad: Completitud (completeness), exactitud posicional (positional accuracy), exactitud temática (thematic accuracy) , entre otros. Sin embargo, estas medidas fueron creadas principalmente para calcular la calidad sobre métodos tradicionales de recolección y codificación de datos geográficos y fueron adoptadas por algunos productores de información VGI para poder medir su calidad. Dada la naturaleza de los datos VGI donde existe una alta heterogeneidad debido a sus diversas maneras de ser producida, autores como (Fonte,2017,p.156) han planteado la necesidad ***de crear nuevas formas de medir la calidad.***

Existen algunas técnicas manuales, semiautomáticas y automáticas que permiten validar algunas de las medidas de calidad VGI, entre ellas se encuentran aquellas relacionadas con búferes de trabajo que seleccionan los datos más próximos entre fuentes (Michael F.

Goodchild & Hunter, 1997) y las automáticas, llamadas Network matching las cuales usan nodos o segmentos "*Node-based algorithm, Segment-based algorithm*" para realizar las respectivas comparaciones(Abdolmajidi, Mansourian, Will, & Harrie, 2015). Sin embargo, aunque algunas de estas técnicas son muy eficientes, aún presentan algunos problemas relacionados a su efectividad y su velocidad(Zhang & Meng, 2008). Por ejemplo, las técnicas manuales consumen una gran cantidad de tiempo y cada vez son menos eficaces debido a la necesidad de entregar a los usuarios en el menor tiempo posible, la información generada por los productores(McKone, Schroeder, & Cua, 2001). Por otro lado, los métodos semi automáticos y automáticos aún presentan problemas en la comparación de geometrías complejas como sistemas de rampas y de manera generar en tiempos de procesamiento(Yang, Zhang, & Lu, 2014), los métodos semi automáticos son ejecutados con un ***buffer estático***, provocando errores de comparación en sectores donde las fuentes de datos varían considerablemente el valor de su posición y quedan expuestos fuera del área de influencia del buffer(Janelle, D.G. and Goodchild, 2011).

Todos estos estudios y técnicas, enfocados a entender la calidad de los datos geográficos codificados por los ciudadanos son realmente escasos en Colombia, dado que aún desconocemos la calidad de los datos VGI. Por otro lado, el incremento de la oferta de datos libres VGI representan una alternativa /oportunidad de uso, donde la integración de los VGI a los datos creados por las agencias nacionales cartográficas serían de gran ayuda para complementar la completitud y actualización de los datos geográficos en el territorio colombiano.

Por ello aquí se evaluó la calidad de los datos VGI de la malla vial de Bogotá colectados a través de la plataforma *OSM* mediante un enfoque multivariado para buscar nuevas formas de medir la calidad de este tipo de datos. A demás se utilizó un algoritmo semiautomático para comparar los datos VGI vs la fuente oficial IDECA. Proponiendo el uso de un buffer móvil a diferencia del buffer estático trabajado tradicionalmente por los métodos semi automáticos.

A partir de este trabajo se encontró que el método semi automático empleado permitió comparar hasta el 85% de los datos. Además, se calculó que la malla OSM tiene una completitud del 85.42%, sobre toda el área de Bogotá, una exactitud posicional de 3.97 m dando como resultado general que la calidad VGI es deficiente respecto a la fuente IDECA, pues el porcentaje de error encontrado fue de aproximadamente 60.3%. Se concluyó que

los datos VGI gozan de una completitud aceptable, una exactitud posicional óptima y una exactitud temática deficiente.

1.2 Antecedentes

Bajo el uso de las medidas expuestas anteriormente se han desarrollado varios trabajos enfocados en medir la calidad VGI. Esta calidad ha sido evaluada comparando objetos referentes a mallas viales respecto a datos oficiales (Dorn, Törnros, & Zipf, 2015; Girres & Touya, 2010; Haklay, 2010; Mahabir, Stefanidis, Croitoru, Crooks, & Agouris, 2017; Mooney et al., 2010a; Zielstra & Zipf, 2010) donde la mayoría de autores han resaltado la presencia de heterogeneidad en la calidad. Estas metodologías se fundamentan en que los datos oficiales se crean bajo altos estándares de calidad(Antoniou & Skopeliti, 2015) y por ello, tiene sentido pretender usarlos como elementos de referencia.

Existen varios trabajos que han usado las medidas cuantitativas de calidad de interés en este trabajo: completitud, exactitud posicional y exactitud temática, para determinar la calidad VGI desde una visión univariada. Por ejemplo, la medida de completitud ha sido ampliamente usada por (M. F. Goodchild & Hunter, 1997) donde a partir de dos fuentes de datos lineales, una denominada oficial y otra como fuente VGI, se crean búferes alrededor de la fuente confiable y se seleccionan todos aquellos elementos VGI que se encuentran dentro del rango de influencia del búfer mencionado, dando como resultado un conteo de elementos y Km que permiten determinar la ausencia o exceso de datos en la fuente VGI analizada. Una modificación a este método ha sido empleada por (Nowak Da Costa, 2016a) donde se evaluó la completitud de los datos cuantificando el exceso y defecto de objetos poligonales. Métodos más avanzados han propuesto evaluar la completitud de los datos por medio de un pareo (matching) semiautomático y automático, tal es el caso del trabajo desarrollado por (Abdolmajidi et al., 2015) donde se evaluó la técnica llamada extended node-based, la cual consististe en comparar la geometría por medio de la coincidencia de nodos y la evaluación topológica de los elementos. Este estudio se basó particularmente sobre el matching de estructuras viales complejas.

Finalmente Jackson et al. (2013) evaluaron la completitud en entidades tipo punto, usando un método más robusto que el desarrollado por (Haklay, 2010) y demostrando que una comparación por conteo no es suficiente para describir las diferencias entre dos fuentes de datos. El método de Haklay consiste en crear una malla regular de tamaño 5x5 Km donde se cuentan elementos por cada cuadrante, para finalmente determinar las diferencias en cada uno de ellos.

Respecto a la evaluación de la exactitud posicional, los métodos más comunes para evaluarla consisten en la creación de pareos geométricos para luego calcular la distancia al centroide de la línea, donde finalmente se calcula la exactitud posicional por medio de la raíz del error medio cuadrático RMSE para la componente deseada. Por ejemplo Haklay (2010) identificó que los datos OSM comparados manualmente con una fuente oficial de Reino Unido contenían un error de 8.5 m, Luding & Krause-Traudes (2010) compararon una malla vial alemana de OSM vs un recurso privado haciendo coincidir de forma automática los objetos lineales para finalmente calcular la exactitud posicional. Por otra parte (Graser, Straub, & Dragaschnig, n.d.) desarrollaron un algoritmo para evaluar la calidad de las redes viales abordando la precisión posicional.

En (Antoniou, 2011) se describe la medición de la exactitud posicional en Inglaterra, haciendo uso de la distancia entre las intersecciones correspondientes de una red de carreteras, obteniendo un error aproximado de 7.9 metros, y concluyendo que esta exactitud permitía comenzar a pensar que los datos OSM podrían ser usados por algunas agencias en su país.

Respecto a la exactitud temática, algunos investigadores miden el porcentaje de clasificación correcta del atributo tipo de vías (Antoniou & Skopeliti, 2015; Stark, 2010). Otros miden la exactitud usando el índice de kappa univariado (Jokar Arsanjani, Mooney, Zipf, & Schauss, 2015). También existen trabajos enfocados a determinar la clasificación correcta de los atributos usando una matriz de confusión y una serie de cruces espaciales (Codescu, Horsinka, Kutz, Mossakowski, & Rau, 2011) donde se han encontrado diferencias importantes de calidad entre áreas urbanizadas y rurales como también sobre el tipo de entidades estudiadas.

Otros trabajos relevantes en cuanto a la calidad de los datos VGI y que se relacionan con las medidas de calidad de interés se encuentran en el campo de la similitud semántica y la aplicación de técnicas de minera de datos (Ballatore, Bertolotto, & Wilson, 2013; Ipeirotis, Provost, Sheng, & Wang, 2014). Dichas técnicas trabajan respectivamente sobre la descripción y clasificación de las etiquetas en OSM usando análisis de texto, así como también descubriendo patrones bajo métodos no supervisados. Una de las ventajas de usar minería de datos es que funciona sobre un enfoque independiente de las leyes y el conocimiento de la geografía, e independiente de los enfoques sociales o de fuentes múltiples para evaluar la calidad de VGI (Mobasheri et al., 2016).

Los otros métodos para asegurar la calidad son los indicadores cualitativos, estos elementos permiten evaluar la calidad VGI desde una perspectiva diferente a la expuesta en los párrafos anteriores, debido a que en muchas situaciones la comparación con datos oficiales no es posible(Fonte, C. C.; Antoniou, V.; Bastin, L.; Bayas, L.; See, L.; Vatseva, 2017). Estos indicadores, conocidos como indicadores VGI, describen el uso y propósito de los datos siguiendo una línea de tiempo desde su recopilación hasta la compilación y uso final. Algunos indicadores VGI son: confiabilidad, credibilidad, calidad del contenido de texto, vaguedad, conocimiento local, experiencia, reconocimiento, reputación (Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2017). Estos Indicadores no hacen parte de este estudio y solo son mencionados aquí a manera de contexto.

En términos generales y siguiendo los resúmenes proporcionados en *A review of volunteered geographic information quality assessment methods* los enfoques más usados para el análisis de la calidad de datos VGI son: Comparación con datos de referencia, consistencia semántica, Análisis de regresión, métodos heurísticos, clasificación de forma, entre otros. Sin embargo, muchas de las medidas de calidad (completitud, exactitud posicional y exactitud temática, etc.) analizadas bajo estos métodos no han sido exploradas en exceso bajo técnicas multivariadas, lo que hace que existan resultados de calidad VGI que describen la calidad variable por variable y no como una característica analizada como consecuencia de la influencia de dos o más medidas de calidad.

1.3 Planteamiento del problema

Si bien la propuesta de dar crédito a los datos aportados por los usuarios que están en contacto directo con su entorno cuenta con el respaldo de expertos en el área(Michael F. Goodchild, 1996), existen posiciones críticas, las cuales afirman que al provenir estos datos de múltiples usuarios su calidad se ve comprometida, lo que afecta directamente la confiabilidad de los datos(Ballatore & Zipf, 2015), haciendo que muchas veces la información no pueda ser usada pues su calidad no está determinada y cuenta con demasiada ambigüedad(Esmaili et al., 2013; Flanagin & Metzger, 2008). Entonces se parte del desconocimiento de la calidad de los datos VGI. ¿Cuál es la calidad de los datos VGI para la malla vial de Bogotá? Aparentemente y según la revisión bibliográfica realizada

para el desarrollo de esta tesis, no se encontró análisis alguno referente al tema. Este desconocimiento hace que no se puedan usar los datos VGI con confiabilidad y por ello es justificable conocerla. Al desconocer la calidad de datos VGI para la malla vía de Bogotá se pierde la posibilidad de incorporar de manera confiable aquella información que posiblemente cumpla con los requisitos de calidad aceptable para entidades gubernamentales y/o desarrollos empresariales que requieran información geográfica de libre de pago. Ahora bien, las medidas de calidad anteriormente expuestas fueron creadas principalmente para operar en relación a métodos tradicionales de recolección y codificación de información geográfica y fueron adoptadas por algunos productores de información VGI para medir su calidad. Dada la naturaleza de los datos VGI, donde existe una alta heterogeneidad es razonable plantear e investigar ***si una nueva forma de medir la calidad es posible***(Fonte,2017,p.156) por ello se formula la siguiente pregunta ¿Existe un camino diferente que permita evaluar de mejor manera la calidad VGI revelando posibles interrelaciones entre las medidas de calidad que no son evidentes con los métodos tradicionales?. Esto permitiría ampliar la investigación referente a la poca exploración de la calidad VGI bajo enfoques multivariados que actualmente se puede evidenciar en la literatura(Senaratne et al., 2017). Partiendo de estas premisas, el propósito de esta investigación es evaluar la calidad de los datos VGI de la malla vial de Bogotá colectados a través de la plataforma *Open Street maps* (OSM) mediante un enfoque multivariado, considerando las medidas de calidad: completitud, exactitud posicional y exactitud temática debido a que la revisión bibliográfica mostró que son las variables más usadas para determinar calidad VGI(Senaratne et al., 2017).

El enfoque multivariado permite abordar una nueva forma de explorar la calidad VGI, mientras que el método empleado para la comparación semi automática propuso el uso de un buffer móvil a diferencia del buffer estático trabajado tradicionalmente por los métodos semi automáticos.

Para desarrollar este objetivo, este documento se estructuró de la siguiente forma: En el capítulo 2 se presenta el marco de referencia donde se expondrán todos los conceptos teóricos usados para el desarrollo de este documento. En el capítulo 3 se expone el estado de arte. En el capítulo 4 se describen los materiales y métodos. El capítulo 5 muestra los resultados junto con su respectivo análisis, en el capítulo 6 se realizará una breve discusión y finalmente en el capítulo 7 se exponen las conclusiones y recomendaciones.

1.4 Objetivos

.

1.4.1 Objetivo General

Evaluar la calidad de los datos VGI de la malla vial de Bogotá mediante un enfoque multivariado.

1.5 Objetivos específicos

- Analizar y estandarizar las relaciones existentes entre los atributos OSM y IDECA
- Examinar y completar los datos de texto VGI para la estandarización
- Comparar y analizar semi automáticamente la información usando muestras estratificadas para clasificar de manera multivariada su calidad.
- Agrupar y clasificar los resultados mediante técnicas multivariadas.

2. Marco teórico

Este marco teórico se divide en 7 secciones, la primera de ellas explica que es dato e información, en que consiste la Información geográfica voluntaria, como funciona y como se obtiene. El siguiente apartado especifica cuáles son las medidas de calidad usadas en esta tesis para medir la calidad de los datos VGI provenientes de OSM. las dos siguientes secciones mostrarán una breve explicación del modelo entidad relación y un desglose de las expresiones regulares. Tema de suma importancia cuando se aborda la estandarización de datos. El siguiente apartado mostrara algunas técnicas de comparación semi automática de atributos encontrada en la literatura, con el fin de poner al tanto al lector de cómo funciona el proceso de comparación de atributos entre dos fuentes lineales de datos, para este caso en particular las fuentes usadas son las redes viales de OSM y IDECA de la ciudad de Bogotá. La sección siguiente mostrará una breve explicación del muestreo aplicado y finalmente se explicará la técnica usada para el análisis multivariado, en este caso un análisis de correspondencia múltiple ACM y análisis de conglomerados.

2.1 Información geográfica voluntaria VGI

Los datos son un conjunto discreto de valores, los cuales corresponden a elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones. Mientras que la información se define como un conjunto de datos procesados que tienen un contexto y un propósito y por tanto dan solución a algún problema. Por lo tanto siempre se referirá a datos a todos los elementos atributivos que componen la

información de la malla vial, mientras que el conjunto de todos ellos se describen como información(Shock, 2003).

La información geográfica voluntaria VGI. Sé puede describir como un movimiento en el cual los datos geográficos son aportados por múltiples usuarios, los cuales actúan como sensores sobre el mundo que los rodea, generando bajo el conocimiento local, información geográfica de manera voluntaria y gratuita (Michael F. Goodchild, 2007). Esta información es creada, editada y compilada por proyectos como OSM donde el objetivo principal es la creación de datos cartográficos de libre uso(Haklay, 2010). Estos datos (VGI), pueden ser agregados o editados por cualquier individuo, lo que implica un alto grado de heterogeneidad, para ello, OSM sigue un modelo de producción creado en Wikipedia(Haklay & Weber, 2008), con el fin de indicar como agregar y editar los datos geográficos sobre la plataforma. En la actualidad cuentan con más de 2 millones de usuarios registrados y más de 20.000 contribuyentes activos(OSM, 2017).

2.2 Medidas de Calidad

La calidad de los datos es definida como el grado en que un conjunto de datos cumple con unos requisitos o especificaciones de producto establecidos previamente para poder ser usados por los usuarios(NSAI standards, 2013).

La calidad de los datos geográficos ha sido enmarcada por la Organización Internacional de Normas ISO, en donde se definieron *PRINCIPIOS y PROCEDIMIENTOS* para la evaluación de la calidad. Anteriormente estos principios, se encontraban registrados bajo los códigos ISO 113 y ISO 114 respectivamente, ahora han sido actualizados bajo la Norma ISO19157:2013. Donde se han reforzado algunos principios para describir la calidad de los datos geográficos(ISO 19157:2013, 2013), allí se identificaron los siguientes indicadores de calidad

2.2.1 Completeness (Haciendo referencia a la totalidad de los datos):

El concepto de totalidad (Completeness "Tomado de la definición en inglés"), fue definido en (Kresse & Danko, 2011) como la presencia y/o ausencia de objetos, atributos y relaciones representadas en el producto respecto a su especificación técnica y a una fuente de mayor exactitud. Sobre este ítem, existen 2 sub elementos de calidad los cuales son conocidos como comisión y omisión. La comisión hace referencia al exceso de datos, al número de elementos dentro del conjunto de datos que no deberían estar codificados pues no existen en la realidad (Número de elementos excedentes) y a una tasa de exceso de elementos, que es calculada como la relación entre elementos en exceso y el número de elementos que deberían haber estado presentes sobre el conjunto de datos. El sub elemento para medir la calidad llamado omisión, cuenta la cantidad de datos ausentes en un producto de acuerdo a lo establecido en la especificación técnica y a una fuente de datos de mayor exactitud.

2.2.2 Exactitud posicional

La exactitud de la posición evalúa cuán bien se relaciona el valor georreferenciado de un objeto con su respectiva realidad en terreno (van Oort, 2006). Sus subelementos de calidad son exactitud absoluta y relativa. La exactitud absoluta hace referencia a la proximidad entre los valores observados vs sus valores verdaderos, mientras que la exactitud relativa se refiere a la posición de un elemento respecto a los demás elementos, contenidos en el conjunto de datos.
Para el cálculo de las diferencias de posición se usa la siguiente fórmula:

$$e_i = \sqrt{(X_{if} - X_{ir})^2 + (Y_{if} - Y_{ir})^2} \qquad (2.1)$$

Donde e representa el error horizontal en cada punto. X_{if}, Y_{if} Son consideradas las coordenadas de la fuente a corroborar, mientras que Xir , Yir son las coordenadas consideradas como verdaderas.

El error medio cuadrático o RMSE consiste en calcular la raíz cuadrada de las diferencias al cuadrado entre los valores de las coordenadas de los datos y las coordenadas extraídas de la fuente de control.

$$RMSE_r = \frac{1}{n}\sum_{i=1}^{n}(e_i) \qquad (2.2)$$

Según *National Standard for Spatial Data Accuracy (*NSSDA) la exactitud posicional con un nivel de confianza del 95%(FGDC, 1998) se puede calcular con la siguiente expresión:

$$Error\ 95\% = 1.7308 * RMSE_r \quad (2.3)$$

2.2.3 Exactitud temática

Evalúa la exactitud de los atributos cualitativos, así como la clasificación de las características y sus relaciones. El sub elemento de exactitud temática involucrado aquí fue exactitud cualitativa de un atributo. El cual consiste en identificar las diferencias entre los valores cualitativos asignados vs sus valores reales. Un ejemplo ilustrativo son las etiquetas de la malla vial. Para el desarrollo de este trabajo, se contemplará la exactitud temática como una sinergia entre la completitud de los atributos y la exactitud de los mismos, siguiendo la definición realizada por Koukoletsos (2012).

La exactitud temática es calculada con la siguiente expresión:

$$\%\ error = \frac{Ne}{N}.100 \quad (2.4)$$

Donde Ne es el número de errores encontrados y N es el número total de elementos.

2.3 Modelo entidad relación

El modelo entidad relación (E-R) permiten observar gráficamente las relaciones entre entidades de una base de datos (Silberschatz, Korth, & Sudarshan, S. (Instituto Indio de Tecnología, 2002). Las entidades son definidas como objetos que son distinguibles de otros objetos y que además poseen atributos. Por otro lado, el modelo (E-R) tiene restricciones de correspondencia de cardinalidad, esto es, el grado de asociación que tienen las entidades. Estas restricciones son definidas de la siguiente forma: Uno a uno 1:1, uno a muchos 1: * y Muchos a Muchos *: *. A continuación, se muestra un breve resumen de los principales conceptos del modelo entidad relación

- Entidad: Es la representación de un objeto acerca del cual se desea guardar información.
- Atributo: característica de la entidad dada por un conjunto de atributos o propiedades.
- Relación: Una relación es una conexión entre dos entidades.

Tipos de relaciones:

- Relación 1:1 (uno a uno): Se representa mediante una línea que une las dos entidades relacionadas. A cada ocurrencia de la entidad A le corresponde una ocurrencia de la entidad B, y viceversa.

- Relación 1:N (uno a muchos): Se representa mediante una flecha que une las dos entidades relacionadas. A cada ocurrencia de la entidad A le corresponden varias ocurrencias de la entidad B, pero a cada ocurrencia de la entidad B sólo le corresponde una ocurrencia de la entidad A.

- Relación N:M (muchos a muchos): se representa mediante una línea con flechas en sus dos extremos que unen las dos entidades relacionadas. A cada ocurrencia

de la entidad A le corresponden varias ocurrencias de la entidad B, y a cada ocurrencia de la entidad B le corresponden varias ocurrencias de la entidad A.

2.3.1 Llaves

La identificación y relación entre entidades viene dada por llaves primarias y foráneas. Las llaves primarias son atributos que identifican de forma única a una entidad/objeto mientras que las llaves foráneas FK son combinaciones de atributos que permiten relacionar a una entidad con otra(Moss, 2012).

2.4 Expresiones regulares

Las expresiones regulares (Regex) son una secuencia de caracteres que forman un patrón representativo de un conjunto de datos mayor, lo que permite comparar dicho patrón con otro conjunto de datos. Esto a su vez permite encontrar, extraer, validar y reemplazar cadenas de caracteres con determinadas características(Lukacs & Bhadra, 2003). Estos patrones pueden estar compuestos por un conjunto de letras, signos y números (ABC123), como también por metacaracteres. Los metacaracteres son símbolos que representan otros caracteres (*.[]$+{}/´…) y tienen como particularidad principal que no se representan a ellos mismos ya que no tienen significado literal, pues son interpretados de una forma diferente por el lenguaje de programación usado(Edition, 2003). Las expresiones regulares son usadas generalmente para procesamiento de texto, detectando números y letras que cumplen con un patrón definido, estas expresiones constituyen en si un lenguaje conocido como *Regex*. Las Regex pueden incluir patrones de coincidencia literal, patrones de repetición, de ramificación entre otras reglas de reconocimiento de texto.

2.4.1 Conceptos de las Regex

Los conceptos expuestos a continuación fueron extraídos de libro Python 3, capitulo 7:*Working with text*. Para simplificar las citaciones, esta será la única referencia realizada para esta sección:

- Metacarácter: Se conoce como metacaracteres aquellos símbolos que dependiendo del contexto tienen un significado especial dentro de la expresión regular. Algunos de los metacaracteres más usados son: ***?. / + [ ] ( ) { } ^ $ | ***
- Meta caracteres de posicionamiento o anclas: Este tipo de meta caracteres permiten delimitar la búsqueda del patrón dentro de la cadena de caracteres.

Algunos metacaracteres de posicionamiento se muestran en la **Tabla 2.1**:

Tabla 2-1: Metacaracteres de posicionamiento.

Meta caracteres de posición	
Meta caracteres	Descripción
$	Fin de línea
\A	Inicio de texto
^	Inicio de línea
\B	Buscar destino a límite de palabra
\b	Busca límite de palabra
.	Comodín para cualquier carácter
:	Coincidencia con cualquier carácter

- Metacaracteres predefinidos: Los metacaracteres predefinidos son funciones creadas previamente para facilitar el uso de las expresiones regulares, permitiendo abreviar el código de las regex. Algunos metacaracteres definidos se muestran en la **Tabla** 2-2.

Tabla 2-2: Meta caracteres de predefinidos.

Meta caracteres predefinidos	
Meta caracteres	Descripción
\w	Detecta letras, números y guion bajo
\W	Detecta caracteres no alfanuméricos
\d	Detecta un carácter numérico
\D	Detecta un carácter NO numérico
\s	Detecta tabulaciones
\S	Detecta No espacios

- Secuencias de escape: Se utilizan para especificar acciones de tabulación o retroceso dentro de una línea de caracteres, algunos comandos de secuencia de escape son ***\n, \t***, los cuales sirven correspondientemente para pasar a la primera posición de la línea siguiente y para pasar a la siguiente posición de tabulación.
- Clases de caracteres: Sirven para buscar caracteres dentro de un rango de posibles opciones. Estas clases son delimitadas por corchetes **[ ]** y dentro de ellas se encuentra el conjunto de caracteres que queremos convertir en un patrón. Por ejemplo, la clase **[^**Avenida**]** encontrará cualquier carácter que no se encuentre en lo descrito en la clase **[^**Avenida**]**. Si tenemos una cadena de caracteres llamada AvC 56, esta clase traerá como resultado C 56. La clase **[**^0-9**]** no mostrará ningún número.
- Rangos: Un rango es una clase de caracteres abreviada que se escribe agregando el primer carácter del rango un guion y el ultimo carácter del rango. Por ejemplo **[**3-7a-c**]** equivale a escribir **[**34567abc**]**. También existen rangos negados los cuales consisten en listar los caracteres que no deben aparecer dentro de la cadena alfanumérica **[^**3-7a-c**]**.
- Cuantificadores: Los cuantificadores son símbolos que permiten definir cuantas veces puede o no aparecer un patrón o cadena de carácter. Un ejemplo de ello es la expresión **[**Car**?**era**]**. esta expresión permite detectar las cadenas de caracteres

Carrera o Carera y desecha las otras posibles opciones. En la **Tabla** (2-3) se muestran algunos de los cuantificadores más usados.

Tabla 2-3: Metacaracteres cuantificadores.

Meta caracteres Cuantificadores	
Meta caracteres	Descripción
{n,m}	Coincide con n,m ocurrencias donde m es el máximo
{n}	Coincide con n ocurrencias del patrón
{n,}	Coincide por lo menos n veces con el patrón
?	Coincide con 0 o una ocurrencia del patrón
+	Coincide con 1 o más ocurrencias del patrón
*	Coincide con 0 o más ocurrencias

2.4.2 Funciones de la librería Re de Python

Compilación del patrón

Las compilación de patrones permiten optimizar el consumo de recursos, por lo que la literatura aconseja crear patrones compilados para grandes cantidades de datos, esta función es definida en Python como **re.compile(Patrón, 'm')** donde m hace referencia a la cadena de caracteres que se quiere compilar.

Metacaracteres de memoria

Los Metacaracteres de memoria se caracterizan por buscar elementos repetidos dentro de las cadenas de caracteres, el código es identificado con de la siguiente forma: **\1**.

Modificación de texto

Por otro lado se tienen los modificadores de texto, estos dividen el texto en una lista, realizando divisiones en donde se logra reconocer el patrón, esta función es conocida

como **split()**. La función **sub()** encuentra todos los subtextos donde existe coincidencia con alguna expresión regular. Esta función es usada para reemplazar caracteres. Finalmente se tiene la función **subn()**: Esta función opera de manera similar a sub() solo varia en que además de regresar el nuevo texto, también informa el número de reemplazos realizados.

Búsqueda:

Los métodos de búsqueda permiten localizar el patrón en la cadena de caracteres, este método retorna un objeto tipo MatchObject el cual contendrá los datos de la búsqueda. Algunas funciones de búsqueda son: **match(,)**, esta función escanea la expresión regular si tiene coincidencias al principio de la cadena de caracteres. **search(,)** busca el patrón en cualquier ubicación de la cadena de caracteres. **findall(,)** devuelve un arreglo con todas las coincidencias encontradas. Una desventaja de esta función consiste en que las coincidencias se devuelven en orden además los caracteres que hagan parte de una coincidencia no puede ser parte de otra. **full.match(,)** devuelve un objeto de coincidencia sí toda la cadena de caracteres coincide con la búsqueda realizada.

Coincidencia:

start(): Retorna la posición inicial de la coincidencia.
end(): Retorna la posición final de la coincidencia.
group(): Retorna el texto que coincide con la expresión regular.
span(): Regresa una tupla con la posición inicial y final de la coincidencia.

2.5 Muestreo simple por asignación proporcional a la localidad

El muestreo estratificado es un diseño de muestreo probabilístico en el que se divide el conjunto de datos en subgrupos o estratos. La estratificación puede basarse en una amplia variedad de atributos o características de la población, como por ejemplo el tamaño, la localización geográfica, la edad, entre otros(Yoshida, 2013). Considerando una población

heterogénea con N unidades, en la que subdividimos L subpoblaciones denominados estratos lo más homogéneos posibles y sin solapamiento(Otzen & Manterola, 2017). En cada uno de los estrados se realiza un muestreo aleatorio simples de tamaño n_i; para finalmente definir cuantos elementos de la muestra se han de seleccionar en cada uno de los estratos. Para crear muestras aleatorias por estrato se dispone de las siguientes opciones:

- Asignación proporcional (el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato que le dio origen, respecto a la población total)
- Asignación óptima (el tamaño de la muestra de cada estrato, es definido por quien hace el muestreo) (Bai et al., 2013)
- Asignación uniforme (Cuando se define el mismo tamaño de muestra para los estratos)

A continuación, se expone la fórmula de tamaño de muestra para realizar un muestreo aleatorio simple MAS.

$$MAS = \frac{NZ^2p(1-p)}{e^2(N-1) + Z^2p(1-p)} \qquad (2.11)$$

Donde Z es la desviación del valor medio que se acepta para lograr el nivel de confianza deseado, aquí usaremos un $Z = 1.645$ que corresponde a un nivel de confianza del 90%. p es la proporción que se busca en el total de la población, e es el margen de error aceptado y N es el tamaño de la población.

2.6 Análisis Multivariado (Análisis de Correspondencia Múltiple ACM)

Cuando se quiere describir y reducir una estructura de dependencia entre variables categóricas, existen técnicas como el análisis de correspondencia que permiten comprender las posibles relaciones existente entre las variables, como también el análisis de su estructura de asociación, describiendo proximidades que permiten la identificación de categorías lo cual produce una síntesis de la información inicial(Montenegro Alvaro, 2005). Esta síntesis se hace en torno a pequeñas dimensiones, permitiendo así la asociación de variables por medio de su similitud(Johnson & Wichern, 1998). Entonces, el análisis de correspondencia es una técnica que nos permite representar atributos de dos o más variables cualitativas en un espacio de pequeñas dimensiones(Jaume & Catalá, 2001), esto involucra que dicha representación se hará agrupando las categorías en función de las semejanzas que presentan las variables relacionadas. Para poder realizar esto, el método se basa en la distancia Ji-cuadrado χ^2, la cual permite ver similitudes entre las variables, la inercia total del sistema que permite determinar la dispersión de los datos respecto a un punto y los valores propios que permiten calcular los vectores los cuales a su vez conformaran la proyección de la nube de puntos dentro de un sistema ortogonal de forma tal que esta explique la mayor parte de la asociación total de las variables.

La recodificación de las variables iniciales, da espacio a la creación de variables factoriales, los factores son una combinación de variables y categorías, las cuales se mezclan permitiendo buscar aquellas combinaciones que permita explicar la mayor parte de las variables en únicamente dos factores, los cuales componen los ejes ortogonales que a su vez permitirán observar las relaciones de todas las variables en un mismo espacio coordenado. A continuación, se muestra gráficamente la composición de las variables factoriales

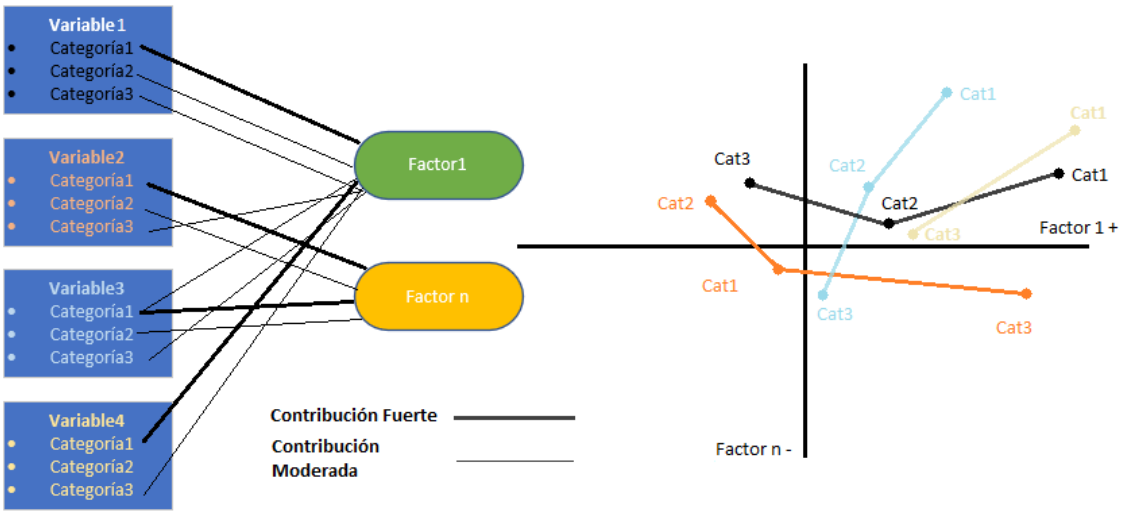


Figura 2-1: Composición de variables factoriales

2.6.1 Análisis de correspondencia múltiple

El análisis de correspondencia múltiple se ocupa de tablas de tres o más entradas, donde se comparan las filas y columnas ocupando todas las posibles combinaciones para determinar las correspondencias que se dan entre las diferentes categorías de cada variable(Montenegro et al., 2005). Para realizar esto, se aplican una serie de cálculos para conocer los perfiles de frecuencia para cada una de las filas y columnas. Se parte de una matriz de datos R, donde las entradas de esta matriz corresponden a los códigos asociados a cada modalidad por variable(Beh, 2012). Esta matriz no es tratable, pues la suma de filas o columnas no posee ningún sentido, por ello es necesario una recodificación, esta recodificación se logra cruzando los n individuos con las modalidades de cada variable(Lozares et al., 1991). A través del uso variables indicadoras se convierte una tabla múltiple en una tabla de doble entrada. Esta tabla posee K variables, donde cada una de las cuales tiene Pi modalidades o categorías (para $i = 1 \ldots K$). Se asocia una variable indicadora por categoría dentro de cada variable de la tabla. El total de categorías se puede calcular bajo la siguiente expresión:

$$\sum_{i=1}^{k} \sum_{j=1}^{p} p_i \quad (2.5)$$

Para un individuo en particular se codifica con 1 si el individuo posee el atributo de la respectiva categoría y con 0 en las demás modalidades de la misma variable ya que se asume que las modalidades son excluyentes(c & Greenacre, 2007). De lo anterior resulta La siguiente matriz **X** de tamaño $n * p$.

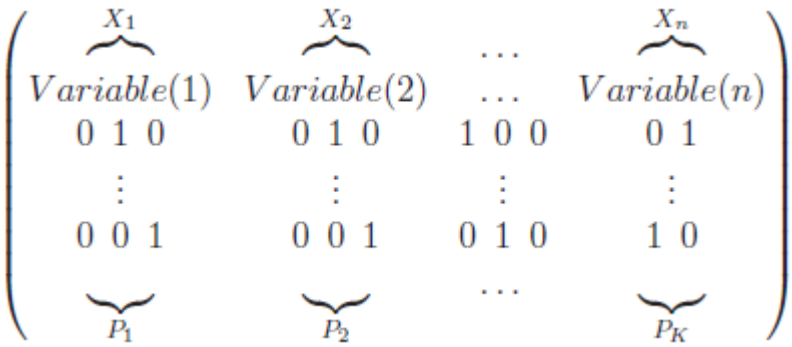


Figura 2-2: Matriz Binaria.

Esta matriz es conocida como tabla disyuntiva o matriz binaria. A partir de esta tabla se calcula la tabla de contingencia de Burt, la cual es desarrollada bajo la siguiente ecuación:

$$B = X'X \quad (2.6)$$

A partir de esta nueva tabla se calculan las frecuencias condicionales el centro de gravedad del sistema, la inercia total y el cálculo de los vectores y valores propios y las distancias Ji-cuadrado χ^2.

2.6.2 Frecuencias condicionales Centro de gravedad e inercias del sistema

Para determinar las distribuciones condicionadas de los valores columna respecto a las filas y filas respecto a las columnas se requiere la matriz de frecuencias condicionales llamada Perfiles. Los perfiles columna respecto a filas $f_{\langle i|j\rangle}$ y filas respecto a columna $f_{\langle j|i\rangle}$ son respectivamente:

$$f_{(i|j)} = \frac{f_{ij}}{f_{.j}}; f_{(j|i)} = \frac{f_{ji}}{f_i} \ \ para\ i = 1, \dots, n\ y\ j = 1, \dots, p. \qquad (2.7)$$

Donde f_{ij} corresponde a cada uno de los valores dentro de la tabla, mientras que $f_{.j}$. es la masa de cada perfil(Lewis, n.d.)

El perfil fila representa las distribuciones muestrales de la variable j condicionadas a cada categoría de la variable i. Ocurre lo mismo con el perfil columna. Los perfiles fila y columna son considerados puntos en el espacio $\Re_p$ y $\Re_n$ respectivamente(Peña, 2002), en consecuencia, los vectores fila y columna que se generan en cada tabla perfil estarán compuestos por sus respectivas componentes. "*Si se consideran dichos puntos y sus pesos como un sistema de masas en el espacio, se encuentra un punto en torno al cual las masas están en equilibrio, este corresponde a su centro de gravedad"(Montenegro, 2005, p. 33).* La fuerza necesaria aplicada en el centro de gravedad para mantener el sistema de masas se denomina inercia y es una medida de la dispersión de la nube de puntos. Si se tiene un sistema de ejes ortogonales, entonces la inercia de la nube respecto a centro de gravedad se descompone como la suma de las inercias a lo largo de cada uno de los ejes(Montenegro, 2005).

El centro de gravedad de la nube de puntos fila se representa por ς_f , sus coordenadas son las frecuencias marginales, es decir, la distribución marginal de la variable que está en columna, $\varsigma_f = f_1 \dots f_p$. De manera similar en la nube de puntos columna, el centro de gravedad está conformado por las frecuencias marginales $\varsigma_f = f_1 \dots f_n$.Restando el centro de gravedad a todos los vectores de los perfiles fila y columna se obtiene una matriz de

perfiles centrados(Murtagh, 2007). Una vez definidos los perfiles fila y columna centrados, se decide como medir la distancia entre ellos. La manera de medirlos será a través de la distancia chi cuadrado χ^2. La distancia entre dos perfiles fila i e i' esta dada por:

$$d^2(i,'i) = \sum_{j=1}^{J} \frac{1}{f_{i+}} . \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2 \quad (2.8)$$

Donde $\frac{f_{ij}}{f_{i+}}$ son las frecuencias para los perfiles fila.

Si los perfiles son similares, la distancia Ji-cuadrado entre cada uno de ellos y su centroide será pequeña. El cálculo de estas distancias da mayor prioridad a la aparición de modalidades (Categorías) raras, por cuanto estas, por su escasez son más diferenciadoras que las otras(Peña, 2002). Si se llegasen a presentar dos perfiles idénticos, entonces ambos perfiles serán representados por las mismas coordenadas en el espacio $\Re_p$ y por lo tanto la distancia ente este punto y los demás no se verá afectada. Lo mismo sucederá con los perfiles en el espacio coordenado $\Re_n$.

El cálculo de la inercia de la nube de puntos respecto a su centro de gravedad se puede computar usando la siguiente expresión:

$$\sum_{i=1}^{n} \sum_{J=1}^{p} \frac{(f_{ij} - f_{i.} * f_{.j})^2}{f_{i.} * f_{.j}} = \frac{\chi^2}{N} \quad (2.9)$$

Donde χ^2 es el estadístico ji-cuadrado, de la prueba de independencia calculado para la tabla de frecuencias relativas y N es el número total de individuos de la tabla.

El valor obtenido en la ecuación (2.9) $\frac{\chi^2}{N}$ es considerado como la inercia total. Esta métrica corresponde a la suma de las distancias calculadas de cada punto al centro de gravedad(Yanai & Ichikawa, 2006).

Esta inercia se descompone un total de k valores propios (que son la explicación de la varianza explicada), cada uno de los cuales constituye la inercia principal de cada dimensión(Montenegro, 2005). El cálculo de los valores propios se realiza a través de la siguiente ecuación:

$$Det(B * -\lambda I) = 0 \quad (2.10)$$

Donde B* es una operación de matrices que contempla la matriz de los perfiles, la matriz diagonal de las marginales y la transpuesta de la matriz de los perfiles.

Esta información permito construir las proyecciones de los perfiles sobre subespacios de dimensión reducida, escogidos de tal forma que dicha proyección conserve la mayor dispersión posible, es decir, en lugar de comparar filas y columnas, este análisis procede derivando un pequeño número de dimensiones(Almeida et al., 2009), de forma que la primera dimensión primer eje o factor explique la mayor parte de la asociación total entre filas y columnas, asociación medida en términos de la inercia total, la segunda dimensión segundo eje o factor explique la mayor parte de la asociación no explicada por el primer factor, y de la misma forma para el resto de dimensiones. Cada uno de los ejes principales caracteriza las categorías de filas y columnas situándolas como coordenadas en el espacio geométrico(Peña, 2002), es decir: las puntuaciones filas son las coordenadas para cada modalidad de la variable fila en las dimensiones de la tabla B y las puntuaciones columna son las coordenadas para cada modalidad de la variable columna en esas mismas dimensiones.

En resumen, este método busca aquellas variables correlacionadas a todos los grupos de modalidades (categorías). Mostrando que las modalidades con apariciones bajas se encontraran más alejadas del origen que las modalidades de mayor frecuencia, este análisis es una descomposición de la nube de puntos de la inercia total del espacio de individuos fila o del espacio de modalidades columna, en ciertas direcciones ortogonales(ITAM, 2015), de tal forma que cada dirección maximice su inercia explicada.

2.6.3 Análisis de conglomerados

Los conglomerados en estadística hacen referencia la agrupación de elementos tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los mismos(Hair et al., 1999). El análisis de conglomerados es una técnica que permite resolver problemas de clasificación. Su principal objetivo es ordenar variables u objetos en grupos que cumplan con cierto grado de similitud o asociación intergrupal(Gondar, 2000).

Este método permite descubrir asociaciones y estructuras en los datos que no son evidentes en un análisis marginal. Los resultados de un análisis de conglomerados pueden contribuir a la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos, o sugerir modelos estadísticos para describir diferentes tipos de fenómenos(Garcia, Juan; Segovia, 2014). Dentro de la vasta gama de algoritmos de clasificación, aquí se resaltan 2: los jerárquicos y los de partición. El Algoritmo jerárquico es un método que tiene una jerarquía de divisiones del conjunto de elementos en conglomerados. parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas están en un único conglomerado, mientras que un método jerárquico de partición sigue el sentido inverso, parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto(R, n.d.). Para el desarrollo de estos métodos se parte de la distancia entre elementos, esta distancia no negativa y simétrica, representa una medida de la diferencia o igualdad entre dos observaciones.

2.6.4 Medidas de similaridad

Las medidas de similaridad permiten asociar elementos según un criterio que por lo general es la distancia(ITAM, 2015, p. 673), el agrupamiento de variables implica una correlación o medida de asociación, la cual tiene muchas veces implicaciones subjetivas, por lo que el conocimiento de la naturaleza de las variables (binarias, discretas o continuas) es imprescindible a la hora de realizar la mejor elección. Hay que tener presente que las variables se agrupan generalmente sobre la base de los coeficientes de correlación o medidas similares de asociación.

Algunas medidas de similaridad son:

- Distancia Minkowiski

$$d_{i,j} = r\sqrt{\left|x_{ij} - x_{jk}\right|^{r}} \qquad (2.11)$$

Donde $d_{i,j}$ es la distancia entre dos individuos $i, j,$ y r es un valor entero que da una variedad en la medida de la distancia. Si r es igual a 1 la anterior ecuación recibe el nombre de distancia Manhatan, mientras que si r es 2 se transforma en una distancia euclidiana.

- Distancia estadística

$$d_{i,j} = \sqrt{(i-j)' - \boldsymbol{A}(i-j)}\,; \boldsymbol{A} = S^{-1} \quad (2.12)$$

- Métrica de Canberra

$$d_{x,y} = \sum_{i=1}^{p} \frac{|xi - yj|}{(xi + yi)} \qquad (2.13)$$

- Coeficiente de Czekanowski

$$d_{x,y} = 1 - \frac{2\sum_{i=1}^{P} \min(xi - yi)}{\sum_{i=1}^{P}(xi - yi)} \quad (2.14)$$

Sin embargo, como la mayoría de variables aquí no pueden ser representadas por mediciones comunes pues hacen referencia a datos binarios (Ocurrencia o no de un suceso), pues se representó la ausencia o presencia de cierta característica de calidad, entonces se usaron medidas de similaridad que permitan medir la relación de las variables por medio de características comunes.

- Distancia Euclidiana cuadrática

$$\left(x_{ij} - x_{kj}\right)^2 = \begin{cases} 0\ si\ x_{ij} = x_{kj} = 1\ o\ x_{ij} = x_{kj} = 0 \\ 1 si\ x_{ij} \neq x_{kj} \end{cases} \qquad (2.15)$$

Donde a mayor distancia cuadrática, mayor es la cantidad de no coincidencias.

Coeficientes de similaridad para variables binarias:

Table Coeficientes de Similaridad

Coefficient	Rationale
1. $\frac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No 0–0 matches in numerator.
5. $\frac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\frac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\frac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

Figura 2-3: Coeficientes de similaridad para variables binarias.

Fuente: Medidas de distancia y similitud Jhonny R. Demey, pág. 50.

Al final del proceso, cuando se hayan conglomerado todos los grupos se procede a mostrar cómo se fueron uniendo dichos conglomerados, para ello se utilizó los árboles jerárquicos. Los arboles jerárquicos o dendogramas son una representación gráfica del resultado del proceso de agrupamiento en forma de árbol, donde la escala horizontal del dendograma corresponde a la distancia en que se produjeron las uniones en cada caso.

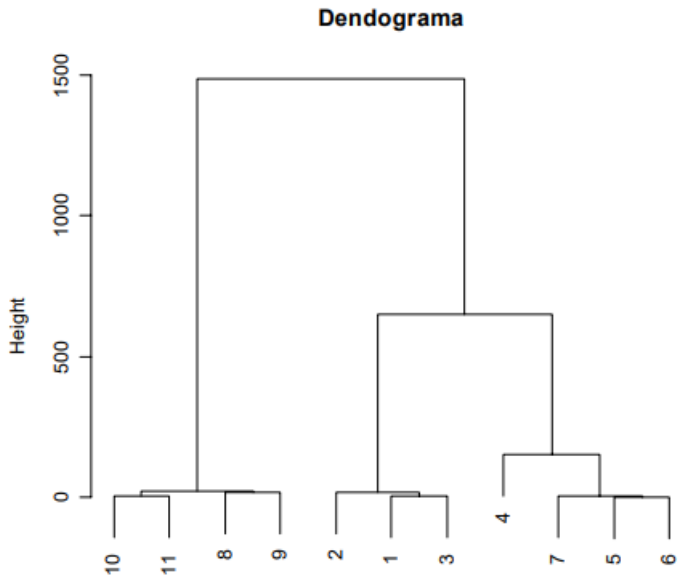


Figura 2-4: Dendograma o diagrama de árbol.

Fuente: https://www.ugr.es/ Departamento de Estadística e Investigación Operativa

2.6.5 Métodos de agrupamiento

Para poder agrupar y diferenciar los clústeres existen una serie de métricas las cuales se exponen brevemente a continuación.

- Método Completo o Vecino más alejado: La distancia entre dos conglomerados es la máxima de las distancias individuales entre puntos del clúster. Tiende a producir grupos alargados. El criterio es invariante ante transformaciones monótonas(Peña, 2002)

- Método Simple o Vecino más cercano: La distancia entre dos conglomerados es la mínima de las distancias individuales entre un punto de un clúster y un punto del otro clúster. Tiende a producir grupos esféricos. El criterio es invariante a transformaciones monótonas.

- Método del centroide: La distancia entre dos conglomerados es la distancia entre sus centroides.

- Método de Ward: Los nuevos conglomerados se crean de tal manera de se minimice la suma de cuadrados total de las distancias dentro de cada clúster. Esto significa que, al aglomerarse dos elementos, la perdida resultante por esta unión sea mínima. Este método también es conocido como método de varianza mínima.(Marín, 2009).

3. Estado del Arte

Bajo el uso de las medidas de calidad VGI se han desarrollado varios trabajos enfocados en asegúrala. Esta calidad ha sido evaluada comparando objetos referentes a mallas viales respecto a datos oficiales (Dorn et al., 2015; Girres & Touya, 2010; Haklay, 2010; Mahabir et al., 2017; Mooney et al., 2010a; Zielstra & Zipf, 2010) donde la mayoría de autores han resaltado la presencia de heterogeneidad en la calidad. Estas metodologías se fundamentan en que los datos oficiales se crean bajo altos estándares de calidad(Antoniou & Skopeliti, 2015) y por ello, tiene sentido pretender usarlos como elementos de referencia. Sin embargo, existen autores que ante este método de comparación subrayan algunos problemas y limitaciones al usar datos de referencia para calcular la calidad de los VGI Mooney et al., (2010). Algunos de los problemas que más resalta Mooney son restricciones de uso, o la incertidumbre en cuanto a la actualización de dichos datos, por ello algunas investigaciones han propuesto métodos que no necesitan datos externos para contratar la calidad VGI(Degrossi, de Albuquerque, Fan, & Zipf, 2016).

Existen varios trabajos que han usado las medidas cuantitativas de calidad de interés en este trabajo: completitud, exactitud posicional y exactitud temática, para determinar la calidad VGI. Por ejemplo, la medida de completitud, el método más usado ha sido el desarrollado por (M. F. Goodchild & Hunter, 1997) donde a partir de dos fuentes de datos lineales, una denominada oficial y otra como fuente VGI, se crean búferes alrededor de la fuente confiable y se seleccionan todos aquellos elementos VGI que se encuentran dentro del rango de influencia del búfer mencionado, dando como resultado un conteo de elementos y *Km* que permiten determinar la ausencia o exceso de elementos en la fuente VGI analizada. Una modificación a este método ha sido empleada por (Nowak Da Costa, 2016a) donde se evaluó la completitud de los datos cuantificando el exceso y defecto de objetos poligonales. Métodos más avanzados han propuesto evaluar la completitud de los datos por medio de un pareo (matching) semiautomático y automático, tal es el caso del trabajo desarrollado por (Abdolmajidi et al., 2015) donde se evaluó la técnica llamada extended node-based, la cual consististe en comparar la geometría por medio de la coincidencia de nodos y la evaluación topológica de los elementos. Este estudio se basó particularmente sobre el matching de estructuras viales complejas. Finalmente Jackson et al. (2013) evaluaron la completitud en entidades tipo punto, usando un método más robusto que el desarrollado por (Haklay, 2010) y demostrando que una comparación por conteo no es suficiente para describir las diferencias entre dos fuentes de datos. El método de Haklay consiste en crear una malla regular de tamaño 5x5 *Km* donde se cuentan elementos por cada cuadrante, para finalmente determinar las diferencias en cada uno de ellos. Con todo ello, estos estudios brindaron conclusiones interesantes en relación a la calidad de estos datos. Por ejemplo, se ha concluido que estos errores no son aleatorios, además presentaban una alta heterogeneidad pues se encuentran mucho más completos en ciudades altamente pobladas, en contraste a lo que ocurre en áreas rurales(Michael F. Goodchild & Glennon, 2010), otra conclusión encontrada en *Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information* indica que el recuento de características simples no evalúan adecuadamente la exactitud y completitud espacial.

Respecto a la evaluación de la exactitud posicional, los métodos más comunes para evaluarla consisten en la creación de pareos geométricos para luego calcular la distancia al centroide de la línea(Michael F. Goodchild & Hunter, 1997), donde finalmente se calcula la exactitud posicional por medio de la raíz del error medio cuadrático *RMSE* para la

componente deseada. los cálculos de desviaciones simples y los métodos de resta de distancias(Fairbairn & Al-Bakri, 2013; ISO 19157:2013, 2013). Por ejemplo Haklay (2010) identificó que los datos OSM comparados manualmente con una fuente oficial de Reino Unido contenían un error de 8.5 m. Luding & Krause-Traudes (2010) compararon una malla vial alemana de OSM vs un recurso privado haciendo coincidir de forma automática los objetos lineales para finalmente calcular la exactitud posicional. Por otra parte (Graser et al., n.d.) desarrollaron un algoritmo para evaluar la calidad de las redes viales abordando la precisión posicional. En (Antoniou, 2011) se describe la medición de la exactitud posicional en Inglaterra, haciendo uso de la distancia entre las intersecciones correspondientes de una red de carreteras, obteniendo un error aproximado de 7.9 metros y concluyendo que esta exactitud permitía comenzar a pensar que los datos OSM podrían ser usados por algunas agencias en su país. Todos estos estudios han permitido concluir que los datos VGI se encuentran correctamente ubicados en algunos sectores, sin embargo, su calidad posicional conserva una estrecha relación entre completitud y la densidad poblacional (Jackson et al.,2013a).

Respecto a la exactitud temática, algunos investigadores miden el porcentaje de clasificación correcta del atributo tipo de vías (Antoniou & Skopeliti, 2015; Stark, 2010). Otros miden la exactitud usando el índice de kappa univariado (Jokar Arsanjani et al., 2015). También existen trabajos enfocados a determinar la clasificación correcta de los atributos usando una matriz de confusión y una serie de cruces espaciales (Codescu et al., 2011) donde se han encontrado diferencias importantes de calidad entre áreas urbanizadas y rurales como también sobre el tipo de entidades estudiadas. Algunos autores que han trabajado la exactitud temática afirman que la mayoría de errores en los datos VGI de OSM son causados por la anotación manual de los contribuyentes que en ocasiones escriben incorrectamente los valores de las características(Mooney & Corcoran, 2012). Otro problema resaltado por la comunidad científica es que las contribuciones que hacen los usuarios en sistemas con pocas restricciones crean condiciones para la heterogeneidad en la información (Nowak Da Costa, 2016b). Muchas veces las comunidades usan etiquetas que tienen sentido para ellos, usan su percepción a cerca de los objetos cartografiarles, pero muchas veces los conceptos alojados en el sistema provienen de otra lengua o incluso son codificados bajo percepciones diferentes sobre un

mismo objeto, lo que implica un problema de clasificación en los objetos(Problemas temáticos) y por ende impactan la calidad(Ballatore & Zipf, 2015).

Por ello algunos trabajos se han enfocado en analizar la calidad desde del campo de la similitud semántica y la aplicación de técnicas de minera de datos (Ballatore et al., 2013; Ipeirotis et al., 2014). Dichas técnicas trabajan respectivamente sobre la descripción y clasificación de las etiquetas en OSM usando análisis de texto, como también descubriendo patrones bajo métodos no supervisados. Una de las ventajas de usar minería de datos es que funciona sobre un enfoque independiente de las leyes y el conocimiento de la geografía, e independiente de los enfoques sociales o de fuentes múltiples para evaluar la calidad de VGI (Mobasheri et al., 2016). Estas investigaciones han ayudado a crear líneas base no solo en el estudio de la semántica de los datos VGI, sino también en estudiar la correspondencia lingüística que existe entre una fuente oficial de datos y VGI, con el fin de poder realizar mejores comparaciones y así llegar a conclusiones más precisas en cuanto a calidad se trata. Por ello, este estudio pretende hacer uso de elementos relacionales de lenguaje para ajustar los datos VGI y hacerlos comparables con la fuente oficial proveniente de IDECA, sin profundizar en todos los aspectos implícitos en los que se puede ahondar tales como la semántica de la plataforma o la relacionada a los procesos de recolección y codificación de la información. Respecto a los procesos de colección de datos, en la búsqueda de entender la calidad de los datos geográficos provenientes de aportes voluntarios, la comunidad científica no solo ha estudiado las propiedades intrínsecas de la calidad descritas al principio de este capítulo, sino también aquellas que de una y otra forma impactan sobre ella(Bordogna, Carrara, Criscuolo, Pepe, & Rampini, 2014). Han sido objeto de estudio las fases de recolección, codificación y edición de datos(C. C. Fonte et al., 2015), es decir, estos estudios se han enfocado en establecer marcos de operación para la adquisición de esa información, con el fin de asegurar la calidad de los datos.

Los otros métodos para asegurar la calidad son los indicadores cualitativos, estos elementos permiten evaluar la calidad VGI desde una perspectiva diferente a la expuesta en los párrafos anteriores, debido a que en muchas situaciones la comparación con datos oficiales no es posible(Fonte, C. C.; Antoniou, V.; Bastin, L.; Bayas, L.; See, L.; Vatseva,

2017). Estos indicadores, conocidos como indicadores VGI, describen el uso y propósito de los datos siguiendo una línea de tiempo desde su recopilación hasta la compilación y uso final. Algunos indicadores VGI son: confiabilidad, credibilidad, calidad del contenido de texto, vaguedad, conocimiento local, experiencia, reconocimiento, reputación (Senaratne et al., 2017). Estos Indicadores no hacen parte de este estudio y solo son mencionados aquí a manera de contexto.

4. Materiales y métodos

4.1 Descripción de la zona de estudio

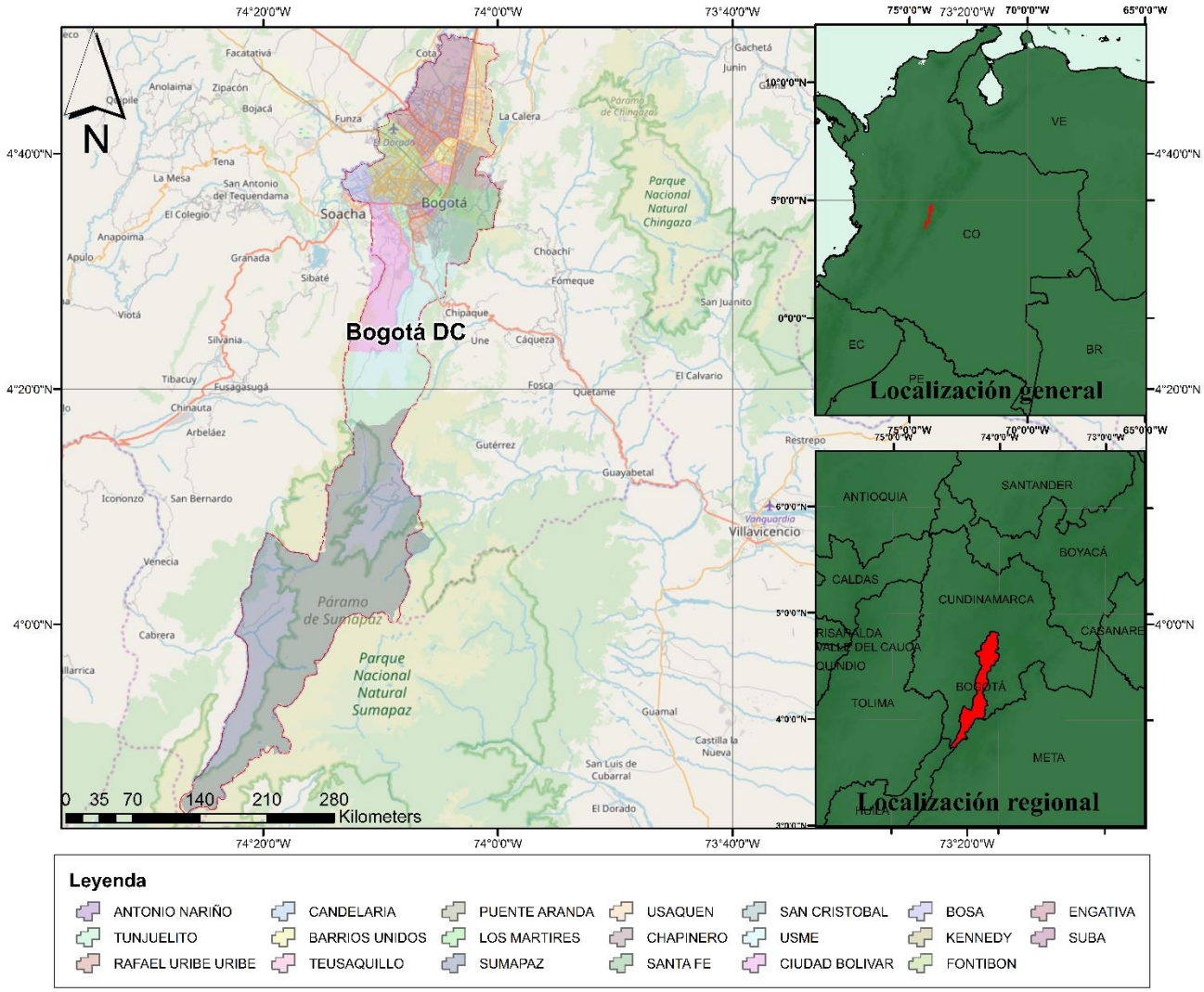


Figura 4-1: Zona de estudio – Bogotá D.C.

La zona de estudio se encuentra localizada sobre la cordillera oriental de los andes a una altitud de 2650 metros sobre el nivel del mar. La ciudad de Bogotá, capital del territorio colombiano posee una extensión de 1732 Km^2 y alberga una población de 7.980.001 habitantes(Triana et al., 2018) lo que la convierte en la más poblada del país. Esta ciudad se encuentra ubicada sobre los 4°36′35″N 74°04′54″W como se puede apreciar en la **Figura** (4-1). Sus límites perimetrales se encuentran inscritos dentro de las coordenadas: Oeste: -74,450; Este: -73,986; Sur: 3,731; Norte: 4,837.

Bogotá está conformada por 20 localidades de las cuales una de ellas es considerada rural. La localidad de Sumapaz se encuentra ubicada en el extremo sur de la ciudad (Ver **Figura** (4-2)), es la menos poblada, pues solo cuenta con 7457 habitantes, pero es la más extensa pues posee una extensión de 307,4 km.

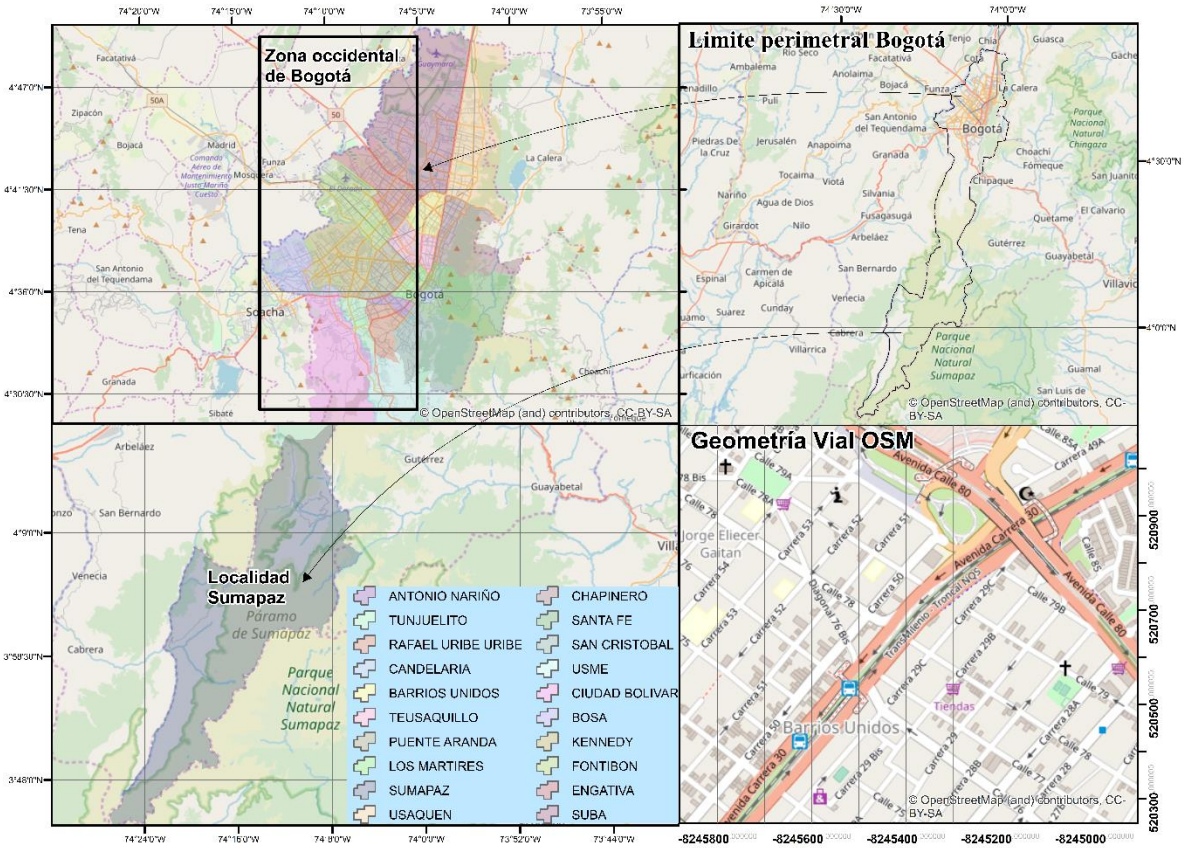


Figura 4-2: Localidades Bogotá.

En contraste la zona más poblada de Bogotá está ubicada en el occidente geográfico. Allí se pueden observar localidades como Suba, Kennedy y Ciudad Bolívar entre otras (Alcaldía Mayor de Bogotá, 2017). Por otro lado, la mayoría de zonas marginales se encuentran ubicadas en el sur occidente y sur oriente de Bogotá, donde los equipamientos brindados por el gobierno son escasos respecto a las otras localidades.

Según la secretaria de planeación distrital, el sistema vial de Bogotá está estructurado bajo la interconexión de cuatro mallas jerarquizadas de acuerdo con sus características funcionales en materia de centralidad(Alcaldía Mayor de bogotá, 2018), la cual se encuentra definida de la siguiente forma:

Malla vial arterial principal: Es la red de vías de mayor jerarquía, que actúa como soporte de la movilidad y la accesibilidad urbana y regional y de conexión con el resto del país.

Malla arterial complementaria: Es la red de vías que articula operacionalmente los subsistemas de la malla arterial principal, facilita la movilidad de mediana y larga distancia como elemento articulador a escala urbana.

Malla vial intermedia: Está constituida por una serie de tramos viales que permean la retícula que conforma las mallas arterial principal y complementaria, sirviendo como alternativa de circulación a éstas. Permite el acceso y la fluidez de la ciudad a escala zonal.

Malla vial local: Está conformada por los tramos viales cuya principal función es la de permitir la accesibilidad a las unidades de vivienda.

Malla vial rural: Vías que comunican los asentamientos humanos entre sí, entre veredas, con la ciudad y la región. En el territorio rural, se definen tres tipos de vías: Principales, Secundarias y corredores de movilidad local rural.

Se estima que la malla vial de Bogotá comprende 8195,90 Km de vías, de las cuales el 1295,6 Km corresponde a tramos tipo malla arterial, 1690,74 Km a tipo malla vial intermedia, 5090,73 a malla vial local, el resto de Km corresponden a malla vial rural sin definir y proyectada lo que significa que estas vías actualmente no existen.

4.2 Datos

Este trabajo contó con el uso de dos fuentes de datos: (i) la malla vial IDECA en su versión 09.17, considerada como fuente de referencia, y (ii) la malla vial VGI proveniente de OSM descargada en septiembre de 2017. Los datos de referencia IDECA fueron levantados bajo las normas de calidad ISO y siguiendo la norma técnica colombiana 5662. Estos datos cuentan con un nivel de detalle 1:2000 y fueron codificados bajo el sistema de referencia espacial MAGNA-SIRGAS (geographic 2D)-4686. Los errores posicionales no superan un metro (1m) de distancia en un nivel de confiabilidad del 95%, mientras que la exactitud temática tiene un porcentaje de clasificación incorrecta del 5% (IDECA, 2017). La malla vial de IDECA consta de 136.958 tramos y 8.195,90 Km. Su sistema de coordenadas geográfico es GCS_MAGNA.

Los atributos a analizar para saber su calidad son descritos a continuación: **Sentido de la vía** (*Desde, Hacia, Doble sentido*). Es una variable indica el sentido de flujo de las vías, *desde y hacia* explican el sentido de flujo desde una sola dirección y doble sentido indica aquellas vías en las que se puede transitar bidireccionalmente; clasificación vial (**Jerarquía via**l), es la clasificación vial de acuerdo a su funcionalidad dentro de la ciudad. Esta se encuentra dividida por IDECA de la siguiente forma: (*Arterial, Intermedia, Local, peatonal, rural, proyectada y sin definir*). Finalmente, la **nomenclatura vial** o Tipo de vía (*AC, AK, CL, DG, KR, TV*) hace referencia al nombramiento de la vía principal. El campo etiquetas contiene la fracción alfanumérica de las vías. En la **Tabla (4-1**) se puede observar un breve resumen de los atributos que componen la malla vial IDECA y que serán objeto de estudio.

Tabla 4-1: Atributos malla vial IDECA.

Atributos malla Vial IDECA			
Objeto geométrico	Atributo	Tipo	Abreviatura
Tramos (LINEA)	-Tipo de vía, -Nombre de la vía	String String	MVITipo MVINombre

	-Etiqueta	String	MVIEtiquet
Tramos (LINEA)	Tipo de clasificación vial	Long	MVITCla
Tramos (LINEA)	Sentido de la vía	String	MVISVia

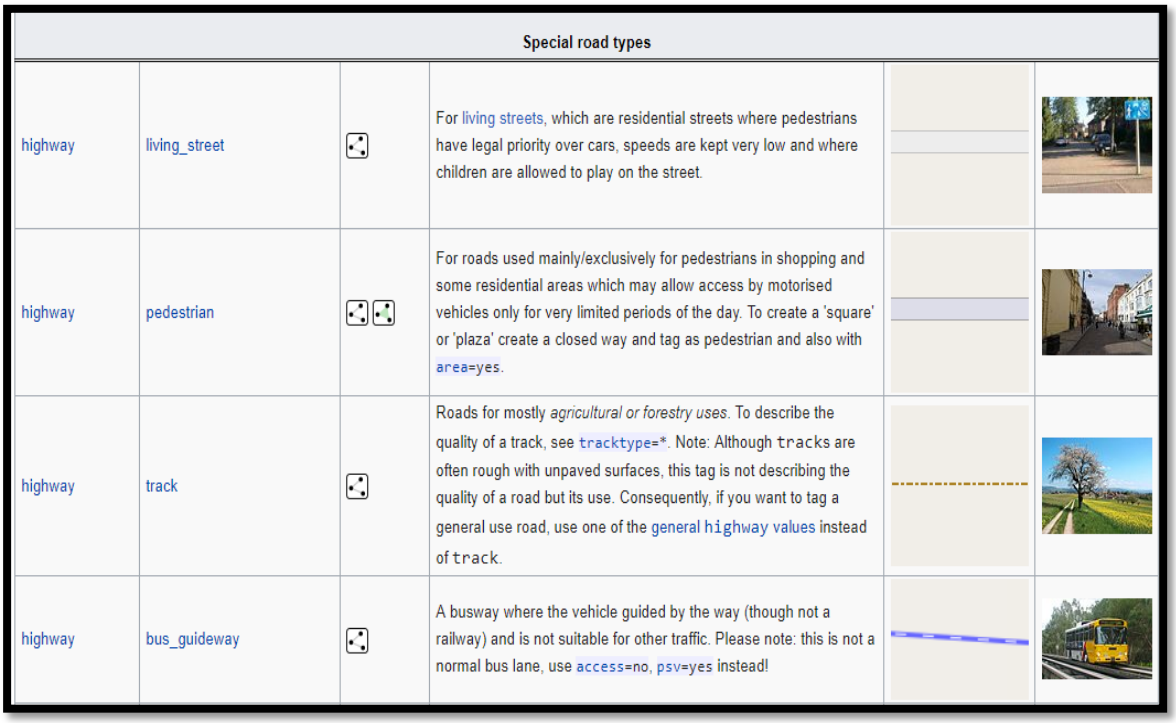
Los datos OSM fueron descargados del sitio geofabrik[2], un servidor que contiene extractos de datos que son actualizados diariamente, dentro de los datos, la capa de geometría vial OSM refiere una serie de atributos muy similares a los expuestos previamente, sin embargo, la clasificación de los atributos llamados (key) en OSM (ver Figura (4-3)) se rigen por los parámetros (values), que son subcategorías para cada elemento dentro de la clasificación dentro del esquema OSM. La información referente a la estructura de la base de datos OSM se encuentra publicada en la página Wiki OSM[3] y fue un insumo de trabajo pues constituían los metadatos para identificar los atributos objeto de análisis.

Key	Value	Element	Comment	Rendering	Photo
Roads					
These are the principal tags for the road network. They range from the most to least important.					
highway	motorway		A restricted access major divided highway, normally with 2 or more running lanes plus emergency hard shoulder. Equivalent to the Freeway, Autobahn, etc..		
highway	trunk		The most important roads in a country's system that aren't motorways. (Need not necessarily be a divided highway.)		
highway	primary		The next most important roads in a country's system. (Often link larger towns.)		

Figura 4-3: Guía de Mapeo OSM.

[2] https://download.geofabrik.de/
[3] https://wiki.openstreetmap.org/wiki/ES:P%C3%A1gina_principal

En la página mencionada se encuentran las guías de mapeo para usuarios. Estas guías explican al usuario como y que debe codificar dentro de la base de datos. Explica la clasificación de los atributos que allí existen, los dominios y sus significados. Por ejemplo, para codificar la Jerarquía vial dentro de *OSM* existen documentos que explican con imágenes que es una vía principal o secundaria, cuál es su categoría correspondiente dentro del sistema y como debe agregar los datos a la base. También explican casos especiales (Ver **Figura** (4-4)), donde se muestran dominios por categorías y breves descripciones de su significado, es así como los usuarios pueden dar clasificación a la malla vial de OSM.



Special road types					
highway	living_street		For living streets, which are residential streets where pedestrians have legal priority over cars, speeds are kept very low and where children are allowed to play on the street.		
highway	pedestrian		For roads used mainly/exclusively for pedestrians in shopping and some residential areas which may allow access by motorised vehicles only for very limited periods of the day. To create a 'square' or 'plaza' create a closed way and tag as pedestrian and also with `area=yes`.		
highway	track		Roads for mostly *agricultural or forestry uses*. To describe the quality of a track, see `tracktype=*`. Note: Although `tracks` are often rough with unpaved surfaces, this tag is not describing the quality of a road but its use. Consequently, if you want to tag a general use road, use one of the general `highway` values instead of `track`.		
highway	bus_guideway		A busway where the vehicle guided by the way (though not a railway) and is not suitable for other traffic. Please note: this is not a normal bus lane, use `access=no`, `psv=yes` instead!		

Figura 4-4: Guía de Mapeo OSM Casos especiales.

La malla vial de OSM contenía en su momento *73.454* tramos y 220.362 nodos, cuenta con el sistema de referencia espacial WGS84. Respecto a los atributos o valores Key tomados para el análisis de la malla vial se encontró que el atributo que contiene las direcciones de flujo vial es llamado **Oneway**. Dentro de este campo se encuentran 3 dominios o valores para el atributo: *both*, *from* y *to*. both corresponden a dirección en doble sentido, los otros explican el sentido de flujo desde una sola dirección. **Fclass**: Este campo contiene la clasificación vial (jerarquía) dentro de OSM, esta clasificación contiene una alta

cantidad de sub categorías (alrededor de 26). Finalmente, el atributo **Name**: El cual contiene las etiquetas del nombre de las vías. En la (Tabla 4-2) se muestra un breve resumen de los atributos OSM usados para el desarrollo de este trabajo

Tabla 4-2: Atributos malla vial OSM.

Atributos malla vial OSM			
Dato	Atributo	Tipo	Abreviatura
Tramos (LINEA)	Name	String	NA
Tramos (LINEA)	Fclass	Long	NA
Tramos (LINEA)	Oneway	String	NA

A demás de estos conjuntos de datos se usaron las localidades de Bogotá, provenientes de IDECA. En la (Tabla 4-3) se muestra un resumen de las fuentes usadas para el desarrollo de este trabajo:

Tabla 4-3: Resumen datos usados.

Elemento	Formato	Descripción	Nota
Datos OSM Malla vial	*.shp	Datos Voluntarios extraídos del servidor de OSM	EPSG:4326 WGS84
Base de Datos IDECA	*.GDB	Datos Oficiales de la ciudad de Bogotá-Malla vial.	EPSG:4686 GCS_MAGNA
Ortofoto Digital de Bogotá 2014	*.wms	Servicio WMS proveniente de IDECA	Resolución espacial 7.5 cm Escala 1:1000. Extensión Geográfica de 49.000 ha
Open Street Map Wiki	*Html	Página Web	Proyectos y lineamientos OSM
Fuente	Open Street Maps 2017		
Fuente	Infra estructura de datos espaciales IDECA 2017		

El procesamiento de los datos fue realizado en el lenguaje Phyton versión 2.7.5 y RStudio 3.4.3. Las salidas graficas fueron creadas en su mayoría en ArcMap 10.4 (Version gratuita para estudiantes). En la **Tabla** (4-4) se pueden observar las librerías y paquetes usados.

Tabla 4-4: Librerías y paquetes usados.

Librerías y paquetes		
Librería/paquete	Descripción	Método
os	Permite usar funciones del sistema operativo	Todos
arcpy	Paquete que permite el uso de funciones GIS.	Auto matching
time	Permite usar funciones relacionadas al tiempo	Auto matching
Urllib2	Permite abrir conexiones URL	Extracción y exploración de datos
math	Permite el uso de funciones matemáticas	Evaluación de la calidad VGI
numpy	Permite generar arreglos multidimensionales e indexación.	Auto matching
pandas	Permite usar herramientas para análisis de datos	Todos
seaborn	Permite el uso de gráficos estadísticos	Extracción y exploración de datos
Re	Permite el uso de expresiones regulares	Estandarización de datos
shapely.geometry	Permite el uso de operaciones espaciales en Python	Estandarización de datos y matching
Nltk	Permite analizar estructuras de lenguaje	Estandarización de datos
Ade4	Permite el análisis de datos Multivariados	ACM
MVA	Métodos gráficos para datos multivariados	ACM
Plugin FactoMineR	Permite desarrollar el Método análisis de correspondencia múltiple por medio de los comandos. MCA	ACM
Stats	Permite el cálculo de distancias para el uso del método Jerárquico	ACM

4.3 Metodología

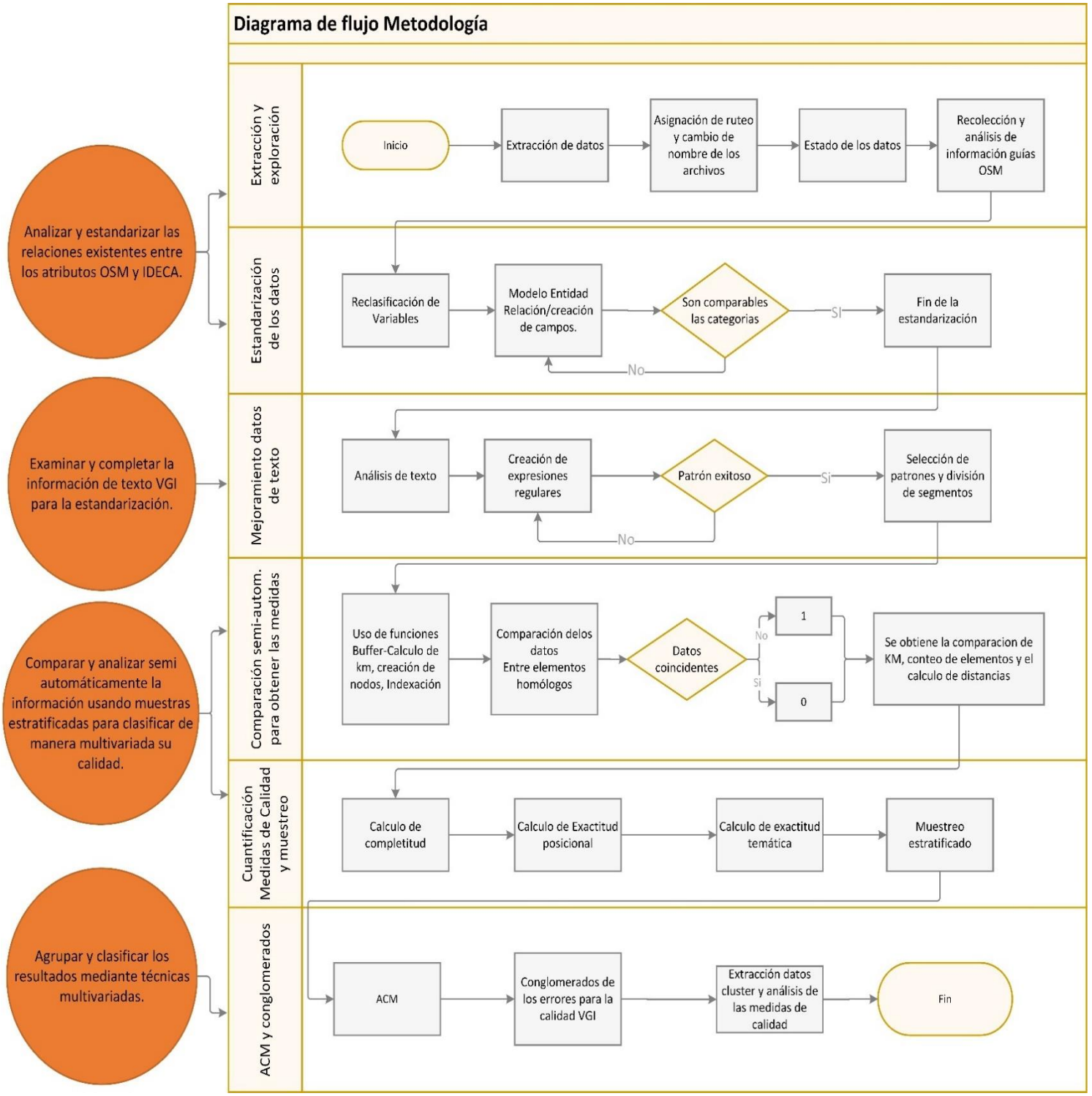


Figura 4-5: Diagrama de flujo de la metodología.

La metodología desarrollada para realizar la evaluación de calidad VGI mediante un enfoque multivariado usando las medidas de totalidad, exactitud posicional y exactitud temática comprendió las siguientes fases divididas por los objetivos específicos propuestos:

Analizar y estandarizar las relaciones existentes entre los atributos OSM y IDECA

- Extracción y exploración de los datos
- Estandarización de los datos por medio de análisis relacional.

Examinar y completar la información de texto VGI para la estandarización

- Mejoramiento de los datos OSM por medio de expresiones regulares.

Comparar y analizar semi automáticamente la información usando muestras estratificadas para clasificar de manera multivariada su calidad.

- Comparación semi automática de los datos para obtener los valores que permitirán evaluar la calidad (tablas que contienen el resultado de la comparación de atributos)
- Evaluación de completitud, exactitud posicional y exactitud temática.
- Muestreo estratificado y reclasificación de las variables

Agrupar y clasificar los resultados mediante técnicas multivariadas.

- aplicación del análisis multivariado
- Agrupación y visualización de resultados por medio de análisis de conglomerados

4.3.1 Análisis y estandarización de las relaciones existentes entre datos y atributos

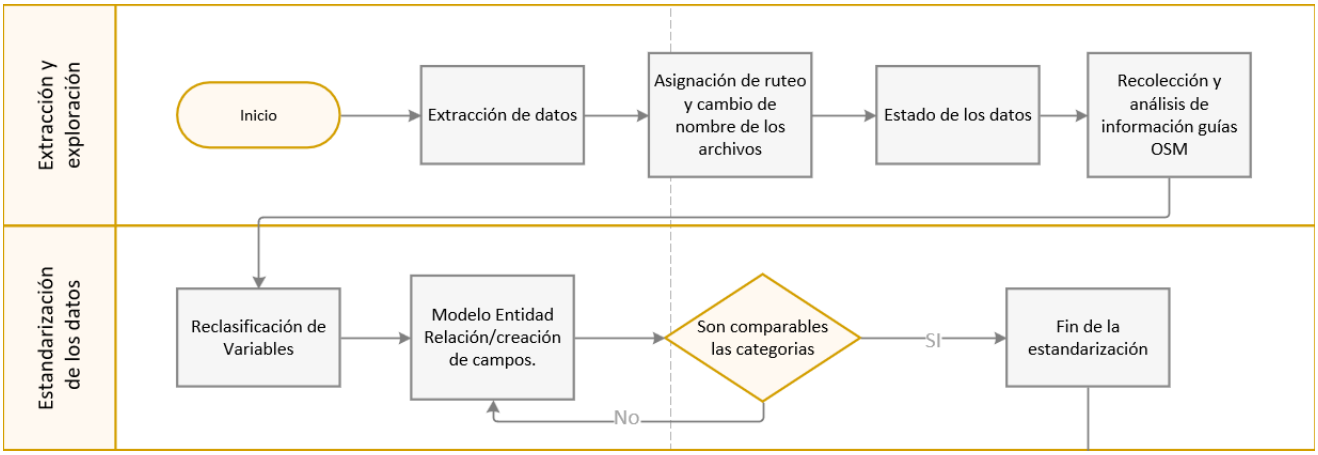


Figura 4-6: Flujo de desarrollo del objetivo1.

Extracción de datos

La extracción de los datos OSM fue realizada usando la librería *urllib2* de Python, esta librería permitió crear una conexión a la página URL (del inglés *Uniform Resource Locator*) que almacena estos datos haciendo posible obtenerlos a voluntad, permitiendo configurar el código Python a la ubicación de descarga de los datos. Aunque el procesamiento semi-automático de los datos no es el objeto principal de esta investigación, si juega un papel importante a la hora de asegurar la credibilidad de los resultados(Abdolmajidi et al., 2015). Como la mayoría de las operaciones espaciales se realizaron usando Python a través de la librería *arcpy*,y *shapely.geometry* fue necesario modificar el nombre original de los shapes files provenientes de OSM, ya que su sintaxis original no permitía que algunas librerías funcionaran adecuadamente y fuesen capaces de reconocer los datos OSM. Todos estos provenían del servidor geofabrik y estaban disponibles a nivel país. Como este articulo uso únicamente los datos OSM de la malla vial de Bogotá, se tuvo que realizar un recorte espacial automático, para seleccionar los datos de interés enmarcados en el área de Bogotá. Para realizar esto, se usaron las siguientes funciones:

- arcpy.MakeFeatureLayer_management()
- arcpy.SelectLayerByLocation_management()
- arcpy.CopyFeatures_management()

Estas funciones, creadas por el grupo (Enviromental system Research Institute; ESRI) permiten crear según el orden expuesto, capas temporales para almacenar resultados, seleccionar datos respecto a un parámetro espacial y finalmente copiar los resultados almacenados en la capa temporal creada. Una vez los datos OSM fueron descargados y extraídos de acuerdo con el área de trabajo, se procedió realizar el mismo proceso para la malla vial de la base de datos de IDECA.

```
arcpy.AddMessage("extract")

Lista_datos=arcpy.ListFeatureClasses()
print("feature_list",Lista_datos)

print("extact data...")

# Seleccionado los datos de OSM en Bogota, criterio nodo principal

arcpy.MakeFeatureLayer_management(r"E:/2018/Tesis_2018/Codig_Python/Data/colombia-latest-free_shp/Dataosm_roads_fre

arcpy.SelectLayerByLocation_management ("osm_roads_free_Bovl", "have_their_center_in", "E:/2018/Tesis_2018/DATA_Com

arcpy.CopyFeatures_management("osm_roads_free_Bovl", "E:/2018/Tesis_2018/Codig_Python/Data/colombia-latest-free_shp

print("Nodos...")
#correction about node intersection split
arcpy.AddMessage("Split Network  by nodes")

arcpy.FeatureToLine_management(osm_roads_free_Bogl, osm_roads_free_spliT, "0.001 Meters", "ATTRIBUTES")

print("Terminando...")
```

Figura 4-7. Segmento de código Python para extraer y recortar datos

Exploración de los datos

La exploración de los datos, importante para dar la primera valoración respecto al estado de las capas se ejecutó usando los atributos descritos en las **Tablas** ((4-1), (4-2)). Donde se realizó un conteo total de registros para cada capa usando una función count(distinc(Attribute). Luego, para cada uno de los atributos se estableció un conteo por

dominio para entender cuan complicada podría ser o no la estandarización de campos entre las dos fuentes de datos. Los conteos también fueron calculados en porcentaje para entender cuáles atributos y dominios tendrían una mayor cantidad de trabajo. Estos conteos fueron realizados usando la siguiente línea de código

```
arcpy.AddField_management(osm_roads_Ped_Removed_shp, "countN", "Int", "", "", "", "", "NULLABLE", "NON_REQUIRED", "")
countN=int(arcpy.GetCount_management("osm_roads_Ped_Removed_shp").getOutput(0))
```

Por otro lado, se calcularon otras estadísticas descriptivas básicas(Walpole, Myers, Myers, & Keying, 2012) como la media (ver Ecuación (4.6)), la mediana (ver Ecuación(4.17)) y la varianza (ver Ecuación (4.18)). A continuación, se muestran las fórmulas usadas en la exploración de datos:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (4.16); \quad \chi med = \frac{1}{2} \quad (4.17)\,; \quad S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (4.18)$$

En la ecuación (4.17) $\bar{x}$ corresponde al valor de la media y x_i al i esimo valor del conjunto de datos, en la ecuación número (4.17) $Fmed$ corresponde a la mediana y finalmente S_x^2 corresponde a la varianza, n es el número de datos y $(x_i - \bar{x})^2$ corresponde a la distancia que hay entre el x_i dato y la media. También se revisó manualmente el estado de los caracteres alfa numéricos en las dos bases de datos, con el fin de entender si existían elementos diferenciadores como lo son tildes comas puntos y espacios en las variables. En la exportación de datos preliminar también se analizaron los tipos de variables (string, integer, etc, boolean) entre las dos fuentes de datos para anticipar posibles discrepancias en el modelo de comparación usado.

Se realizó una revisión manual al estado de los metadatos IDECA, con el fin de comprender cual había sido la calidad calculada por la organización distrital. Durante este proceso se encontraron elementos que definieron los valores máximos de calidad en la malla vial IDECA con los cuales se definen los criterios límites de calidad. Lo que significó que, sí los datos VGI de OSM analizados superaban estos umbrales entonces podríamos haber concluido que los datos VGI de OSM tendrían mejor calidad que los encontrados en la malla vial de IDECA. Para realizar esto, se tomó el documento de especificación técnica de mapa de referencia IDECA (Unidad Administrativa Especial de Catastro Distrital -

Infraestructura de Datos Espaciales para el Distrito Capital -Gerencia IDECA, 2013)creado para la versión de base V17. y se analizaron todas las variables de calidad, partiendo desde el nivel de detalle de la cartografía y pasando por todas las medidas de calidad que a este trabajo competen.

Debido a que OSM contiene una serie de reglas para la codificación de los atributos viales en cada país, fue necesario estudiar la guía de mapeo OSM creada para Colombia[4] , junto con la definición de parámetros globales creados en OSM (Map_Features). Estos parámetros asignan etiquetas a cada elemento codificado dentro de OSM, cada etiqueta contiene un *key* que representa las subcategorías para cada clase. A continuación, se pueden observar los criterios usados para la evaluación de la guía de OSM.

Criterios revisados dentro de la Guía para mapear en Colombia y Map features (OSM)

- Etiquetado de las vías para su clasificación

El etiquetado de los segmentos de línea fue revisado manualmente uno a uno, buscando tanto el nombre y significado de las etiquetas como los parámetros de restricción que se tienen para ingresar datos a OSM. Para ello se usaron las listas de etiquetas creadas por OSM para ubicar y facilitar la posterior comparación.

- Nombramiento de las vías.

Para el nombramiento de las vías en OSM se buscaron y analizaron todas las etiquetas relacionadas al nombramiento, también se encontraron todas las restricciones creadas por OSM para el ingreso de datos.

- Clasificación vial OSM propuesta en Map features page publicada en la página de Wiki OSM.

Las siguientes Jerarquías creadas por OSM fueron revisadas de manera manual buscando aquellas que no fueran permitidas en la base de Colombia OSM.

[4] https://wiki.openstreetmap.org/wiki/Guide_for_mapping_in_Colombia

highway=motorway, highway=trunk, highway=primary, highway=secondary highway=tertiary, highway=residential, highway=living_street, highway=unclassified ,highway=service, highway=track

- Errores Comunes en codificación detectados por OSM

Todas las restricciones de codificación en relación a el nombramiento, y la jerarquización vial fueron colectadas a través de tablas, las cuales permitieron analizar el estado actual de los datos.

Estandarización de los datos por medio de análisis relacional.

Lo primero que se realizó fue una transformación en los datos a un único sistema de coordenadas, el sistema de coordenadas elegido fue el de GCS_MAGNA EPSG 4686. Luego y una vez comprendidos los lineamientos de codificación para OSM e IDECA, se pudieron identificar claramente los campos que necesitaron una reclasificación con el objetivo de poder ser comparados. Esta comparación se hizo mediante una correspondencia entre las clasificaciones dadas por OSM y IDECA, tomando como referencia el tipo y el uso de los atributos. Por ejemplo, si IDECA tuvo una clasificación de vías en 6 categorías, entonces los datos de OSM deben migrar a 6 categorías de acuerdo al uso y al tipo de atributo definido, asegurando que las muchas categorías viales de OSM encajaran en al menos una de las categorías creadas por IDECA. Esto se realizó creando modelo entidad relación *E-R* siguiendo en términos generales el trabajo de(Nowak Da Costa, 2016b) y usando las variables descritas en la **Tabla**(4-5). Algunas categorías dentro del campo *Fclass* de OSM no pudieron ser relacionados a una categoría en IDECA, por lo que debieron ser excluidos del estudio.

Tabla 4-5: Atributos revisados para la estandarización.

Atributos revisados para la estandarización			
Campos	Descripción	Sub categorías	Data
MVITipo	Tipo de vía (Cl, KR, TV, DG)	No	IDECA
MVINombre	Nombre de la vía	No	IDECA
MVISVia	Sentido de la vía	Si	IDECA
MVIEtiquet	Nombre de la vía principal	No	IDECA
Tipo de clasificación	Jerarquía de la vía	Si	OSM
Fclass	Clasificación de la vía	Si	OSM
Name	Nombre de la vía	No	OSM
Oneway	Sentido de la vía	Si	OSM

Para crear el modelo inicial (ver **Figura** (4.8)), se tuvieron presentes las siguientes directrices. Relación en los campos de sentido vial: donde un sentido vial en IDECA (B, FT, N, SD, TF, NULL) podría tener más de un valor en OSM (B, F, T), respetando los valores Hacia, Desde y Doble sentido. Con esta directriz se estandarizó la dirección de

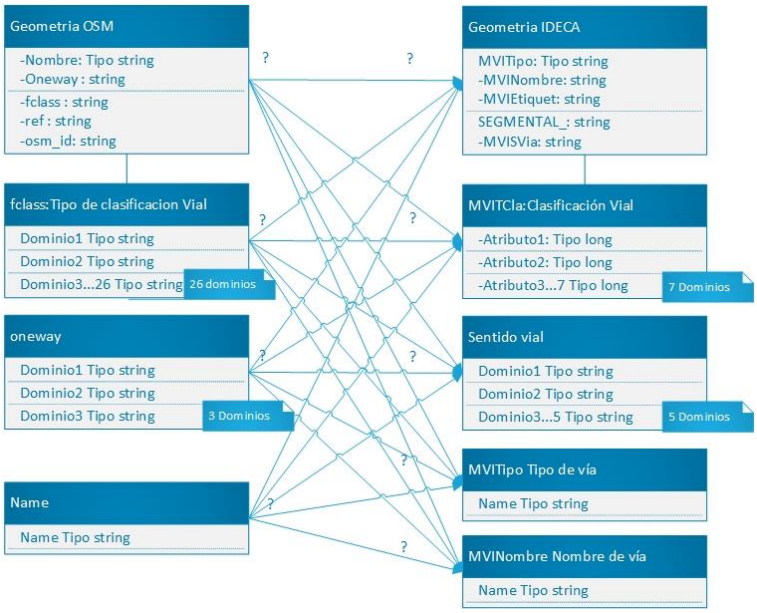


Figura 4-8: Modelo E R Inicial.

flujo para IDECA usando los valores: B F T y se ajustaron los valores de IDECA, los valores N se ajustaron como B (doble sentido de circulación). Jerarquía vial: Se estableció que la malla vial OSM tenía demasiadas categorías, las cuales podían ser reducidas y estandarizadas a las codificadas en IDECA, Respetando que las muchas clases dentro de la jerarquía vial OSM podrían corresponder a un único valor en la clasificación jerárquica de IDECA. La relación entre campos se puede observar en la **Tabla** (4-6).

Tabla 4-6: Relación de Jerarquía vial IDECA OSM.

Relación de jerarquías por uso y tipo entre IDECA y OSM	
Dominios en jerarquía vial IDECA	Dominios en Jerarquía vial OSM
Malla vial arterial	primary + primary link + trunck
Malla vial intermedia	Secundary + secondary_link + tertiary + tertiary_link + trunck_link
Malla vial local	residencial + living_street*
Malla vial rural	track1, track 2, track 3, track 4, track5
Sin definir	unclasified + unknow + services
Proyectadas	Na
Peatonal	Na

Nombramiento de las Vías: Todos los datos de OSM debían ser estandarizados de acuerdo a los datos *IDECA*, separando el tipo de vía y el nombre correspondiente. Respetando la relación uno a uno. Sin embargo, aquí existieron problemas debido a que existían muchos datos errados.

Con el modelo entidad relación definido, se pudieron materializar las relaciones anteriormente expuestas. Para luego crear los campos de estandarización en las respectivas capas, para lograrlo, se crearon consultas automáticas donde se seleccionaban y agrupaban elementos de acuerdo a las reglas establecidas previamente. Algunas funciones usadas para automatizar el proceso fueron:

- arcpy.Select_analysis(): Función para seleccionar datos usando consultas Structured Query Language SQL.

- arcpy.AddField_management() Función para crear nuevos campos, los cuales contenían la nueva clasificación de los datos de acuerdo a el tipo de atributo descrito en el modelo E-R.

4.3.2 Examen y complementación de los datos de texto VGI para la estandarización

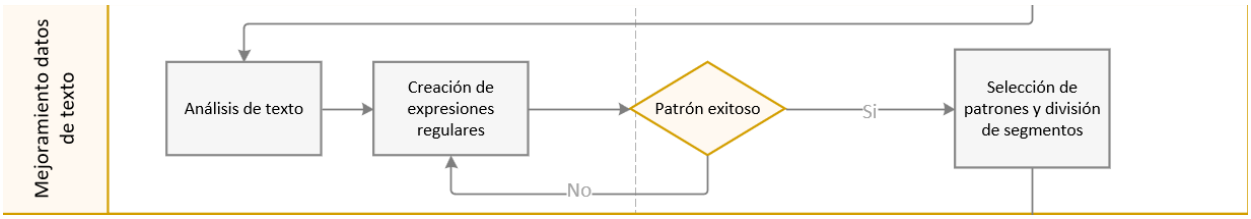


Figura 4-9: Flujo de desarrollo del objetivo2.

Mejoramiento de los datos OSM por medio de expresiones regulares

Definida y estandarizada la estructura de comparación, el siguiente desafío consistió en poder comparar campos compuestos por cadenas de caracteres, como se tenía previsto los datos OSM contenían inconsistencias en cuanto a digitación de texto lo que forzó a estandarizar los nombres de acuerdo a tipo de calle (*Cl, KR, TV, DG*) y nombre de calle. Este estudio no analizó el nombre compuesto por las dificultades en cuanto a la búsqueda de patrones más completos.

La estandarización de nombres se logró usando expresiones regulares conocidas como Regex, Las expresiones regulares son un lenguaje de descripción de texto, que permite encontrar patrones, comparar y dividir segmentos coincidentes de texto(Nelli, 2015, p. 155).Para todas las cadenas de caracteres relacionadas a los nombres viales (Calle, Carreras, Diagonales, transversales, Avenidas) fueron creados patrones con el fin de mejorar el contexto de los nombres viales en *OSM*. Pues, aunque no estuviesen bien

escritos, sabemos por conocimiento local el significado del texto, entonces cuando se encontraron cadenas de caracteres como *cale, carera, Cal, Carrerraaaa*, *Acle*,, Dgnal, tsversal, etc, estas fueron mejoradas a partir de algunos de los siguientes patrones:

- **[**Ca**.**le**]+***$*: Busca cualquier patrón relacionado con la palabra calle. inclusive si existen espacios entre las letras. *Klle, Cale, Cll, Clle*, son elementos detectados como pertenecientes al patrón de calle. Lo que se hizo fue encontrar el patrón y convertirlo a la cadena de caracteres que se requiere, en este caso CL.

- **^**Ca.le**{**0,1**}**: Busca el patrón Calle sin que existan repeticiones al final de la línea, entonces. Si existen palabras como *Calleeeesss,* el patrón seleccionará únicamente los caracteres hasta que la letra *e* exista unasola vez al final de la línea.
- **^**Ca.?le**{**0,1**}**: Permite encontrar la palabra Calle mal escrita, por ejemplo cuando se escribe la palabra *Cale* o *Cael.* Con este patrón sabemos que el posible significado era probablemente Calle.
- **(**^Ca.**?**le**{**0,1**})*****: El patrón se puede volver más robusto, aquí se creó uno por grupos, eliminando los demás caracteres con el fin de poder localizar fácilmente los aciertos.

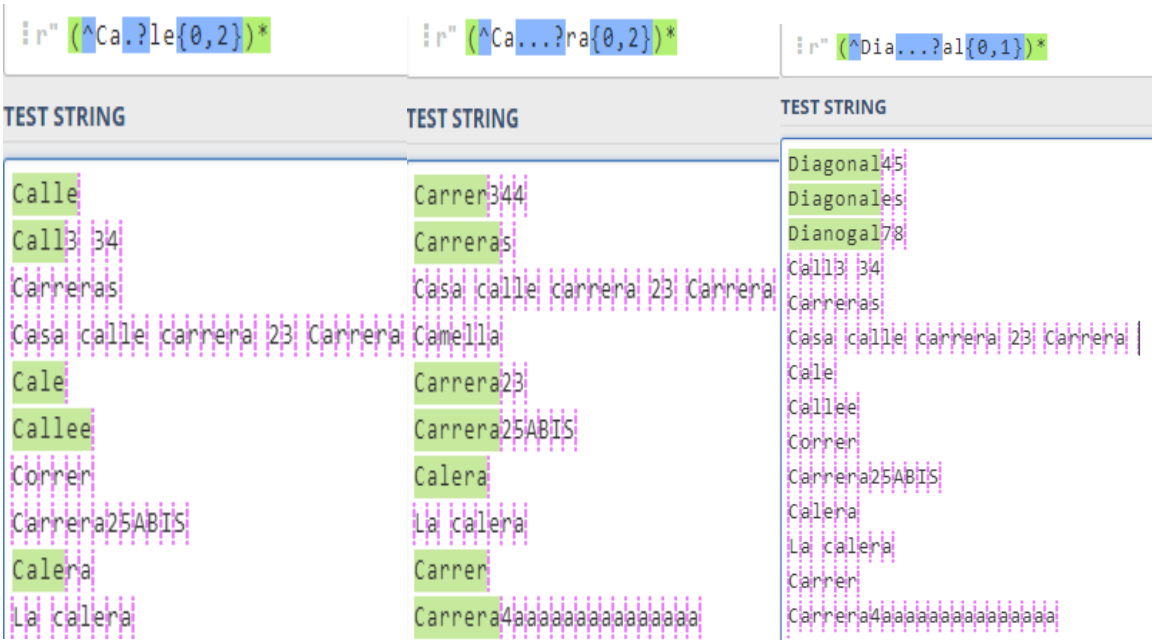


Figura 4-10: Ejemplo Patrón de búsqueda.

- **^(**Calle**|**calle**|**Cl**{**2**}|**cl**|**AvCalle**|.**venida **.**alle**)**: Con este patrón podemos seleccionar todas las palabras que tengan avenidas y estén relacionadas con la palabra Calle/calle.
- **^(**Carrera**|**carrera**|**Cl**{**2**}|**cr**|**AvCarrera**|.**venida **.**era**)**: Con este patrón podemos seleccionar todas las palabras que tengan avenidas y estén relacionadas con la palabra Carrera/carrera.
- **[**0-9**]+[**A-Za-z**]*** selecciona expresiones que inicien con números y continúen con letras, esto es apropiado para seleccionar los nombres de las vías con números contiguos al texto.
- **^([**0-9**]+)$** Patrón para seleccionar expresiones que inicien con números pero que no tengan letras.
- **^([**0**+)$**Selecciona las cadenas que inician con 0 y que no tienen asociado otro número.
- **^ [**A-Z**]+$**selecciona expresiones que inicien con letras y continúen con letras.

Todos los patrones creados y usados podrán ser consultados en el **Anexo A** “Patrones creados a partir de expresiones regulares “

Con la ayuda de estas expresiones se encontraron patrones generales para detectar el tipo de via (*Cl, KR, TV, DG*) inclusive, como se mencionó anteriormente, si se encontraba mal digitado. También se logró estandarizar el nombre de la vía principal, limpiando espacios al inicio y fin de cada registro.

Para la edición de las cadenas de caracteres de OSM se corre el siguiente código, el cual selecciono, reemplazo y edito las cadenas de caracteres con el fin de mejorar y estandarizar los datos. Algunas cadenas de caracteres no pudieron ser ajustadas debido a su complejidad:

```
import re
import sqlite3

con=sqlite3.connect('osm_roads_Named.db')
query =  "SELECT name FROM osm_roads_free_spliT WHERE name> IS NOT NULL;"
```

```
name = pd. read_sql_query ( query , conn )
t=len(name)
r=1
Nnames = []
patron1=re.compile(r'^(Calle|calle|Cl{2}|cl|AvCalle|.venida .alle)')
patron2=re.compile(r'^(Carrera|carrera|Cl{2}|cr|AvCarrera|.venida .era)')
patron3=re.compile(r'(^Ca.?le{0,1})*')
patron4=re.compile(r'(^Ca...?a{0,1})*')
patron5=re.compile(r'(^Dia...?al{0,1})*')
.
.
.

t=0
patronx=re.compile(r'Ca.le]+$:') #hasta los n patrones creados

        for i in range(0,t):

        Coincidencias=re.findall(patron+ , [name[i]]))
        Nname[i]= Coincidencias.append(r)
```

De manera similar se corrió código para todas las sentencias de búsqueda match(,),

search(,) , full.match(,) y coincidencia start(), end(), group(). Luego se depuro el vector de caracteres con las funciones siguientes: replace*(), .index(), .Lstrip(). Rstrip()* contenidas en el lenguaje Python.

Estas funciones permitieron hacer respetar las siguientes reglas:

(I) No pueden existir dos tipos de vía en una misma etiqueta: Calle 34; calle 23. (II) No pueden existir espacios al inicio, (III) No puede existir caracteres especiales = / # -$.
(IV) No puede existir ninguna combinación de Calle, Carrera Avenida que no terminé en vocal.
(V) Nomenclatura y nombres deben estar separados para su comparación. (VI) Bis debe ir en mayúsculas y sin espacio entre el número: 4 Bis: 4BIS.

Finalmente, todas las cadenas de caracteres quedaron normalizadas para su posterior comparación (*KR #, CL #, TV #, DG #...*).

4.3.3 Comparación y análisis semi automático usando muestras estratificadas para clasificar de manera multivariada la calidad

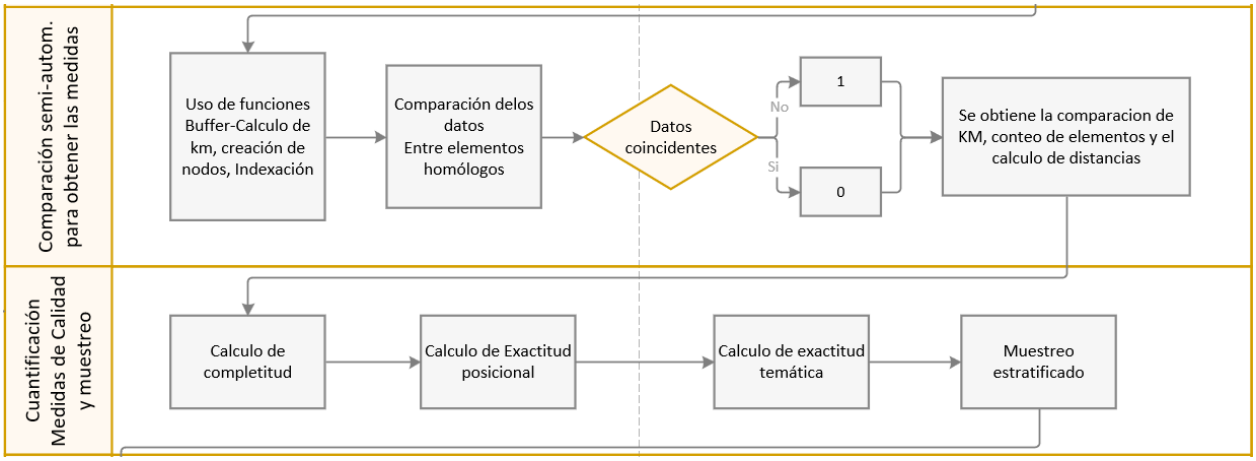


Figura 4-11: Flujo de desarrollo del objetivo 3.

Comparación semi automática de los datos para obtener los valores que permitirán evaluar la calidad

Una vez se estandarizaron los datos y con las columnas listas para almacenar los resultados se hizo uso de una serie de funciones que permitieron crear relaciones espaciales por medio del uso de técnicas de indexación, cálculo de mínima distancia, conteo de nodos, uso de buffers y cálculo de kilómetros. Esto se pudo realizar siguiendo en términos generales las metodologías creadas por Goodchild et al. (1997)y Hakaly (2010) , usadas para medir la exactitud posicional y la completitud. Por un lado, el método del primer autor consistió en crear un buffer de ancho fijo alrededor de la fuente de referencia donde el porcentaje de objetos que cayeran dentro de esta zona de influencia eran comparados y evaluados.

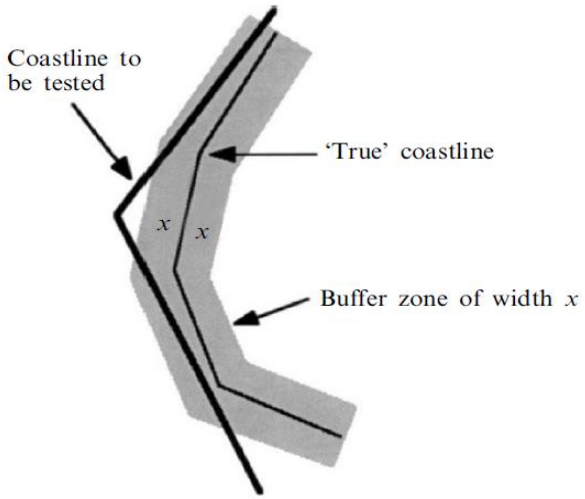


Figura 4-12: Método de búfer Goodchild 1997.

Fuente: Goodchild 1997.

Aquí se aplicó la metodología de goodchild, con las siguientes modificaciones.

Primero se cortaron las vías *OSM* de acuerdo a cada una de sus intersecciones. Esto hizo que los segmentos quedaran con longitudes lo más parecidas posibles a la fuente de *IDECA*, ya que la fuente de control tenía todas sus vías segmentadas de acuerdo a sus intersecciones con otras vías. Después se crearon y calcularon los campos para medir Km en cada una de las capas (mallas viales *OSM -IDECA*). Se aclara que en este paso todos los campos de jerarquía vial, dirección de flujo y nombre de las vías ya estaban estandarizados y listos para ser comparados. Después, en lugar de usar la geometría tipo línea y realizar un búfer estático, aquí se usaron los nodos centrales de cada uno de los segmentos viales de *OSM* e *IDECA*, transfiriéndoles a estos todos los atributos que les correspondían, entre ellos estuvieron los campos con los valores estandarizados de nombramiento de las vías y jerarquía vial y cálculo de longitud del segmento. Luego se aplicó un buffer móvil que dependiera de la distancia del objeto más cercano al nodo central de cada vía (Ver **Figura** (4-11)). Para ello se calculó la mínima distancia que existía desde un nodo central hacia otro nodo central de la fuente de control, que en este caso fue *IDECA*, como se puede apreciar, (En Rojo el nodo central y las vías *OSM* y en color verde los nodos y las vías *IDECA*.)

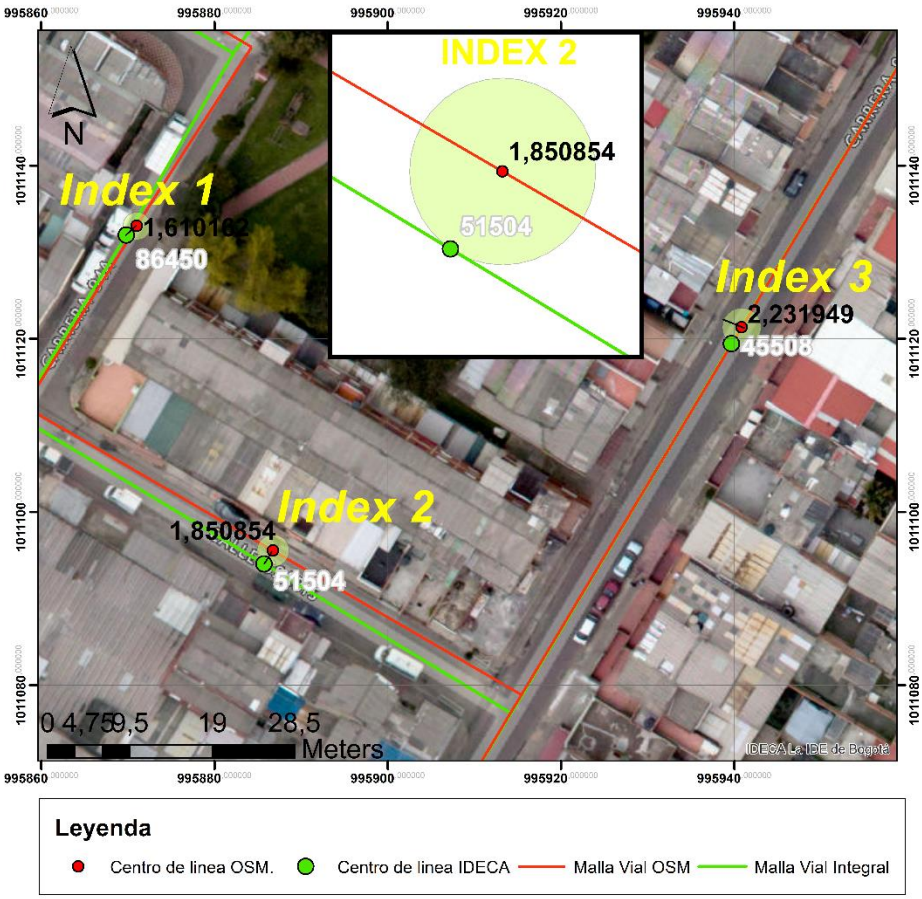


Figura 4-13: Comparación automática de atributos.

Cada nodo fue identificado por medio de una función de indexado, las funciones de indexado permiten ordenar los datos de acuerdo a ciertas características, Esta indexación permitió saber que pareja de nodos comparar, por ejemplo, si vemos la (**Figura** (4-13)) y tomamos como referencia el nodo indicado como Index 3, podremos observar en color negro, el valor de la mínima distancia a la cual el algoritmo encontró un nodo. Para el ejemplo el valor fue de 2.3 metros. Aquella fue la distancia donde el buffer dejo de buscar un nodo cercano. En color blanco y sobre el nodo con el rotulo Index 3, podemos observar el número 45508. Este número corresponde al indexado de estos dos nodos (*OSM - IDECA*). Con este número (ID) el algoritmo entiende que para esos nodos la comparación de atributos tuvo sentido, pues los dos nodos poseen ese mismo identificador. En este paso los atributos fueron comparados uno a uno, tomando la siguiente regla, sí el valor es

igual en los dos nodos para el atributo revisado, entonces el campo que aloja la comparación tendrá el valor 0, indicando ningún cambio (valor correcto), de lo contrario, sí el algoritmo encuentra que existe una diferencia entre los atributos en los nodos, entonces se obtendrá el valor 1 (incorrecto) indicando que sí existió un cambio. Si los atributos de *OSM* se encuentran vacíos y por el contrario los de *IDECA* contiene información, entonces el algoritmo contará el valor como 1 para *OSM* indicando que el nodo existe (pero los atributos no). Cuando el algoritmo no puede detectar su nodo par, entonces asume que la geometría en OSM está faltando, por ello cuenta esto como 1.

El proceso del algoritmo se puede observar en la (**Figura** (4.14))

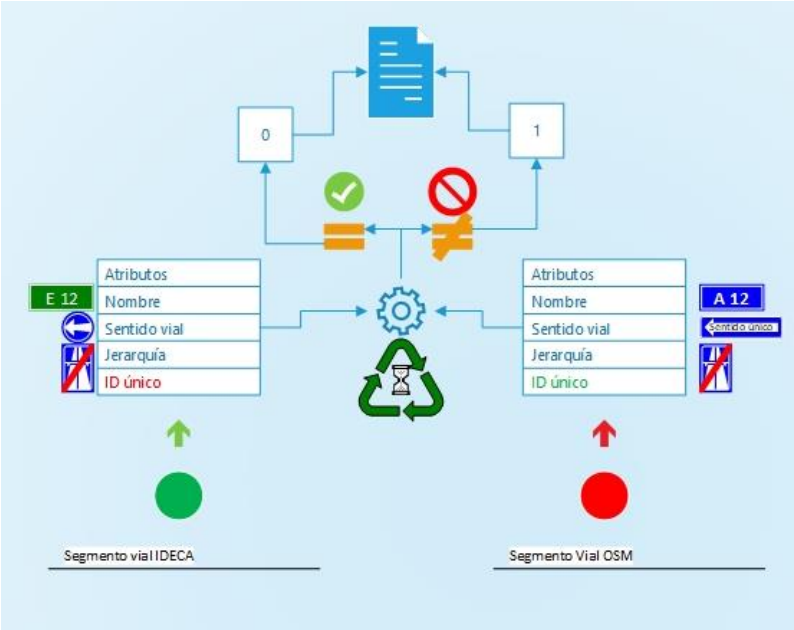


Figura 4-14: Diagrama del algoritmo para la comparación de atributos.

El cálculo de la distancia mínima a el objeto más cercano se puede apreciar en la siguiente imagen. Lo que hace este cálculo es buscar el nodo próximo y luego calcula la distancia a este elemento, un problema encontrado al usar el cálculo de la distancia a el nodo más cercano consiste en que muchas veces, cuando no hay un nodo cercano como se puede apreciar en la figura, el cálculo de la distancia se incrementa hasta que finalmente encuentra uno, esto podía llegar a ocasionar problemas cuando se indexaran nodos con esta característica, pues se compararían atributos que no corresponden a la geometría. Por esta razón se tuvo que realizar una revisión superficial evitando que cálculos mayores

a 30 metros hicieran match con otros nodos pues aquí ya se confirmaba la ausencia de geometría por lo cual el valor de geometría faltante sería 1 en todos los casos.

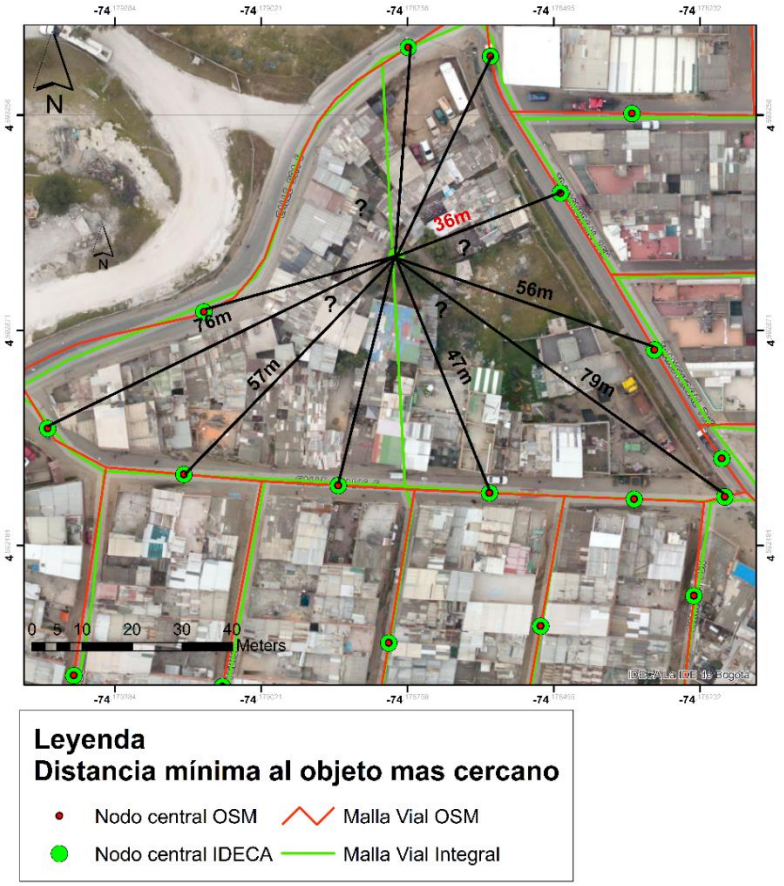


Figura 4-15: Problema de nodos y distancias

Por otra parte, el método de Haklay usado para medir la completitud de los datos consistió en crear grillas de 5X5 *Km* donde se contó la diferencia de Km en la malla vial de Londres(Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2016a). De manera similar, aquí se usaron las localidades de Bogotá en lugar de grillas para contar los Km y nodos de los segmentos línea, y así detectar ausencia o no de elementos. En la **Figura** (4-16) se pueden observar algunas localidades con sus respectivos nodos. El pareo de estos nodos dentro de cada localidad permitió contar la cantidad de elementos por comisión y omisión.

Con este proceso culmina la comparación automática, dando como resultado una tabla que contiene los valores comparados por localidad. Esta tabla fue el insumo usado para

calcular la calidad VGI usando los indicadores expuestos en el capítulo dos: sección medidas de calidad. Este insumo también fue usado para analizar la calidad VGI usando el análisis de correspondencia múltiple y método de conglomerados. Esta tabla puede ser consultada en el **Anexo B** “Tabla comparación de atributos “.

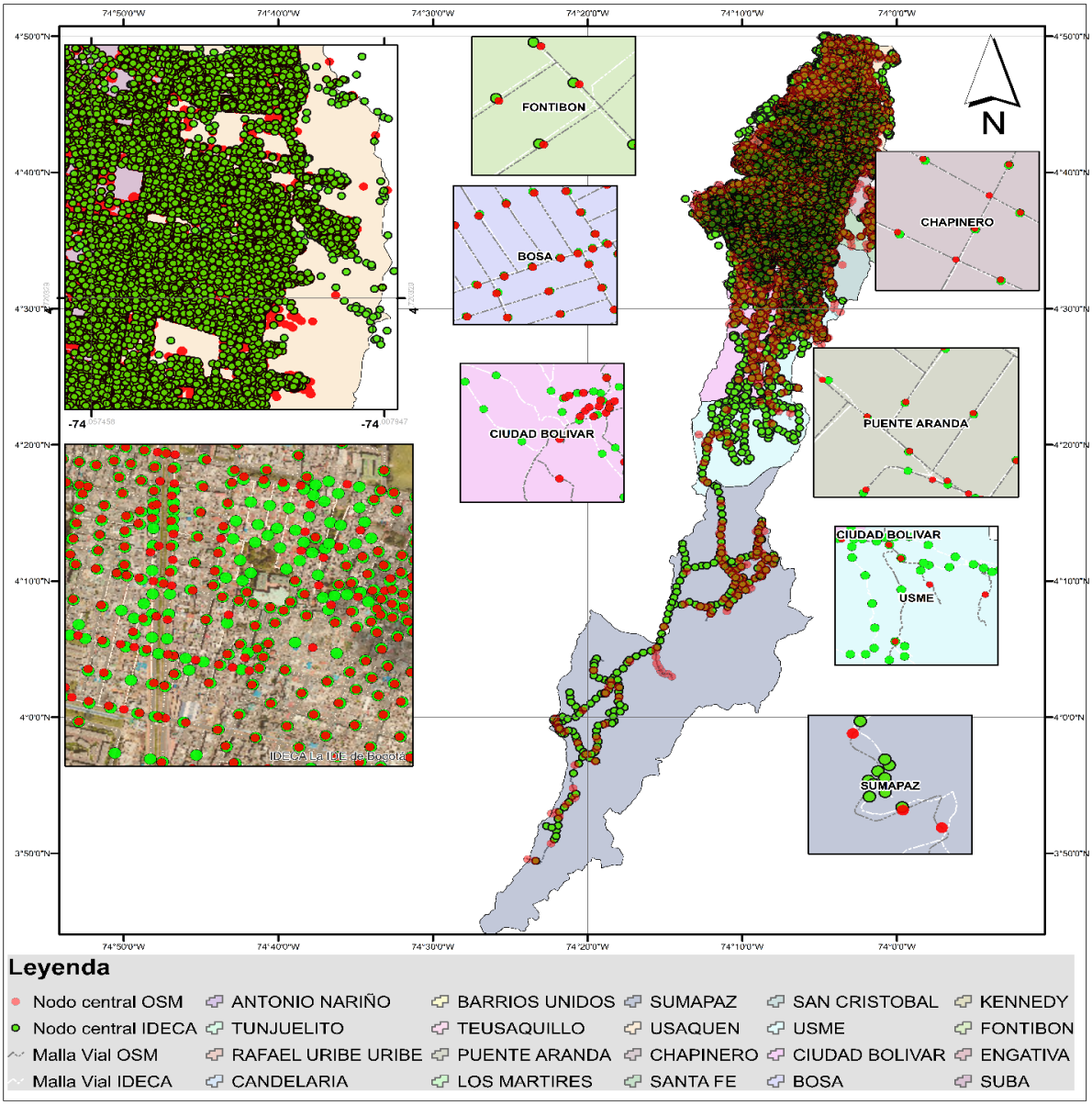
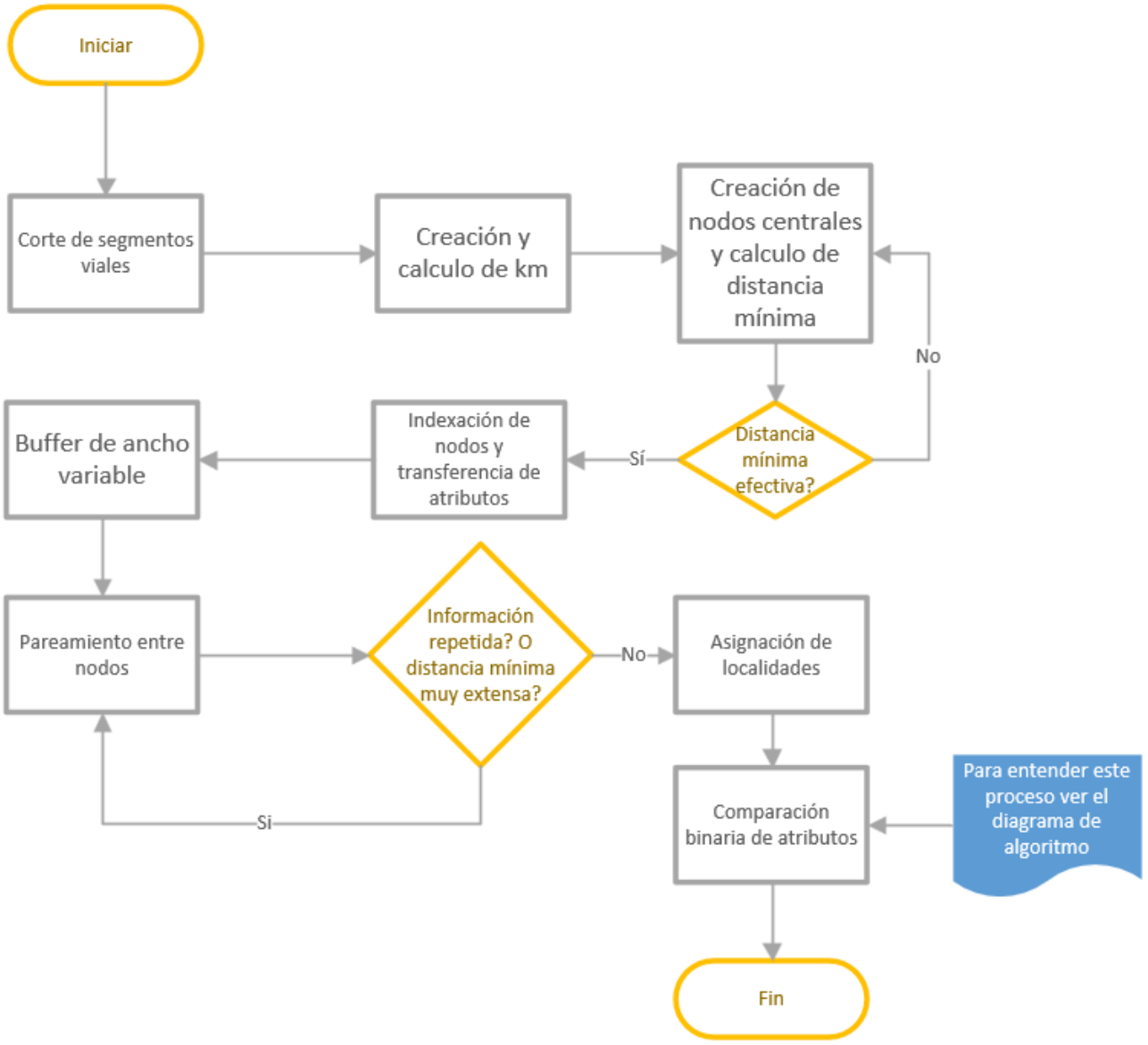


Figura 4-16: Creación de nodos dentro de las localidades.

A continuación, se expone el diagrama del match semi automático:



Evaluación de completitud, exactitud posicional y exactitud temática

Cálculo Completitud (omisión o comisión)

La evaluación de la completitud de los datos relacionada a la ausencia o exceso de elementos se calculó usando el método automático expuesto anteriormente, el cual creó una relación entre nodos de línea más cercanos para las dos fuentes de datos. Se calcularon los *Km* y número de nodos centrales de cada línea. Estos resultados se agruparon por localidad, donde se comparó: número de elementos y Km para determinar la ausencia o exceso de elementos en la malla *OSM* respecto a la fuente de referencia.

Calculo exactitud posicional

La exactitud posicional absoluta, la cual hace referencia a la exactitud de la posición de un elemento con respecto a otro de mayor precisión, fue calculada mediante la (Ecuación (4.19)). Se usaron las coordenadas para cada uno de los nodos centrales de la malla vial *OSM* y *IDECA*, luego se midieron las diferencias en la componente horizontal *X, Y*, como se muestra a continuación:

$$e_i = \sqrt{(X_{if} - Xir)^2 + (Y_{if} - Yir)^2} \qquad (4.19)$$

Donde e representa el error horizontal en cada punto, X_{if}, Y_{if} son consideradas las coordenadas de la fuente a corroborar, en este caso OSM, mientras que $Xir\ , Yir$ son las coordenadas consideradas como verdaderas.

El error medio cuadrático o *RMSE* se calculó como:

$$RMSE_r = \frac{1}{n}\sum_{i=1}^{n}(e_i) \qquad (4.20)$$

Siendo n igual al número de registros observados.

Finalmente, para calcular la exactitud posicional con un nivel de confianza del 95% según National Standard for Spatial Data Accuracy NSSDA y observando que los errores medios cuadráticos en la componente de las *X* son considerados iguales a los encontrados en el eje de las *Y* (Greenwalt and Schultz, 1968). Finalmente, se obtuvo que:

$$Error\ 95\% = 1.7308 * RMSE_r \quad (4.21)$$

Calculo exactitud temática

La evaluación de la exactitud temática se realizó de manera binaria. Si el dato estandarizado en *OSM* no coincidía exactamente con el valor codificado en *IDECA* entonces la casilla tomaba el valor de 1, de lo contrario tomaba el valor de 0. Al final de esta comparación se obtuvo un número de ítems de cada objeto incorrectamente clasificados por localidad para cada uno de los atributos comparados, los cuales fueron dirección de flujo, clasificación vial (jerarquía vial) y nombramiento de la vía principal.
Para establecer la medida de exactitud en la clasificación se tomó el número de errores encontrados Ne divididos sobre el número total de elementos N :

$$\%\ error = \frac{Ne}{N}.100 \quad (4.22)$$

Se estableció un criterio de aceptación de hasta el 5% de elementos mal clasificados. Este valor es tomado con base a los parámetros de calidad ya calculados por IDECA donde el 5% de los datos mal clasificados resulta ser el umbral de aceptación de calidad para la exactitud temática(C et al., 2012).

Muestreo simple por asignación proporcional a la localidad y re clasificación de variables

Muestreo simple por asignación proporcional

Como la tabla comparación de atributos **Anexo B** obtenida en la sección 4.6 y de la cual se obtuvieron los valores de calidad desarrollados en el apartado anterior, contenía demasiados registros. Se optó por crear un muestreo estratificado que representara los datos (errores y aciertos) encontrados en cada una de las localidades. A demás de la cantidad de datos, la justificación para realizar un muestreo estratificado radicó en que el algoritmo de ACM no fue capaz de procesar los datos de la población, haciendo que existiera un procesamiento de más de 30.8 Gb.

```
Consola R. Error: cannot allocate vector of size 30.8 Gb
```

Como se conocía la cantidad de nodos dentro de cada localidad N, se procedió a calcular los tamaños de muestra n para cada una de ellas. La fórmula usada para realizar el muestreo aleatorio simple MAS dentro de cada estrato fue

$$MAS = \frac{NZ^2p(1-p)}{e^2(N-1)+Z^2p(1-p)} \qquad (4.23)$$

Este tamaño de muestra aleatorio se calculó con un $Z = 1.96$, que corresponde a un nivel de confianza del 95%, para una proporción p del 50% y un margen de error aceptado del 5%. Este proceso se repitió para cada una de las localidades. La razón de no hacer un muestreo aleatorio proporcional como comúnmente se usa, radico en que la muestra final bajo ese método era muy pequeña lo que afectaba el análisis de conglomerados. En la (**Figura** (4.17)) se puede apreciar el tamaño de muestreo para cada una de las localidades.

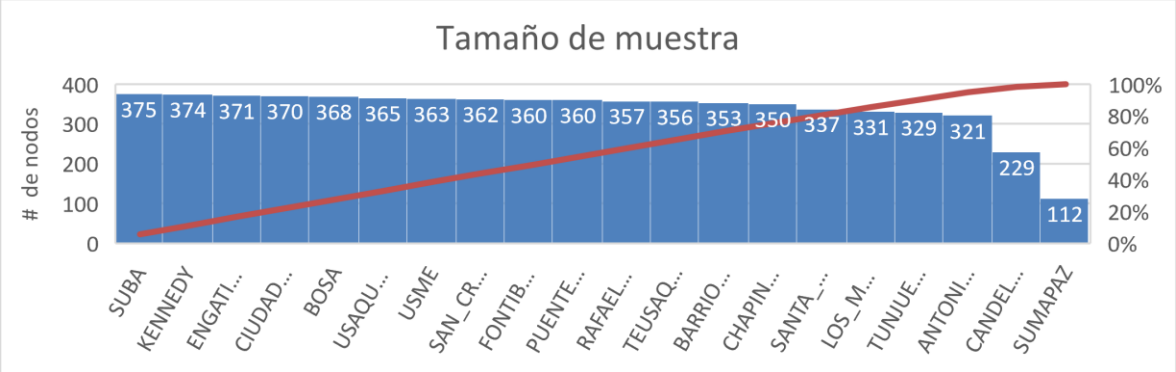


Figura 4-17: Tamaño de muestra.

Después de obtener el tamaño de muestra para cada uno de las localidades se usó una función aleatoria (*Random*(x,y)) para seleccionar los nodos dentro de cada una de las localidades, de acuerdo con el número de muestra obtenido. Con este muestreo disminuyo considerablemente la cantidad de registros (94.5% menos) y se pudo correr el Análisis de correspondencia múltiple.

Re categorización de variables

El muestreo estratificado no fue el único cambio que tuvo la tabla de comparación de atributos adjunta en el **Anexo B**. Como puede recordar el lector, esta tabla contuvo la comparación de los atributos objeto de estudio entre *OSM* y *IDECA*, donde las variables obtenidas eran de tipo binario *(0,1),* exceptuando el cálculo de distancia y el nombre de las localidades. Las variables binarias fueron reescritas a categorías, básicamente por dos razones, la primera de ellas, consiste en que el ACM trabaja con variables categóricas y aunque los valores 0 y 1 corresponden a ese tipo categórico binario, el análisis visual y de relación se hace muy complicado de desarrollar. La segunda, es para brindar una mejor visualización de resultados. En la (**Tabla** (4.7)) se puede ver con más detalle la recategorización mencionada.

Tabla 4-7: Re categorización de las variables

Re categorización de las variables		
Variables	Valor	Re categorización
Jerarquía_Vial	1	Error_Jerárquico (E_J)
	0	No_error_Jerárquico (N_3_J)
Sentido_Vial	1	Error_Sent_Vial (E_S_V)
	0	No_error_Sent_Vial (N_e_S_V)
Nombre_vial	1	Error_Nomencla (E_N)
	0	No_Error_Nomencla (N_e_N)

Exac_posicional	Variable continúa transformada en categórica	GQ, MQ, WQ

La tabla llamada *Tabla_OSM_IDECA.txt* que contiene los datos del muestreo y la reclasificación de variables puede ser consultado en el **Anexo C** “Muestreo por estrato “.

	row.names	Exac_po	Localidades	Jerarquia_Vial	Sentido_Vial	Nombre_vial
1701	26730	2.16>Exac_po<=2.54	SANTA_FE	No_error_Jerarquico	No_error_Sent_Vial	No_Error_Nomencla
1702	26733	1.97>Exac_po<=2.16	SAN_CRISTOBAL	No_error_Jerarquico	No_error_Sent_Vial	No_Error_Nomencla
1703	26736	Exac_po<=1.88	RAFAEL_URIBE_URIBE	Error_Jerarquia	No_error_Sent_Vial	No_Error_Nomencla
1704	26737	Exac_po>2.54	TUNJUELITO	No_error_Jerarquico	No_error_Sent_Vial	No_Error_Nomencla
1705	26806	2.16>Exac_po<=2.54	TUNJUELITO	No_error_Jerarquico	No_error_Sent_Vial	Error_Nomencla
1706	26833	Exac_po<=1.88	RAFAEL_URIBE_URIBE	Error_Jerarquia	No_error_Sent_Vial	Error_Nomencla
1707	26842	Out_of_scope	TUNJUELITO	Error_Jerarquia	No_error_Sent_Vial	Error_Nomencla
1708	26852	Out_of_scope	CIUDAD_BOLIVAR	Error_Jerarquia	No_error_Sent_Vial	No_Error_Nomencla
1709	26862	Exac_po>2.54	TUNJUELITO	No_error_Jerarquico	Error_Sent_Vial	No_Error_Nomencla
1710	26880	Out_of_scope	SANTA_FE	No_error_Jerarquico	No_error_Sent_Vial	No_Error_Nomencla
1711	26893	Exac_po<=1.88	SANTA_FE	Error_Jerarquia	No_error_Sent_Vial	No_Error_Nomencla
1712	26903	Exac_po>2.54	CIUDAD_BOLIVAR	No_error_Jerarquico	No_error_Sent_Vial	Error_Nomencla
1713	26915	Exac_po<=1.88	ANTONIO_NARIÑO	Error_Jerarquia	Error_Sent_Vial	No_Error_Nomencla
1714	26933	Exac_po<=1.88	ANTONIO_NARIÑO	No_error_Jerarquico	No_error_Sent_Vial	No_Error_Nomencla
1715	26936	Out_of_scope	SANTA_FE	No_error_Jerarquico	No_error_Sent_Vial	Error_Nomencla
1716	26948	Exac_po>2.54	TUNJUELITO	No_error_Jerarquico	Error_Sent_Vial	No_Error_Nomencla
1717	26949	Exac_po>2.54	TUNJUELITO	No_error_Jerarquico	No_error_Sent_Vial	Error_Nomencla
1718	26956	Out_of_scope	SANTA_FE	Error_Jerarquia	No_error_Sent_Vial	No_Error_Nomencla
1719	26974	Out_of_scope	SANTA_FE	No_error_Jerarquico	No_error_Sent_Vial	Error_Nomencla
1720	26976	Exac_po<=1.88	ANTONIO_NARIÑO	Error_Jerarquia	Error_Sent_Vial	No_Error_Nomencla

Figura 4-18: Tabla de datos.

Agrupación y clasificación de los resultados mediante técnicas multivariantes (ACM)

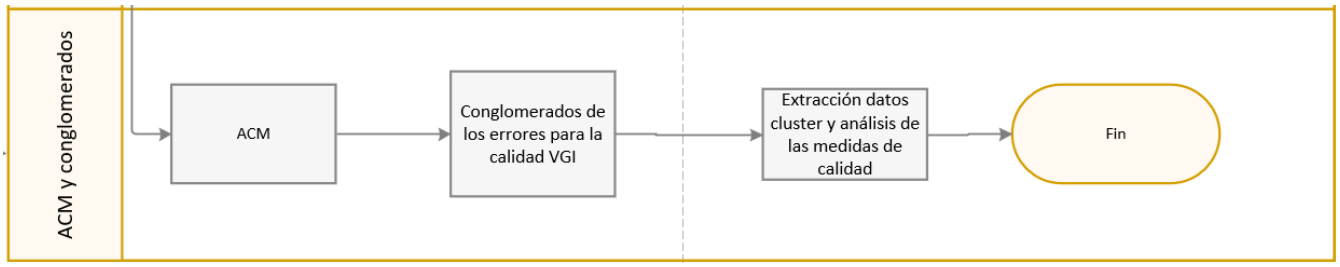


Figura 4-19: Flujo de desarrollo del objetivo 4.

Análisis de Correspondencia múltiple ACM

Con el muestreo de los datos y las variables reescritas como categóricas, se procedió a aplicar el ACM. Este análisis fue desarrollado en *RStudio* en su versión 1.1.423, usando principalmente el plugin de *FactoMine R* y la librería *ade4*. Todo el código elaborado, junto con las salidas gráficas puede ser consultado en el **Anexo G** "Codigo R ".

Este análisis tomó como datos de partida la tabla *Data_Comparation_OSM_IDECA_RS (***Anexo C***)*. Desde esta tabla (Ver **Figura** (4.18)) y con ayuda de la función. mca*( )* se calculó la tabla binaria X. Esta tabla cruza los $n = 6424$ individuos con las categorías de cada variable (Ver ecuación (4.24)), donde el número de variables K=6 y cada k contiene P_i Categorías.

Para calcular el número de categorías se usó la siguiente ecuación:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} p_i = 34 \quad (4.24)$$

La tabla binaria registró la aparición de las categorías por cada variable, lo que significa que la tabla asignó el valor de 1 si el individuo (filas) poseían el atributo (*Categoría de la variable objeto de estudio :En este caso se refiere a la aparición o no de un error*) y asignó 0 a las demás categorías dentro de esa variable. Aquí la suma de columnas mostró la cantidad de individuos que participaron en cada una de las categorías (errores), mientras que la suma de las filas fue una constante que mostró la cantidad de categorías por individuo (nodo).

A partir de la tabla binaria X (ver **Figura** (4.20)) se construyó la tabla de Burt, que es el resultado de todas las posibles tablas binarias(Murtagh, 2007) de las 6 variables tratadas aquí. Estas variables fueron: (Localidades y numero de registros) que constituyen los datos para analizar la completitud de los datos, (errores de jerarquía vial, errores sentido vial, nombre de vía) que constituyen los datos para calcular la exactitud temática y finalmente la variable exactitud posicional.

La tabla de Burt $B = X^{'}X$, permitió calcular los perfiles fila y columna que son las frecuencias marginales de cada una de las categorías por variable. Para realizar

	row.names	Exac_po.1.88>Exac_po<=1.97	Exac_po.1.97>Exac_po<=2.16	Exac_po.2.16>Exac_po<=2.54	Exac_po.Exac_po<=1.88
131	484	0	0	0	0
132	486	0	0	0	0
133	498	1	0	0	0
134	511	0	0	0	1
135	536	0	0	0	0
136	552	0	0	0	0
137	577	0	0	0	1
138	589	0	0	0	1
139	645	0	0	0	0
140	650	0	0	0	0
141	657	0	0	0	1
142	669	0	1	0	0
143	699	0	0	0	1
144	704	0	0	0	0
145	732	0	0	1	0
146	734	0	0	1	0
147	754	0	0	0	1
148	760	0	0	0	0

Figura 4-20: Tabla Binaria X.

este calculó se usó el tamaño de la tabla binaria $(n * i) = 38544$, donde n corresponde al número de individuos e i al número de variables. Por tanto, cada elemento dentro de la tabla de Burt, quedo dividido por $\frac{1}{38544}$ que es el peso de cada observación en el perfil fila.

Entonces, la suma de estas filas es igual a $\frac{1}{n}$ que es el peso de todas las observaciones en el perfil. De la misma forma, el peso para el perfil columna se calculó como $\frac{1}{nj.}$. Esto permitió calcular las frecuencias marginales fila y columna $f_{\langle j|i\rangle}$, las cuales vienen dadas por la siguiente ecuación:

$$f_{(i|j)} = \frac{f_{ij}}{f_{.j}}; f_{(j|i)} = \frac{f_{ji}}{f_i} \ \ para\ i = 1, \dots, n\ y\ j = 1, \dots, p. \qquad (4.25)$$

Donde f_{ij} son las frecuencias relativas.

Construidas las frecuencias marginales, se calcularon las distancias entre perfiles fila y columna mediante el uso de la ecuación Ji -Cuadrado χ^2. La interpretación de estas distancias muestra que valores pequeños entre perfiles equivalen a filas y o columnas similares y por ende variables o individuos homogéneos(Montenegro Alvaro, 2005). Esta distancia también muestra que si dos individuos (nodos) son cercanos en términos de distancia χ^2 entonces seguramente tienen en común las mismas categorías, que en nuestro caso son la aparición o no de ciertos tipos de error. Entonces, cada perfil fila representa la distribución de frecuencias con que cada tipo de error aparece en cada una de las variables analizadas (completitud, Exactitud posicional, temática). A continuación, se muestra la ecuación con las que se realizaron los cálculos.

$$d^2(i,'i) = \frac{1}{s}\sum_{j=1}^{p} \frac{n}{n_j}.\left(Z_{ij} - Z_{i'j}\right)^2 \quad (4.26)$$

Donde $\frac{f_{ij}}{f_{i+}}$ son las frecuencias para los perfiles fila/columna.

En la ecuación de distancia expuesta anteriormente s representa el número de variables (6), p representa el número de categorías (34), n es el número de individuos y n_j corresponde a la cantidad de individuos dentro de determinada modalidad. $Z_{ij}, Z_{i'j}$ son los valores que toman los perfiles que se esté comparando, estos valores varían entre 0 y 1.
A partir de las distancias entre perfiles calculadas con ayuda de R y la función inertia.dudi( ) se calculan los valores totales de inercia, para obtener los eigenvectores que construyeron cada uno de los ejes principales o factores, los cuales caracterizaran a cada

una de las categorías de las variables de la tabla binaria situándolas como coordenadas en el espacio geométrico(Lewis, n.d.). Las dimensiones "planos factoriales" que acumularon mayor varianza fueron escogidas para la generación del espacio geométrico que mejor representa la asociación entre perfiles.

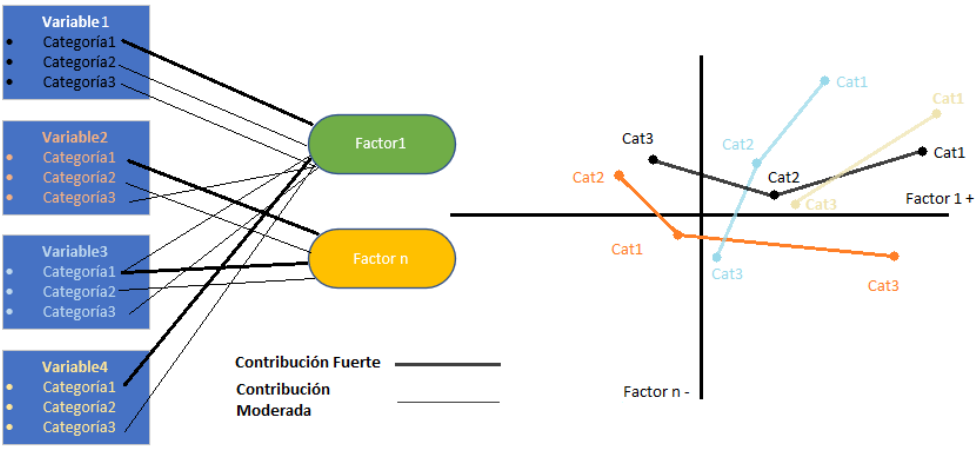


Figura 4-21: Construcción de planos factoriales

Se compararon los planos factoriales 1VS2, 2VS3, 3VS4 y 4VS5 para establecer cual combinación de estas brindaba una mejor representación. La primer dimensión o eje principal, produce un valor propio o eigen valor el cual representa la mayor parte de la inercia total y constituye el primer conjunto de coordenadas. La segunda dimensión, produce un nuevo eigenvalor que explica la mayor parte de la inercia residual y constituye el segundo conjunto de coordenadas. Después de escoger el mejor plano que representa las variables usando como ayuda del histograma de valores propios (**Figura** (4.22)) el cual muestra el porcentaje de

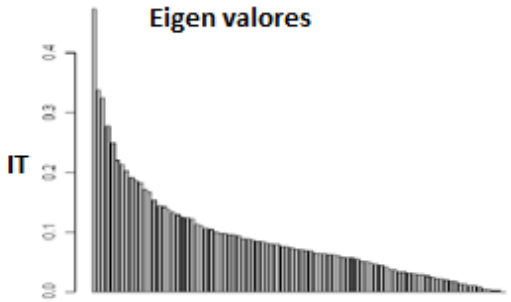


Figura 4-22: Histograma de eigen valores.

varianza acumulada (inercia total) para cada una de las dimensiones. Como la mejor representación de las variables y sus atributos se encuentra relacionada no solo a la distancia entre el centro coordenado y la variable, sino que también depende de los aportes individuales que hizo cada una de las variables para construir los ejes coordenados. En el **Anexo D**" se podrán consultar los aportes individuales que cada variable contribuyo para la creación de los ejes, como también se podrán consultar cada uno de los eigen valores generados.

Para fortalecer el análisis se trabajó con una variable de control la cual contenía la unión de algunos tipos de errores encontrados, por ejemplo, la etiqueta *3E_JE_SV_NV.* Dentro de esta variable artificial se indica que el individuo especifico contaba con la aparición de 3 errores, los cuales eran *JE:* error en la clasificación Jerárquica de la vía, *SV:* Error en la clasificación de sentido vial y *NV:* Error en la clasificación de Nombramiento de vía. Esto permitió identificar de manera fácil las relaciones entre los errores que mostraba el plano coordenado creado para representar las variables. En total se crearon 8 categorías para esta variable suplementaria. Las categorías agregadas se muestran en la (**Tabla** (4.8)) Esta nueva variable fue incorporada a la tabla usando la función ind.sup() del paquete FactoMineR. Se aclara que la variable suplementaria no hace parte de la construcción de los planos coordenados para la representación de las variables, por ello no influye en los cálculos de Inercia del sistema.

Tabla 4-8: Variable de control.

Variable de control	
1E_JER	Error en Jerarquía
1E_NV	Error en Nombramiento
1E_SV	Error en Sentido vial
2E_JER_NV	Error en Jerarquía y Nombramiento
2E_JER_SV	Error en Jerarquía y sentido vial
2E_SV_NV	Error en Sentido vial y nombramiento

3E_JE_SV_NV	Error en Jerarquía, Sentido vial y Nombramiento
N_E	Ningún Error

Finalmente se graficaron las variables proyectadas con el fin de determinar las relaciones entre los errores encontrados. Las variables fueron graficadas en los espacios de dimensión (1,2), (2,3), (1,3) y finalmente (16,17). Esto tuvo como finalidad ver la proyección de puntos (Coordenadas perfil columna y fila) para determinar numéricamente, a partir del mejor plano cartesiano aquel que representara mejor las variables (exactitud posicional, temática, exactitud posicional.) y sus sub categorías (Error encontrado o no encontrado) y así poder encontrar relaciones que no fuesen visibles fácilmente de manera marginal(Montenegro Alvaro, 2005).

Analizando las variables y los individuos proyectados en el espacio coordenado junto con la inercia calculada, se pudo establecer una super posición de variables que se encontraban aportando la misma información pues sus frecuencias aparecían en los mismos individuos. Si estas variables hubiesen sido una combinación de errores, entonces sería exactamente lo que se busca (agrupar errores según su frecuencia de aparición) pero en este caso eran otro tipo de variables, por ello y en base al análisis visual, dichas variables fueron excluidas de análisis. De esta forma se obtuvieron mejores resultados en cuanto al ACM.

Una vez analizados los gráficos ACM para encontrar las posibles relaciones de los errores y sus medidas de calidad VGI, se procedió a agrupar los resultados usando la metodología de conglomerados expuesta en la siguiente sección.

Agrupación y visualización de resultados por medio de análisis de conglomerados

El método seleccionado para agrupar los datos fue el jerárquico aglomerativo, este método se aplicó usando la función *hcpc().*Esta función requiere de los datos ACM, junto con una

medida de distancia y un método para aglomerar los datos. La medida de distancia es usada para ver qué tan cerca o lejos se encuentran los elementos testeados, en nuestro caso fueron los nodos centrales de línea. Estos individuos tienen coordenadas calculadas a partir del ACM. Pues como recuerda el lector, el ACM re-proyecta las variables objeto de estudio en un nuevo plano cartesiano.

Las distancias aquí testeadas fueron, distancia euclidiana, distancia máxima, Manhattan, Minkowsky y Jaccard(R, n.d.).Para encontrar la mejor clasificación se probaron varios métodos de aglomeración “vecino más cercano, vecino más lejano, método de Canberra, método de Ward”(Garcia, Juan; Segovia, 2014), de esta manera se pudo comparar varios tipos de aglomeración. Para la interpretación de los la formación de conglomerados se utilizaron los dendogramas permitiendo ver las distancias a las cuales se comenzaron a formar cada uno de los grupos. En el dendograma, la escala vertical representa la distancia y se selecciona el fin de la unión de los conglomerados a partir de ella.

Uno de los mejores métodos para la agrupación de datos utilizado fue el método de Ward. Este método utiliza la distancia intra grupal que cumple con el objetivo de buscar grupos que posean menos inercia dentro de dichos conjuntos como criterio de homogeneidad(Bridges, 1966). La idea del método es unir en cada paso de la clasificación los dos grupos que menos incrementen la inercia intergrupal, esto se traduce a menos dispersión en cuanto a la conglomeración se refiere. A continuación, se muestra la ecuación distancia de Ward entre dos grupos con pesos iguales.

$$W(i,j) = \frac{1}{2n} d^2(i,j) \quad (4.27)$$

Definido el algoritmo de agrupación y conociendo que medida permitirá agrupar los datos, se utiliza el dendograma para seleccionar los clústeres finales. La letra **H** en la figura 4.18, representa la altura a la cual los clústeres pueden ser cortados.

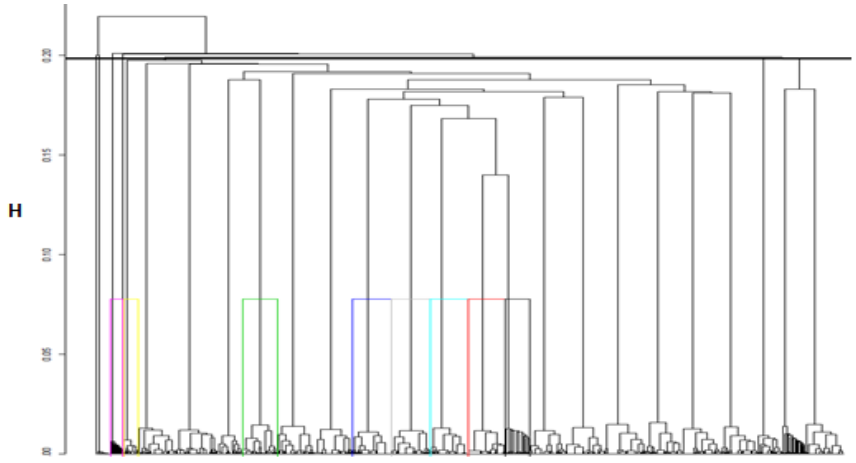


Figura 4-23: Dendograma para la selección de clústeres.

Para cada clúster formado se tienen las distancias a las cuales fueron creados, como también se conoce cuales nodos pertenecen a qué tipo de clúster. Este cálculo se realizó usando la siguiente línea de código:

```
res2.hcpc<-HCPC(res2 ,nb.clust=0,consol=FALSE,min=3,max=7,graph=TRUE)
res2.hcpc$data.clust[,ncol(res2.hcpc$data.clust),drop=F]
res2.hcpc$desc.var
res2.hcpc$desc.axes
res2.hcpc$desc.ind
```

Cada conglomerado contiene las posibles relaciones que existen entre los diferentes tipos de errores. Con estos datos se calculan las nuevas medidas de Calidad VGI (completitud, cxactitud posicional y exactitud temática). Cumpliendo así con el objetivo de este trabajo. La (**Figura** (4.24)) muestra el árbol jerárquico con la cantidad de clúster calculados, cada clúster tiene información de cuáles y cuantos nodos lo conforman. Esta información puede ser consultada en el **Anexo E**:" Clústeres"

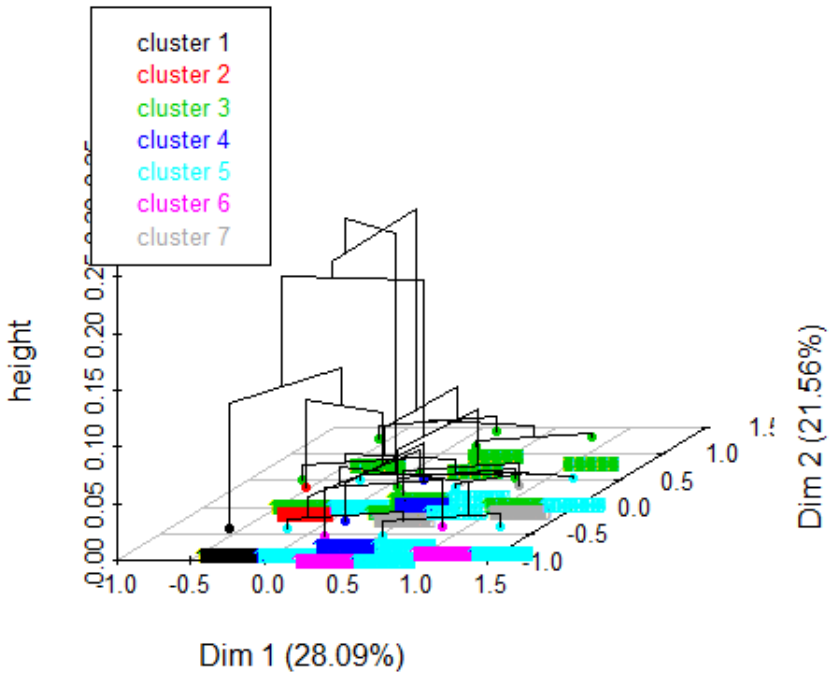


Figura 4-24: Método jerárquico para la creación de clústeres.

A diferencia de la evaluación de calidad VGI de manera univariada, donde se calculó la completitud de los datos, exactitud posicional y exactitud temática por separado, los clústeres calculados permitieron analizar estas medidas de calidad de manera conjunta, pues cada uno de ellos contenía la combinación de errores y las estadísticas de aparición de los nodos, inclusive mostrando la combinación de hasta dos y 3 tipos de errores dentro de un mismo clúster. Para poder ver y clasificar los resultados y conocer la calidad VGI se extrajeron los valores de cada uno de los clústeres generados. Y se realizó un conteo por porcentaje según la agrupación de errores asignando a cada localidad el porcentaje y conteo de error encontrado. Con esta información se crearon graficas para cada una de las localidades en donde se muestra la distribución de los errores encontrados.

Estos errores contemplan la exactitud temática y la exactitud posicional. Respecto a la medida de calidad completitud, esta no cambia pues se mantienen los valores encontrados de manera univariada, la razón es que la evaluación de la completitud de los datos que relaciona la ausencia o exceso de elementos es la primera información que se obtiene

pues viene dada por la comparación en sí y como el ACM no realiza de nuevo la comparación geométrica entre OSM e IDECA entonces no puede ser recalculada.

Tabla 4-9: Tabla ejemplo Extracción de datos clústeres por Localidad VGI.

Tabla ejemplo Extracción de datos Clústeres por Localidad								
Localidades	Clúster Error1	Clúster Error2	Clúster Error3	Clúster Error12	Clúster Error13	Clúster Error123	Clúster Error23	Clúster Sin error
1								
2								
3								
n								

A partir de la tabla obtenida se evaluó la calidad VGI mediante el enfoque multivariado contando los errores de acuerdo al tamaño de muestra calculado para cada una de las localidades. Los resultados multivariados por localidad provenientes de los clústeres se compararon con los resultados univariados con el fin de generar conclusiones más precisas. Con esta comparación, realizada a partir de gráficas y análisis finaliza la evaluación de calidad VGI mediante el enfoque multivariado para la malla vial de Bogotá.

5. Resultados

Para poder medir la calidad VGI de una forma multivariada se han desarrollado una serie de pasos que permitieron explorar preparar, mejorar, comparar y analizar la información OSM, a continuación, se muestra el orden en el cual los resultados serán expuestos. En la sección de resultados para el ACM muchas graficas no podrán ser totalmente claras debido al volumen y a la escala de las gráficas, por ello y ante cualquier duda, estas graficas serán enviadas como **Anexo G** *"Graficas ACM"* en un formato de hoja que permita disipar cualquier duda.

Los resultados son presentados bajo el siguiente esquema:

Para cumplir el objetivo específico 1:

- Extracción y exploración de los datos
- Estandarización de los datos por medio de análisis relacional.

Para cumplir el objetivo específico 2:

- Mejoramiento de los datos OSM por medio de expresiones regulares para poder comparar el nombre de las vías en las dos fuentes.

Para cumplir el objetivo específico 3

- Comparación semi automática de los datos para obtener los valores que permitirán evaluar la calidad (tablas que contienen el resultado de la comparación de atributos)
- Evaluación de completitud, exactitud posicional y exactitud temática.
- Muestreo y reclasificación de variables.

Para cumplir el objetivo específico 4

- aplicación del análisis multivariado.
- Agrupación y visualización de resultados por medio de análisis de conglomerados.

5.1 Extracción y Exploración de los datos

La extracción de los datos OSM fue realizada usando el siguiente código:

```
import os
import time
from urllib2 import urlopen


def osm_donwload():
    url = ('http://download.geofabrik.de/south-america/colombia-
    URL_OSM = urlopen(url)
    Nombre_Archivo=url[43:]

    print ('Download %s...' % Nombre_Archivo)
    # Archivo local
    fichero = open(Nombre_Archivo, "wb")
    # Fichero
    fichero.write(URL_OSM.read())
    # Cerrar ficheros
    fichero.close()
    URL_OSM .close()
    print ('%s descargado correctamente.' %Nombre_Archivo)
    Fecha_Hora = time.asctime(time.localtime(time.time()))
    print ('Fecha y hora de descarga:',Fecha_Hora)
if __name__ == "__osm_donwload__":
    osm_donwload()
```

Figura 5-1: Extracción de datos OSM.

Como resultado se descargaron los datos OSM colombia-latest-free. Los datos OSM obtenidos se ajustaron al área de estudio de acuerdo al perímetro oficial de Bogotá(IDECA,2013). El área de trabajo estuvo compuesta por 20 localidades incluyendo la sección rural de Bogotá. Una parte de los datos crudos puede ser observada en la **Figura** (5.9).

Respecto a la exploración preliminar de los datos se analizaron los siguientes campos: Numero de registros, nombre principal de las vías, dirección de flujo y jerarquía vial. Se concluyó que OSM tenía 73.454 segmentos viales mientras que IDECA contaba con 136.958. Esta diferencia dependía en gran medida en que OSM no contaba con una división por cruces viales. En la (**Figura** (5.2)) la línea amarilla hace referencia a un segmento digitalizado en OSM, el cual no se encontraba segmentado por la intersección con otras líneas, mientras que en los datos IDECA existan para este mismo elemento 4 segmentos diferentes. Por esta razón se tuvo que hacer una operación de Split (Nelli, 2015)a los datos OSM. Asegurando que la comparación entre nodos centrales y atributos de cada línea fuera más efectiva pues se aseguraba que un nodo central pudiese encontrar a su homologo. En la figura de la derecha se puede apreciar parte de la geometría OSM, en color rojo con los segmentos fragmentados por cada nodo inicial y final de línea.

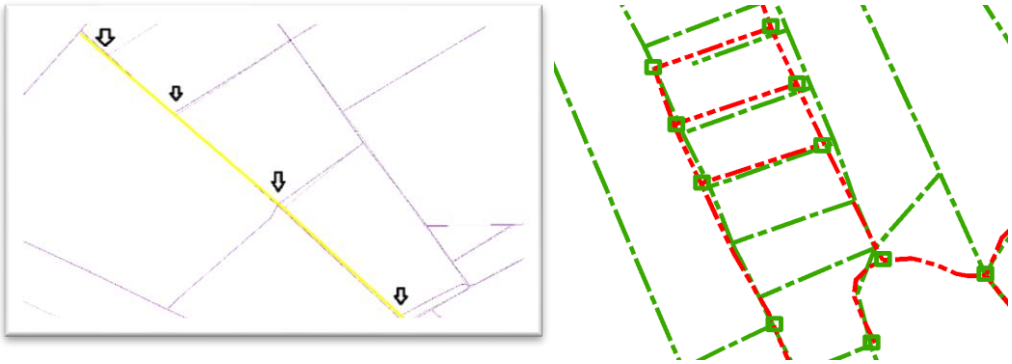


Figura 5-2: Representación geométrica OSM- IDECA.

Por otra parte, se encontró que 28.29% de los registros OSM no poseían nomenclatura vial asignada, mientras que solo el 5.3% de los datos IDECA se encontraban en la misma condición. Aparentemente los datos OSM se encontraban menos completos en cuanto a nombramiento. Sin embargo, más adelante se pudo observar que muchos de estos segmentos viales OSM si tenían datos solo que no habían sido asignados correctamente debido al Split de geometría realizado en la preparación de los datos. Los datos "*tipo de vías"* expuestos en la siguiente figura tienen nomenclaturas diferentes debido a que pertenecen a fuentes diferentes y aún no han sido estandarizados.

Respecto al sentido vial se encontró que IDECA tenía 5 clases definidas para este atributo: *B - doble sentido, FT- desde el sentido de digitalización - sin sentido vial, SD-Sin definir, TF hacia el sentido de digitalización*.

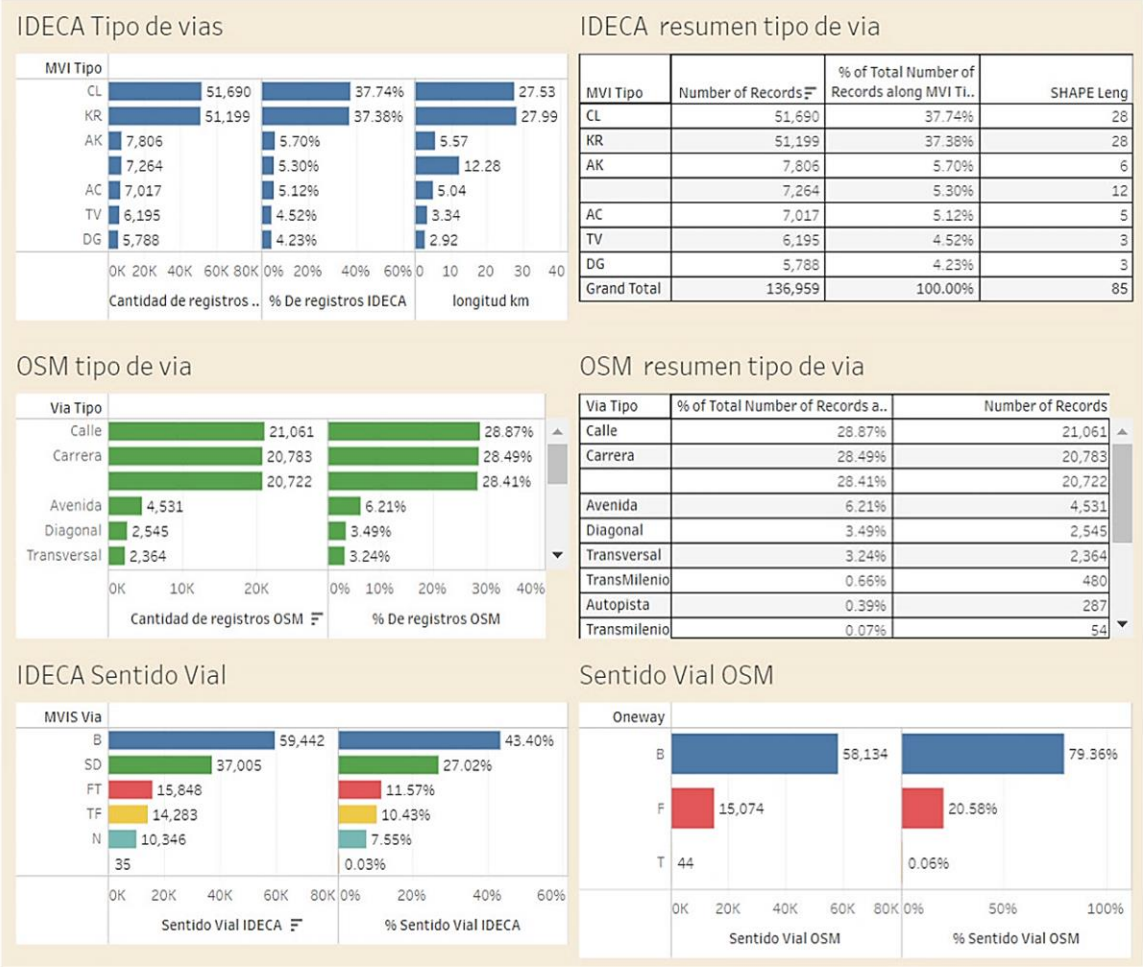


Figura 5-3: Análisis exploratorio de datos IDECA VS OSM.

Donde el 43.4% de los datos hacían referencia a el sentido vial *B* y el 27% se encontraba sin ninguna definición a cerca del sentido de flujo SD. En lo que respecta a OSM el sentido vial estaba determinado bajo 3 categorías *B, F, T* donde el 79.36% de los datos estaban registrados con sentido *B*. En la exploración de datos se pudo observar que la asignación de dirección de flujo para IDECA había sido creada desde el sentido de digitalización, sin embargo, algunos errores producto de esa interpretación pudieron afectar los resultados finales. En la (**Figura** (5.4)) se muestra un resumen de lo expuesto anteriormente:

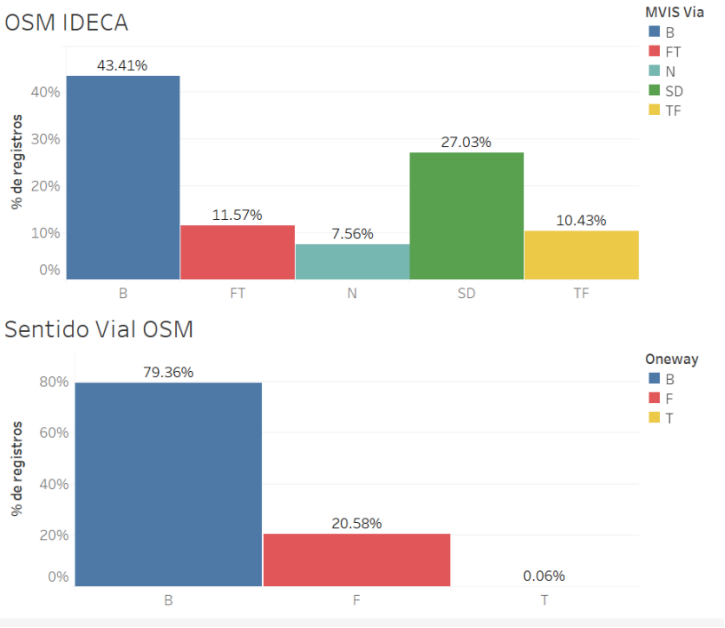


Figura 5-4: Clasesy porcentaje de sentido vial encontrado.

Respecto a la jerarquía vial (elemento que categoriza la red vial de acuerdo a sus funciones arteriales) se encontró que OSM contaba con una clasificación de 26 categorías, donde el tipo residencial predominaba con 36980 registros. Seguidos de la categoría *Footway* con 8398 registros. Por otro lado, la jerarquía vial IDECA solo contaba con 7 categorías. Esto finalmente repercutió en la necesidad de crear un estándar para aquellos atributos que podían pertenecerá a más de una clase. En la (**Figura** (5.5)) se puede apreciar simbólicamente la diferencia entre categorías viales, el tamaño de los círculos indica cantidad datos.

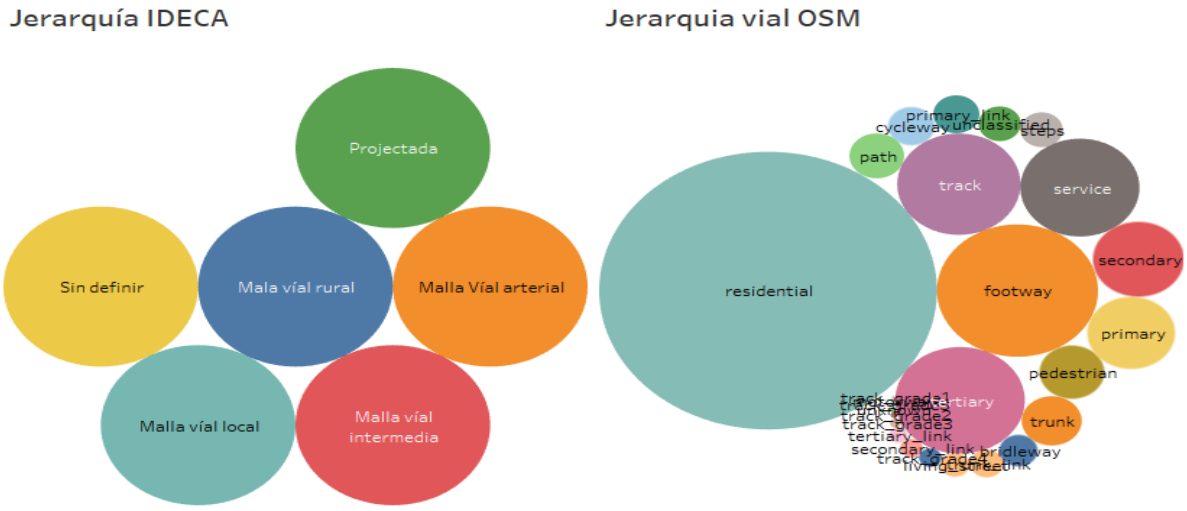


Figura 5-5: Jerarquía malla vial IDECA vs OSM.

Para entender mejor los datos encontrados, se usaron los elementos de estadísticas descriptivas básica para el conteo de geometría, y los demás atributos explorados. Encontrando que en promedio existen 2400 elementos con sentido vial en OSM con un mínimo de 44 registros en una clase y un máximo de 58134 para otra. Respecto al tipo de vía los números encontrados fueron, 21061 datos con tipo de vía asignado, estos corresponden al nombramiento calle, en promedio 2432 datos tienen algún tipo de nombramiento asignado. Finalmente, las estadísticas básicas para Jerarquía vial mostraron que existe una clase que contiene 36980 registros, en contrapartida, existe una clasificación que solo posee 4 registros dentro de la base de datos. Clases con pocos registros fueron candidatas a fusionarse con una clase más grande. En promedio la Jerarquía vial tiene 2817 elementos asignados por clase. Algunas estadísticas descriptivas son mostradas a continuación:

Summary	
Count:	3
SUM(Number of Records)	
Sum:	73,252
Average:	24,417.33
Minimum:	44
Maximum:	58,134
Median:	15,074.00

Summary	
Count:	30
Measure Values	
Sum:	72,946
Average:	2,432
Minimum:	0
Maximum:	21,061
Median:	6

Summary	
Count:	26
SUM(Number of Records)	
Sum:	73,252
Average:	2,817.38
Minimum:	4
Maximum:	36,980
Median:	593.50

Figura 5-6: Estadísticas básicas.

EL último paso para finalizar la exploración de datos consistió en revisar y analizar la guía para mapeo de OSM. Como resultados se encontró que para la clase *highway de OSM*, la clasificación *motorway* no debe ser codificada en Colombia, ya que OSM estableció que estas vías no existen en el país. Esto efectivamente es correcto pues no contamos con vías de acceso controlado(Alcaldía Mayor de bogotá, 2018), por lo que toda codificación encontrada con esos valores fue contada como un error. También se encontró que las carreteras nacionales troncales son asignadas a la etiqueta highway=trunk, según OSM las vías NQS, Calle 80, Autopista Norte, Avenida el dorado representan esta categoría lo cual no es totalmente cierto sí se examina y compra con la clasificación vial que tiene IDECA(C et al., 2012). En términos generales se encontró que OSM aún tiene problemas en la documentación para la codificación de datos de jerarquía vial, tal como afirma(Loai Ali, 2016) OSM aún maneja términos ambiguos y poco claros a la hora de realizar recomendaciones de codificación a los usuarios.

Los siguientes atributos OSM fueron excluidos del estudio debido a que no tenían un elemento de correspondencia en los datos de la malla vial IDECA: *('bridleway', 'cycleway', 'footway', 'pedestrian', 'path' , 'steps')")*

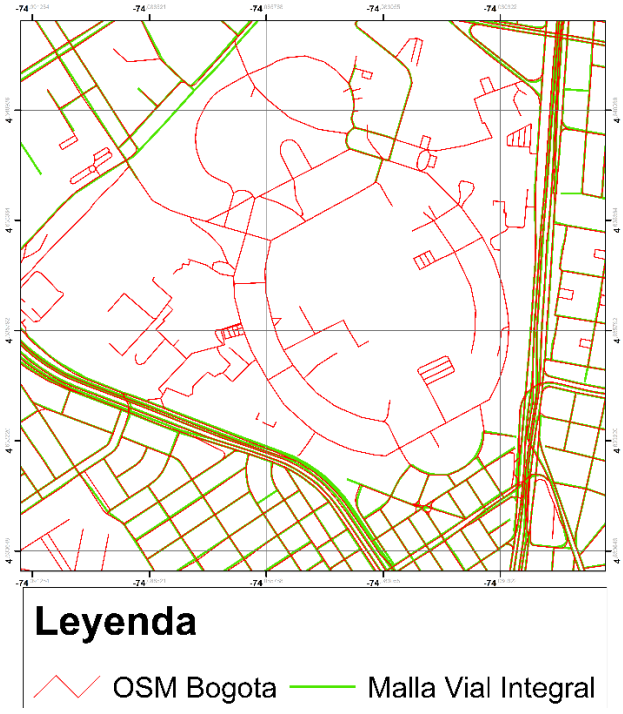


Figura 5-7: Líneas peatonales IDECA vs OSM

Esto se debe a que IDECA ha separado la información de red de Bici usuarios y la de peatones en otras capas de datos. Es decir, la malla vial integral de IDECA solo contempla vías navegables por vehículos.

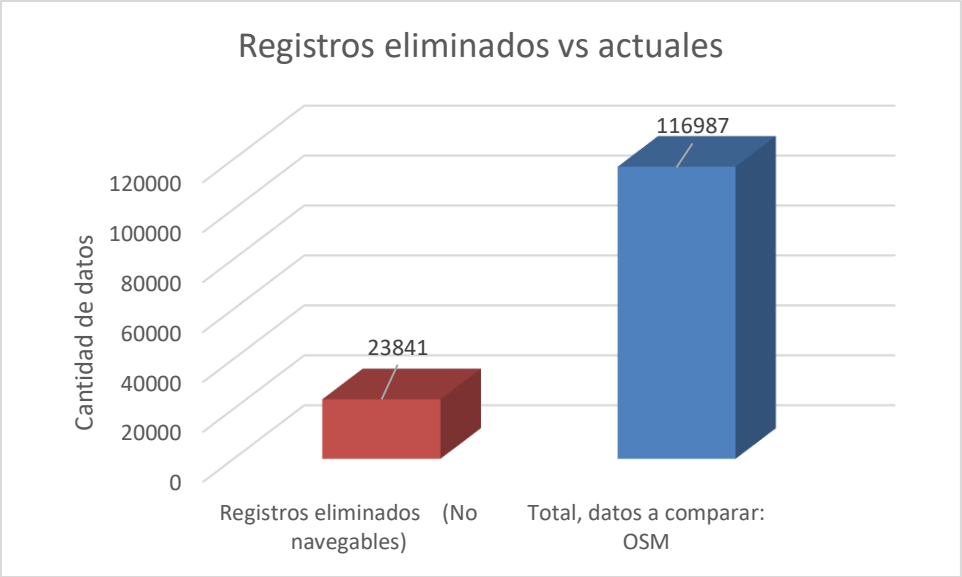


Figura 5-8: Datos eliminados OSM

Estos datos fueron removidos usando la función *arcpy.Select_analysis()* de la librería *arcpy*. Haciendo que este proceso sea siempre automático y repetible. Todos los demás atributos referentes a clasificación vial se establecieron como comparables si se realizaba una estandarización. Respecto al nombramiento de las vías y la codificación de dirección de flujo NO se encontraron restricciones o reglas que impidieran la comparación entre los datos.

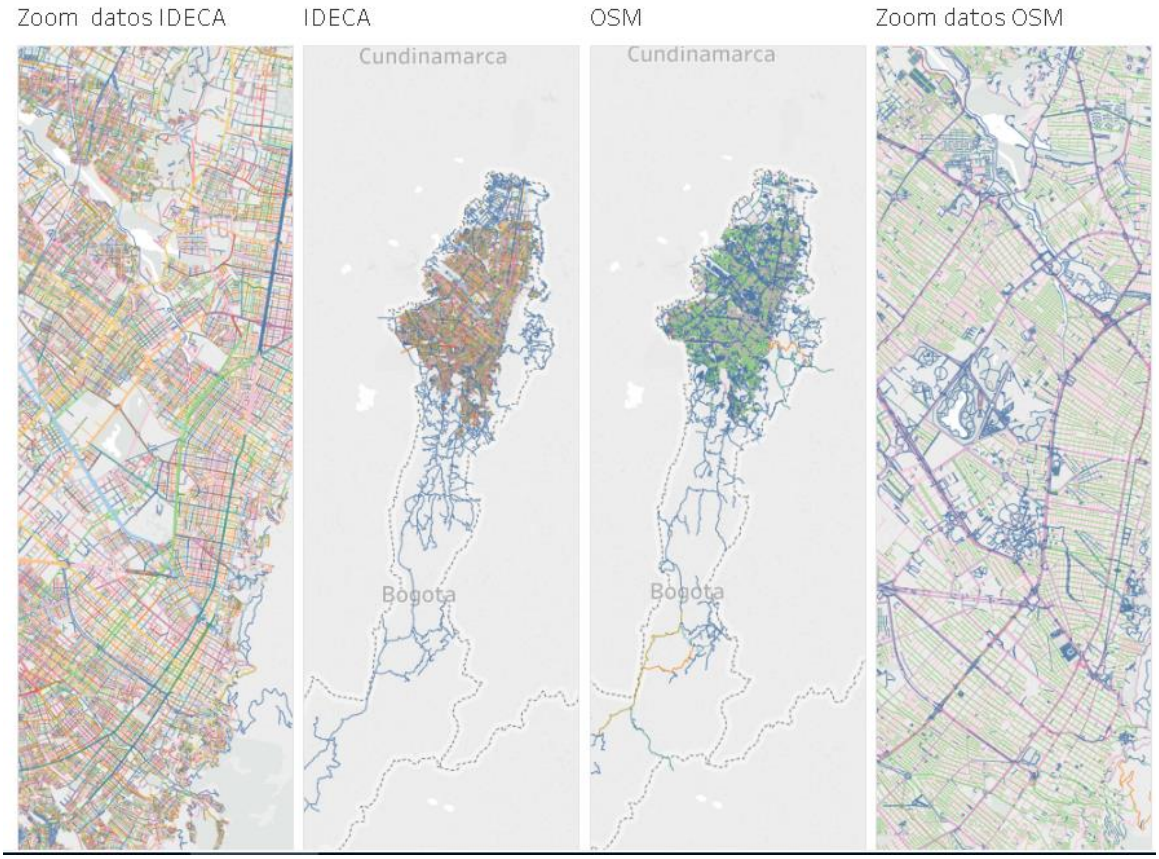


Figura 5-9: Información publicada en la web sin procesar IDECA-OSM.

5.2 Estandarización de los datos

Como se había mencionado previamente, para poder comparar los datos era necesario estandarizarlos, esta estandarización dio como resultado un modelo entidad relación que permitió entender qué tipo de variables y que clases debían ser transformadas de manera similar a como fue realizado por (Nowak Da Costa, 2016b) en *Towards Building Data Semantic Similarity Analysis:OpenStreetMap and the Polish Database of Topographic Objects pag. 273*. Aquí también se propuso una alteración en la clasificación de datos mediante la fusión de algunas de las calases existentes para minimizar y optimizar la comparación entre las fuentes de datos obteniendo buenos resultados.

Como resultado se muestra el modelo generado donde se ilustra las relaciones entre tablas de atributos.

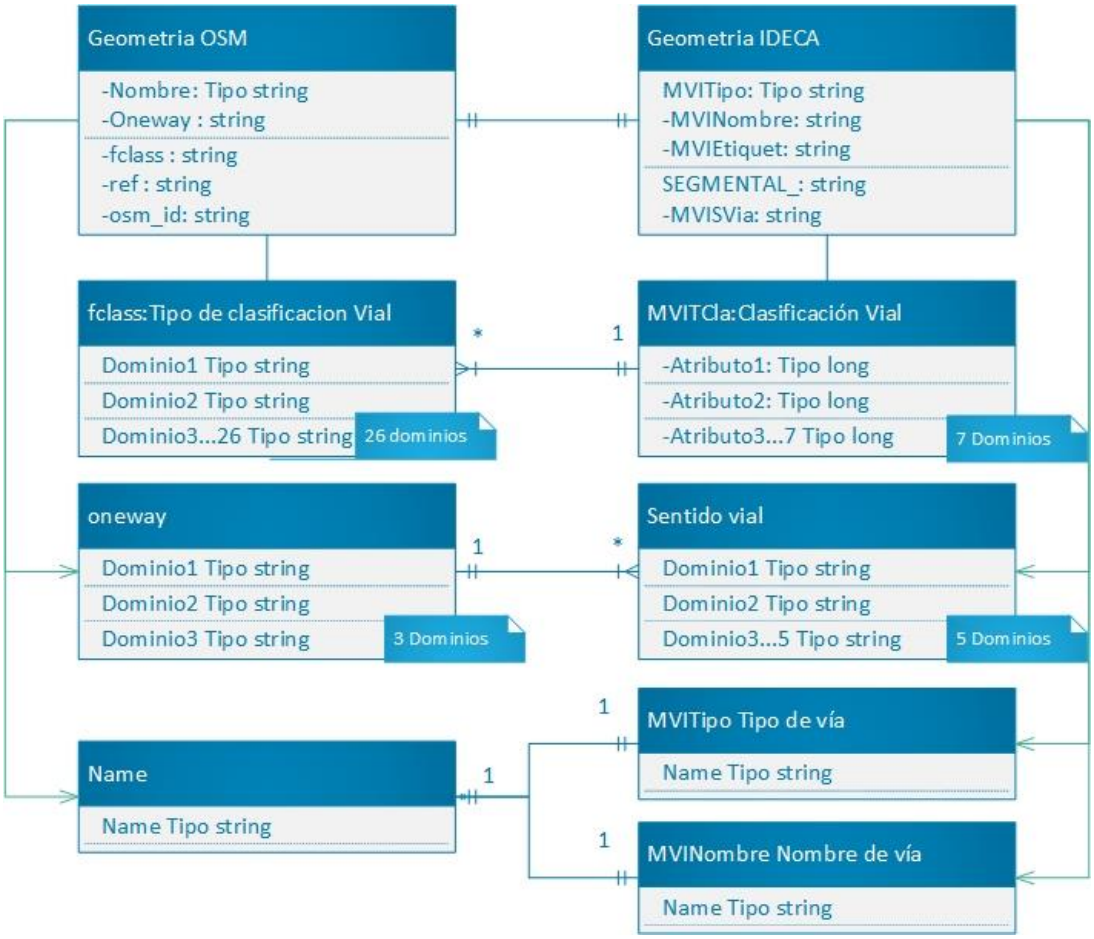


Figura 5-10: Modelo entidad relación.

Este modelo permitió entender que la geometría debía tener una relación uno a uno, es decir, en lo posible un único link o centro de nodo en *OSM* debería tener único link o centro de nodo en *IDECA*. Obviamente esto no se cumplió completamente pues se tuvieron algunos problemas relacionados a geometrías complejas cuando se aplicó el algoritmo de comparación. En algunos casos, el algoritmo no supo que elementos indexar dando como resultado una comparación aleatoria que dependía de la vecindad generando resultados imprecisos. Este tipo de inconvenientes hicieron que autores como Joanna Nowak Da Costa, Ehsan Abdolmajidi, Ali Mansourian entre otros, investiguen mejoras formas de matching geoespacial usando indexación. Inclusive el algoritmo trabajado aquí, el cual se basa en la selección de nodos, comparte características similares al algoritmo llamado *Extended node-based matching algorithm* aplicado en Matching authority and VGI road networks using and extended node-based matching algorithm con la diferencia de que el

desarrollado en este trabajo usa un buffer móvil mientras que del autor Abdolmajidi hace uso de topología.

Continuando con la clasificación vial, se puede ver en la figura previa que muchas clases en *OSM* tenían una única clase de destino en *IDECA* sin embargo también se puede evidenciar un problema, las variables en *IDECA* eran de tipo numérico (*long)* mientras que las de OSM eran de tipo *string*. Esto obligó a transformar los dominios numéricos en *IDECA* para luego poder reclasificar los datos en OSM.
Observando el modelo respecto a la tabla de sentido vial, ocurrió que *OSM* tenía una clasificación más breve para describir la dirección de flujo, por ello se estableció que muchos tipos de sentido vial en *IDECA* podían corresponder a una única clase en OSM. Finalmente, el nombramiento de las vías trato de mantenerse 1:1, pues para términos de este estudio Una vía solo podía tener un UNICO nombre y un elemento en OSM debía corresponder únicamente a un elemento nombrado en *IDECA*. Nótese que en *IDECA* el nombramiento de las vías se encontraba separado mientras que en OSM existía solo una clase.

Basado en el modelo entidad relación planteado y la exploración de los datos realizada en el numeral anterior, se construyó la siguiente estructura (ver **Tabla** (5.1)) siguiendo los lineamientos de(Nowak Da Costa, 2016b), donde cada atributo analizado contenía las clases que la componen y el tipo de relación que se creó. Por ejemplo, para jerarquía vial *OSM* existían 3 categorías que pasaron a ser solo una, esta nueva categoría asumió el nombre de Malla vial arterial, categoría que pertenece a *IDECA* y que engloba las características de las 3 clases que antes contenía OSM. De la misma forma, se puede observar que para la clasificación de la Malla vial intermedia existían muchas clases en OSM, las cuales fueron reemplazadas por una sola. Así se procedió con cada uno de los atributos objeto de estudio los cuales son (jerarquía vial, sentido vial y nombre de las vías). La tabla 5.1 contiene todas las relaciones creadas entre los campos y permite entender qué tipo de relación se tuvo presente para la homologación de datos.

Después de tener la estructura de relaciones entre campos clara, se crearon las columnas que alojaron los resultados de la estandarización y también se crearon los campos que

contendrían los resultados de la comparación: estos campos alojarían valores binarios que representaran la ausencia o no de error.

Después de tener el modelo se materializo esta estandarización, este resultado se muestra en la (**Tabla** (5.1))

Tabla 5-1:Resultado Modelo entidad relación.

Jerarquía Vial OSM	IDECA	Relación
Primary	Malla vial arterial	Muchos a uno
Primary link		
Trunck		
Secundary	Malla vial intermedia	Muchos a uno
Secundary Link		
Terciaria		
Terciary link		
Trunck link		
Residencial	Malla vial local	Muchos a uno
living street		
track 1…5	Malla vial rural	
Unclasified	Sin definir	Muchos a uno
Unknow		
Services		
Sentido vial OSM	**IDECA**	**Relación**
B	B	Uno a Muchos
	N	
	SD	
	Null	
F	FT	Uno a Uno
T	TF	
Nombre de vía OSM	**IDECA**	**Relación**
Name	name	Uno a uno

Los campos creados para los datos estandarizados fueron:
New_Jerc: Nueva jerarquía vial OSM de 26 categorías a 7.
Tipo_Viaos: Nombre de las vías estandarizadas *OSM* (campo vacío, paso siguiente en el proceso).
Dir_ViaIDE: Estandarización sentido vial en *IDECA*: 5 categorías a 3 categorías.
New_Jerc: Jerarquía vial: Cambio de dominio, ahora la jerarquía vial se rige por números del 1-7.

Campos creados para guardar los resultados de la comparación:

NEAR_DIST (Distancias m): Campo para calcular las distancias a los centroides de cada objeto (Completitud).
Exac_po: Campo para alojar el cálculo de exactitud posicional.
Ch_New_Jer: Campo booleano para clasificar la comparación referente a Jerarquía Vial.
Ch_Sent_F_: Campo booleano para clasificar la comparación referente a Sentido Vial
Fc_Name_: Campo booleano para clasificar la comparación referente a Nombre de vía principal.

5.2.1 Estandarización de cadenas de caracteres

Muchos de los datos alojados en OSM referentes al nombramiento de las vías tenían problemas asociados con espacios, texto incompleto etc., esto puede ser causado por múltiples factores, entre ellos el descuido de los usuarios, la diversificación con que fueron colectados los datos o incluso ediciones deliberadamente codificadas de manera errónea para afectar las bases de datos (Vandalismo informático) OSM (Neis & Zielstra, 2014). Para no perder esta información y lograr una mejor comparación se usaron expresiones regulares donde se seleccionan y se estandarizar los datos por medio de patrones. Trabajos similares fueron desarrollados por (Ballatore et al., 2013) donde el lugar de texto se trabajaron etiquetas de elementos geográficos para reclasificar elementos de tipo geográfico ayudando a reducir el ruido y la ambigüedad de la información OSM.

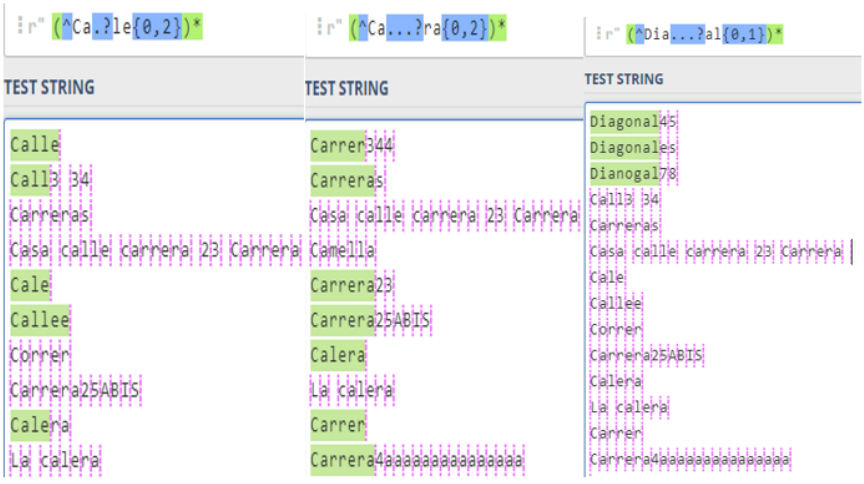
En la **Tabla** (5.2) se puede observar una sección de la estandarización de nombre en vía principal, realizada sobre OSM. 85.125 registros fueron estandarizados. Se encontraron cadenas de caracteres que no pudieron ser tratadas debido a que no se encontró un patrón asociado. Algunas expresiones que tuvieron que corregirse manualmente fueron aquellas que iniciaban con número, pero no tenían asociado ningún tipo de nomenclatura como calle o carrera o cadenas de caracteres mayores a 20 registros. Esta es una limitante de este trabajo pues únicamente se pudieron trabajar las cadenas de texto más simples y no tan extensas.

El patrón para todos los tipos de nombramiento relacionados a Calle, Carrera, Avenida, Diagonal, trasversal fueron encontrados y son expuestos a continuación:

Patrón para la palabra Calle-Carrera-Diagonal y trasversal:

- ^([0-9]+)$ -^(Calle|calle|Cl{2}|cl|AvCalle|.venida .alle)
- ^(Calle|calle|Cl{2}|cl|AvCalle|.venida .alle)
- [Ca.le]+$: (^Ca.?le{0,1})*
- [Ca…ra]+$: (^Ca..?ra{0,2})*
- ([^e]+[0-9]$)*
- ^(Carrera|carrera|Cl{2}|cr|AvCarrera|.venida .era)
- ^ ([0-9]+)$ -^(Carrera|crrera|CR{2}|cl|AvCarrera|.venida .arrera)
- ^(Carrera|carrera|KR{2}|kr|AvCarrera|.venida .arrera)
- [Ca.le]+$: (^Ca.?le{0,1})*
- (^Dia…?al{0,1})*
- (^Tran…?al{0,2})*

Todos los patrones creados y usados podrán ser consultados en el **Anexo A**.



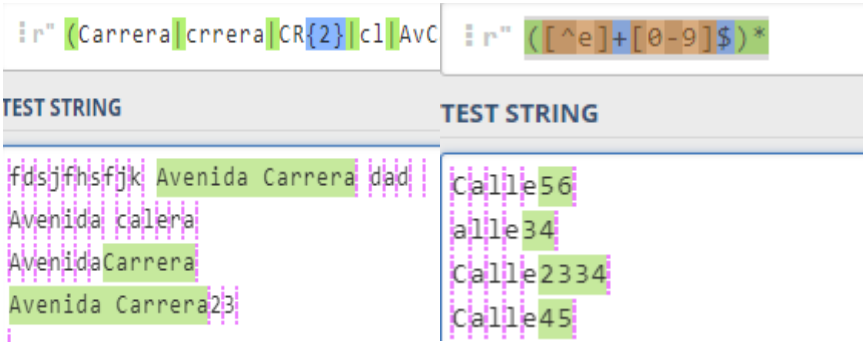


Figura 5-11: Patrones para el mejoramiento de texto.

Estos patrones permitieron extraer de los datos originales de estas cadenas sin importar el lugar donde estuviesen. Los patrones tienen como regla que el nombre buscado o cualquiera de sus posibles combinaciones (Cale, Cll, Cl, Call,) tuviese un numero contiguo, pero nunca un numero antes de esta cadena, ya que, si ocurría esto, posiblemente el nombre buscado se establecía como el nombre compuesto de la vía, por ejemplo: Carrera 23 Cll 5.

Algunos resultados producto de esta estandarización se pueden observar en la **Tabla** (5.2)

Tabla 5-2: Estandarización de nombres.

Estandarización de nombres	
Original	Estándar
AveniA CArr000012	KR 12
Transv 9 Bis Este	TV 9BISE
eTransversal 9 Bis Este	TV 9BISE
Transversal 9B Este	TV 9B E
Variante Bosa San José - Travsnersal 80I	TV 80I
Akenida Agoberto Mejía - Transversal 80G	TV 80G
Calle 32 Callw33	CL 32
Transversal 7 Bis Este	TV 7BISE
Avenida Tintal - Carrera 89	KR 89
Carrera 78H Bis Sur	KR 78HBIS S
Carrera 53F	KR 53F
Calera 15	KR 15

Calle 45/sur	CL 45 S
TransMilenio - Patio Calle 80	CL 80
TransMilenio - Intercambiador CL 6	CL 6
CL-57B	CL 57B
AveniA CALLE 00001	CL 1

Finalmente se corrigieron y estandarizaron todos los registros de OSM bajo la siguiente nomenclatura (KR #, CL #, TV #, DG #...etc.). Todos los datos estandarizados cumplieron las siguientes reglas: (I) No pueden existir dos tipos de vía en una misma etiqueta: Calle 34; calle 23. (II) No pueden existir espacios al inicio, (III) No puede existir caracteres especiales = / # -$. (IV) No puede existir ninguna combinación de Calle, Carrera Diagonal que no terminé en vocal. (V) Nomenclatura y nombres deben estar separados para su comparación. (VI) Bis debe ir en mayúsculas y sin espacio entre el número: 4 Bis: 4BIS. Los datos estandarizados se copiaron automáticamente en la columna creada para alojar esta estandarización.

5.3 Comparación semiautomática de los datos para obtener las medidas de calidad: Exactitud posicional, Completitud y exactitud temática

Por medio del matching al objeto más cercano, el cual consistía en un buffer móvil que dependía de la distancia al nodo más próximo, se lograron comparar 114183, registros, lo que significó un match de 85.42%. Muchos nodos no se encontraron cerca de sus homólogos, pues alguna geometría en *OSM* tenía desplazamientos. Solo el 14.58% de los datos no pudo ser unido debido a problemas con geometría. Comparando este pareo semi automático con el de otros autores como(Zielstra & Zipf, 2010), aparentemente el método empleado permite obtener resultados más precisos pues Ziliestra realizaba la comparación de km usando una grilla con un determinado tamaño mientras que aquí la comparación se realizó uno a uno con un porcentaje de fracaso relativamente bajo (Aprox 15%). Por otro lado, comparando este trabajo con el de autores que realizaron pareo automático como lo son (Steffen Volz, 2006) y (Abdolmajidi 2015) se puede observar que el método empleado

aquí es de menor calidad que el del autor previamente citado pues aún no se superan los errores de unión en geometrías complejas, sin embargo es mucho más rápido en términos de procesamiento y es de mejor calidad que la metodología de búfer estático usada con frecuencia. Este tipo de errores aparecen sobre aquellos elementos complejos como deprimidos, glorietas y en términos generales sobre vías con geometría complejas. Este tipo de problemas ya habían sido reportados en la literatura(Michael F. Goodchild & Hunter, 1997). El problema del match de geometría usando el nodo central radico en que no todos los nodos pudieron ser localizados cerca a su par.

Figura 5-12: Nodos OSM-IDECA.

En las siguientes figuras se puede observar en color negro los nodos que lograron hacer match y permitieron la comparación de los atributos objeto de estudio **Figuras** (5.13), en color verde se puede observar la geometría IDECA y en color rojo OSM. Nótese que, para esta región el match fue muy bueno en las vías principales mientras que en algunas vías internas no se consiguieron resultados tan precisos.

En la **figura** (5.13), en el cuadro color azul, se muestra una coincidencia perfecta, pero tiene una particularidad. El algoritmo no evalúa topología. Aunque en este ejemplo en particular se pudo comparar el nodo central con su homónimo, esto no quiso decir que la topología este correcta. Los cuadros color naranja muestran dos estados, primero resaltando que OSM tenía geometría que IDECA no, y que por ende no podía tener ningún

match y segundo, muestra buenos resultados en términos de comparación para segmentos de línea cortos.

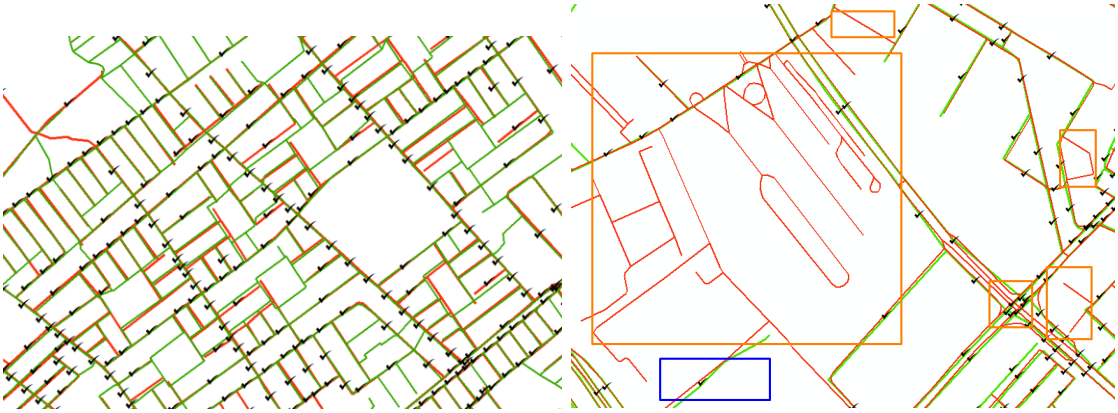


Figura 5-13: Proceso de comparación usando nodos centrales

En ALGUNAS geometrías complejas donde no había superposición de nodos la comparación fue exitosa:

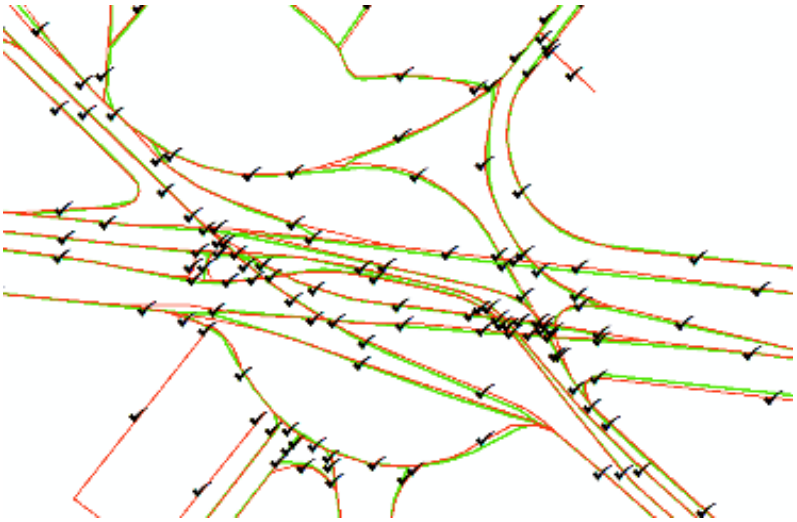


Figura 5-14: Proceso de comparación automática de datos.

A continuación, se muestran algunos ejemplos donde no fue posible generar el pareo de datos.

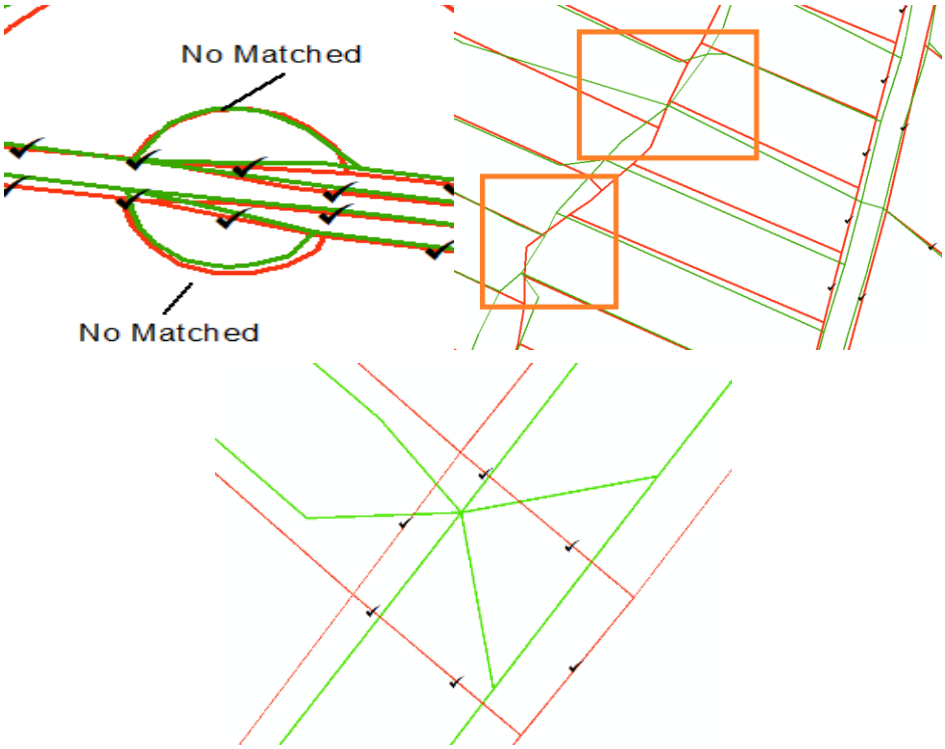


Figura 5-15: Problemas con el match en algunas geometrías.

Los problemas en el match como los mostrados en la figura 5.15 NO pudieron ser resueltos por ello sabemos que al menos 14.58 % de la información se perdió y no pudo ser analizada. Este dato es muy importante pues el 14.58% de los datos tubo que se comparado manualmente.

En términos de rendimiento, se invirtieron 1.5 horas para realizar el mach geométrico y 30 minutos por atributo para realizar la comparación de campos, lo que significó un total de 3 horas de procesamiento, otros algoritmos invierten hasta 5 hora en la ejecución de esta tarea, como lo reporta (Abdolmajidi et al., 2015).

5.3.1 Completitud

La completitud se evaluó en términos generales, sobre toda la ciudad de Bogotá y particularmente sobre cada una de las localidades que la conforman. Como se puede observar en la **Tabla** (5.3), se encontró que *OSM* había omitido 1183,29 Km, lo que equivale a un 12,6% menos respecto a los *Km* reportados por *IDECA*. Por otra parte, el conteo derivado de la comparación automática, mostró que *OSM* contiene -14.6% elementos lineales que los encontrados en *IDECA*.

Tabla 5-3: Tabla datos Omisión.

Resultados Completitud nivel Bogotá				
Variables	IDECA	OSM	Delta	% Omisión
KM	9379,21	8195,91	1183,29	12,6%
Elementos	136958	116964	199,94	14,6%

Analizando los datos por comisión se encontró un exceso en valor de 36.42 *Km* correspondientes a 1266 objetos lineales en OSM. En términos de porcentaje esto correspondió respectivamente al 0,44 y 1,11%. El resumen de los valores de comisión es expuesto en la **Tabla** (5.4). Se resalta que el valor real de datos contenidos en OSM fue de 114183, debido a que se le eliminaron los valores de comisión.

Tabla 5-4: Tabla datos Comisión.

Resultados Completitud nivel Bogotá				
Variables	IDECA	OSM	Delta	% Comisión
KM	9379,21	8195,91	36.42	0,44%
Elementos	136958	116964	1266	1,11%

Por otro parte, evaluando los resultados de completitud por localidad **Figura** (5.16) se encontró que las localidades en color azul contienen los valores más altos en cuanto a falta de segmentos en la malla vial de *OSM* (3205-5978). En contra parte, las localidades

(achuradas) en la imagen 5.16 “segunda figura a la izquierda”, hacen referencia a la presencia de exceso de segmentos viales (776). Por último, la imagen situada a la derecha hace referencia a cantidad de Km por cada localidad, donde se puede observar que la zona rural ubicada al sur de la ciudad posee la mayor cantidad de Km cuantificados por exceso (-19 *Km*) mostrando un comportamiento diferente al comúnmente encontrado en zonas rurales(Michael F. Goodchild & Glennon, 2010). Las localidades de tono rojo, contienen la mayor cantidad de kilómetros faltantes (252-306 km*).*

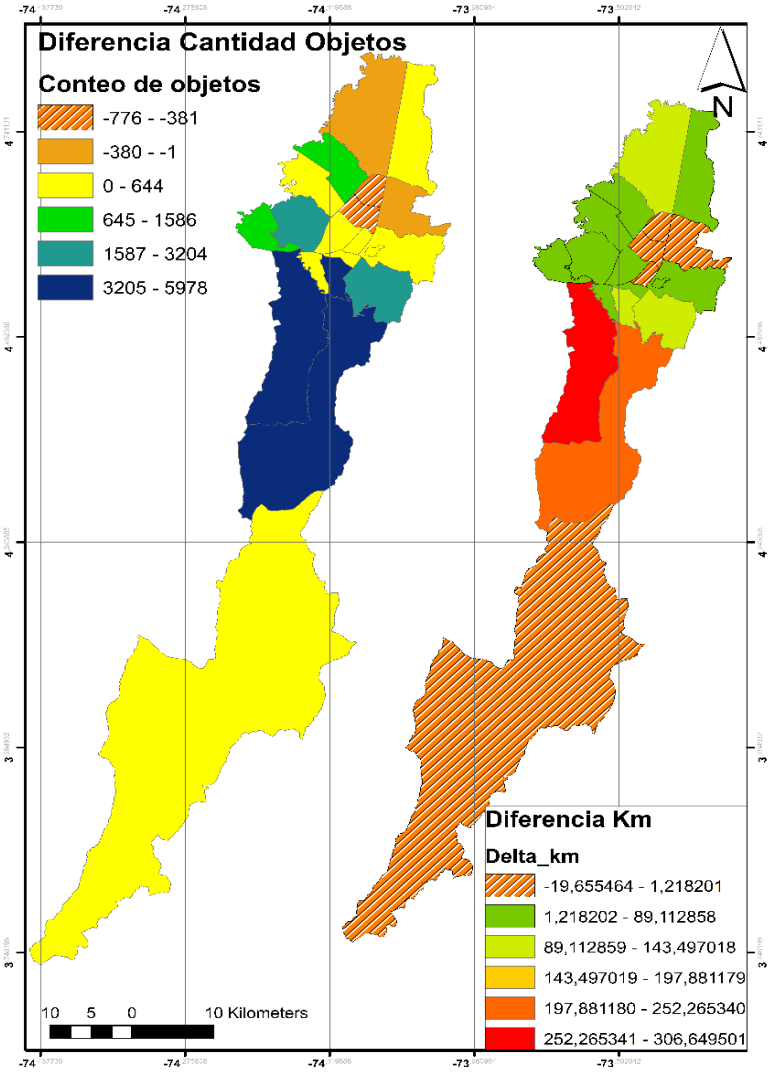


Figura 5-16: Completitud por localidad.

5.3.2 Exactitud posicional

El resultado de la sumatoria de los errores medios cuadráticos encontrados fue: $RMSr = 2{,}252$. Los valores extremos al evaluar la distancia entre nodos relacionados tuvieron un valor mínimo encontrado de: 0,008162 metros y un valor máximo de: 5,00 metros. El factor usado para computar la exactitud posicional con un intervalo de confianza del 95% fue 1,7308, considerando que los errores $RMSx$, $RMSy$ tuvieran un comportamiento similar en su componente vertical y horizontal. De este modo, el error posicional encontrado fue: Error posicional horizontal (95%) =3.98metros.

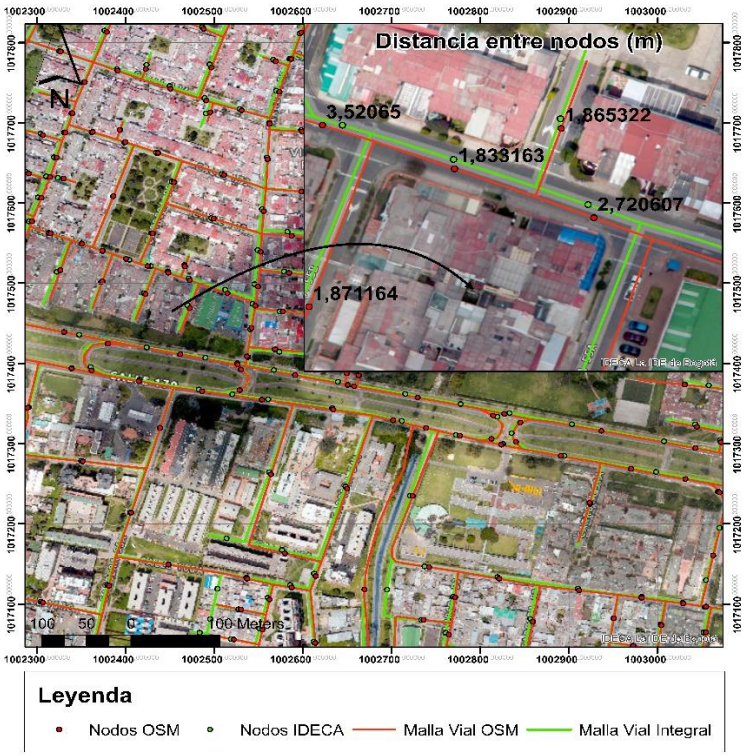


Figura 5-17: Muestra de distancias calculadas.

En la **Figura** (5.17) se puede apreciar una muestra de las distancias calculadas a los nodos relacionados. En términos generales no se encontraron diferencias de distancia superiores a los 5 metros.

Comparando el resultado obtenido (3.98m) con una investigación realizada en Inglaterra donde se obtuvo un error aproximado de 7.9 metros (Antoniou, 2011) y se concluyó que esta exactitud permitía comenzar a pensar que los datos OSM podrían ser usados por algunas agencias en su país lleva a plantear la misma inquietud para Colombia pues la exactitud encontrada fue mucho mejor que la mencionada en el estudio citado.

Los resultados de la evaluación de la exactitud posicional por localidad fueron los siguientes:

Tabla 5-5: Exactitud posicional 95%.

Exactitud posicional 95%				
ID	Localidad	Conteo elementos	RMSr	Error 95%
1	ANTONIO NARIÑO	1937	1,867	3,23
2	TUNJUELITO	2263	2,547	4,41
3	RAFAEL URIBE URIBE	4989	2,121	3,67
4	CANDELARIA	561	1,927	3,34
5	BARRIOS UNIDOS	4273	1,883	3,26
6	TEUSAQUILLO	4729	1,976	3,42
7	PUENTE ARANDA	5473	2,041	3,53
8	LOS MARTIRES	2358	1,783	3,09
9	SUMAPAZ	157	3,525	6,10
10	USAQUEN	7178	2,168	3,75
11	CHAPINERO	3831	2,108	3,65
12	SANTA FE	2685	2,075	3,59
13	SAN CRISTOBAL	5980	1,888	3,27
14	USME	6557	2,074	3,59
15	CIUDAD BOLIVAR	9983	1,812	3,14
16	BOSA	8276	2,109	3,65
17	KENNEDY	10951	2,349	4,07
18	FONTIBON	5487	2,251	3,90
19	ENGATIVA	10783	2,255	3,90
20	SUBA	15709	2,466	4,27

En cuanto a los resultados del error posicional al 95% se encontró, como se puede apreciar en la **Tabla** (5.5), que estos errores se duplican respecto al error posicional global calculado (3,98 m), Mostrando, por ejemplo, que localidades como Tunjuelito y Sumapaz sobrepasan los 4 metros en el error posicional evaluados al 95% de confianza.

Los valores máximos encontrados corresponden a la localidad rural Sumapaz, $RMSr = 3.525$. La **Figura** (5.18) muestra en rojo los valores $RMSE$ más elevados, Los colores amarillos muestran valores $RMSEr$ dentro del rango 1,98 y 2,16 metros, y finalmente los colores verdes muestran los valores $RMSE$ más bajos.

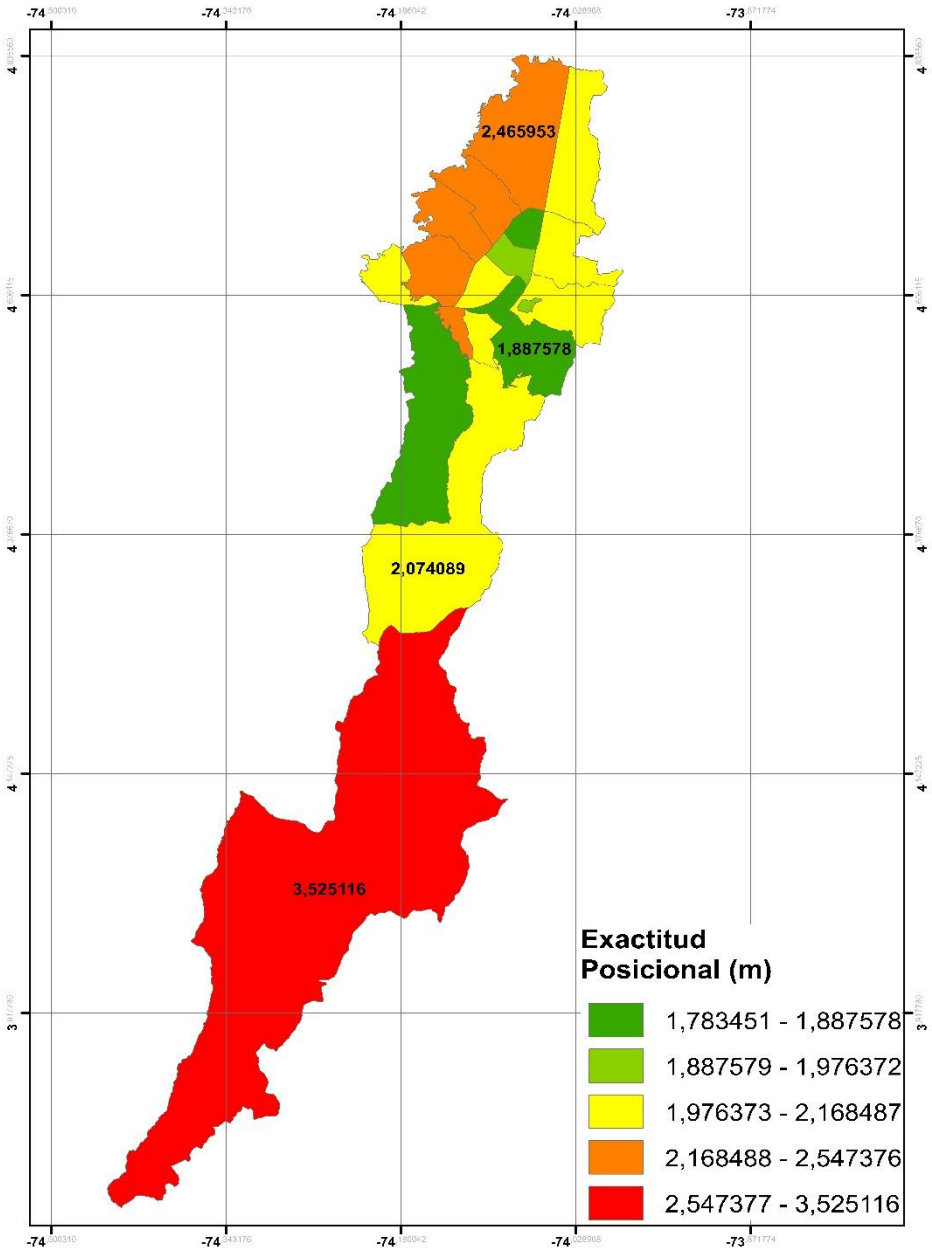


Figura 5-18: Exactitud posicional por localidad.

5.3.3 Exactitud temática

Los resultados de exactitud temática para los atributos de Jerarquía vial, Dirección de flujo y nombramiento principal de la vía se muestran a continuación.

- Jerarquía Vial

La comparación de atributos para jerarquía vial encontró que todos los valores de error estaban sobre el 20%. Las localidades con más errores en la clasificación vial fueron Sumapaz con 98.7% de elementos incorrectamente clasificados, seguida de La candelaria con 52,9%, y Ciudad Bolívar con un 48.2%. En contraste y como se puede observar en la **Tabla** (5.6), localidades como Bosa y Kennedy mostraron los valores más bajos en cuanto a conteo de errores. Esto debido a que son localidades que presentan valores altos en cuanto a vías tipo residencial, categoría que resulto menos afectada en cuanto a la cantidad de errores compilados. Este resultado es compatible con lo descrito por(Jackson et al., 2013) el cual aseguraba que áreas más poblas tienden a tener mostrar mejores resultados y que estos estaban relacionados con la completitud de los datos.

En general se encontró que el 35.8% de los datos evaluados se encontraban mal clasificados respecto a la Jerarquía vial. En la siguiente tabla se pueden observar los resultados de la evaluación de la clasificación jerárquica por cada una de las localidades.

Tabla 5-6: Exactitud temática- Jerarquía vial.

Exactitud temática- Jerarquía vial			
Localidad	#Elementos evaluados	# errores encontrados	% error
ANTONIO NARIÑO	1937	662	34,2
TUNJUELITO	2263	700	30,9
RAFAEL URIBE URIBE	4989	1625	32,6
CANDELARIA	561	297	52,9
BARRIOS UNIDOS	4273	1650	38,6
TEUSAQUILLO	4729	1978	41,8
PUENTE ARANDA	5473	1586	29,0

LOS MARTIRES	2358	962	40,8
SUMAPAZ	157	155	98,7
USAQUEN	7178	2427	33,8
CHAPINERO	3831	1713	44,7
SANTA FE	2685	1059	39,4
SAN CRISTOBAL	5980	2112	35,3
USME	6557	3654	55,7
CIUDAD BOLIVAR	9983	4811	48,2
BOSA	8276	1642	19,8
KENNEDY	10951	3016	21,9
FONTIBON	5487	2371	43,2
ENGATIVA	10783	3164	29,3
SUBA	15709	5262	33,5

La **Figura** (5.19) ilustra gráficamente la información expuesta en la tabla anterior.

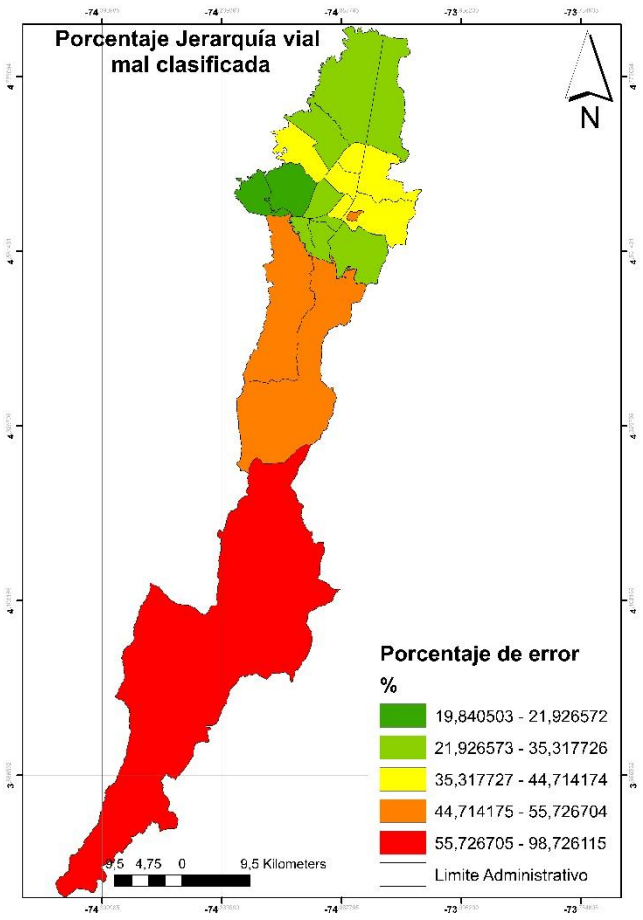


Figura 5-19: exactitud temática Jerarquía vial.

- Sentido vial

La comparación del atributo sentido vial dio como resultado que existe una distribución espacial del error, localizando la mayoría de desaciertos en la parte central de la ciudad. El primer rango encontrado respecto al porcentaje de errores (ver **Figura** (5.20)) muestra que entre el 0 y 10% de estos están ubicados en el sur de la ciudad. En color Naranja y amarillo se encuentra el 10 – 27% y finalmente en color rojo, encontramos los errores sobre el 27%. En contraste y como se puede observar en la **Tabla** (5.7), localidades como Sumapaz y ciudad Bolívar mostraron los valores más bajos en cuanto a conteo de errores. Esto debido a que son localidades que presentan valores altos en doble sentido de circulación, categoría que resultó menos afectada en cuanto a la cantidad de errores.

Tabla 5-7: Exactitud temática sentido vial.

Exactitud temática sentido vial			
Localidad	Elementos	# errores encontrados	% error
ANTONIO NARIÑO	1937	490	25,3
TUNJUELITO	2263	621	27,4
RAFAEL URIBE URIBE	4989	775	15,5
CANDELARIA	561	179	31,9
BARRIOS UNIDOS	4273	1105	25,9
TEUSAQUILLO	4729	1212	25,6
PUENTE ARANDA	5473	1144	20,9
LOS MARTIRES	2358	869	36,9
SUMAPAZ	157	0	0,0
USAQUEN	7178	932	13,0
CHAPINERO	3831	1004	26,2
SANTA FE	2685	636	23,7
SAN CRISTOBAL	5980	428	7,2
USME	6557	213	3,2
CIUDAD BOLIVAR	9983	344	3,4
BOSA	8276	415	5,0
KENNEDY	13755	1440	10,5
FONTIBON	5487	907	16,5
ENGATIVA	10783	1589	14,7
SUBA	15709	1503	9,6

La localidad con más errores encontrados corresponde a los Mártires, con el 36,9% de sentidos viales no concordantes, Sin embargo, no se esperaban valores tan elevados en cuanto la calidad de esté atributo por lo que la validación del método empleado será discutida posteriormente. En general se encontró que el 15% de los datos evaluados se encontraban mal clasificados respecto a su sentido vial. Este valor es relativamente alto dado el resultado de otros autores como(Michael F. Goodchild & Li, 2012; Haklay, 2010; Zielstra & Zipf, 2010)

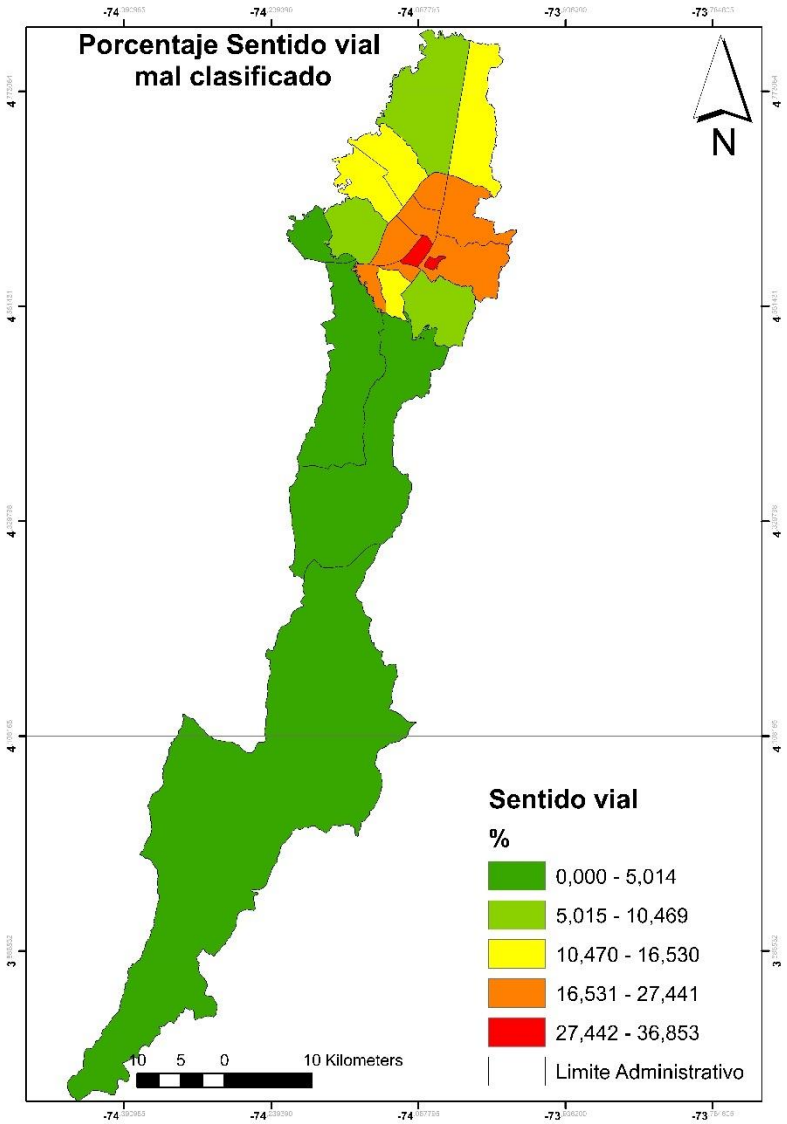


Figura 5-20: Porcentaje sentido vial mal clasificado.

- Nombre vial

Sobre este atributo, se encontró (ver figura 4.21 y tabla 5-8) que el rango de error se ubicó entre 22.9 y 100%. Las localidades con más errores se ubicaron al sur de la ciudad, Entre ellas se destacó Sumapaz con el 100% del nombramiento errado y Usme con el 71.8%, Por otro lado, los porcentajes más bajos en relación a la cantidad de errores encontrados se ubicaron en las localidades de Bosa 32.1%, Kennedy 24.9% y Engativá con él 25.5%.

Tabla 5-8: Exactitud del nombre vial.

Exactitud del nombre vial			
Localidad	Elementos evaluados	# errores encontrados	% error
ANTONIO NARIÑO	1937	598	30,9
TUNJUELITO	2263	849	37,5
RAFAEL URIBE URIBE	4989	2358	47,3
CANDELARIA	561	150	26,7
BARRIOS UNIDOS	4273	1230	28,8
TEUSAQUILLO	4729	1755	37,1
PUENTE ARANDA	5473	1468	26,8
LOS MARTIRES	2358	539	22,9
SUMAPAZ	157	157	100,0
USAQUEN	7178	2341	32,6
CHAPINERO	3831	1240	32,4
SANTA FE	2685	993	37,0
SAN CRISTOBAL	5980	2563	42,9
USME	6557	4711	71,8
CIUDAD BOLIVAR	9983	3392	34,0
BOSA	8276	1908	23,1
KENNEDY	13755	3431	24,9
FONTIBON	5487	2408	43,9
ENGATIVA	10783	2748	25,5
SUBA	15709	5827	37,1

En términos generales, el 38.16% de todas las vías se encontró con porcentaje de clasificación errado.

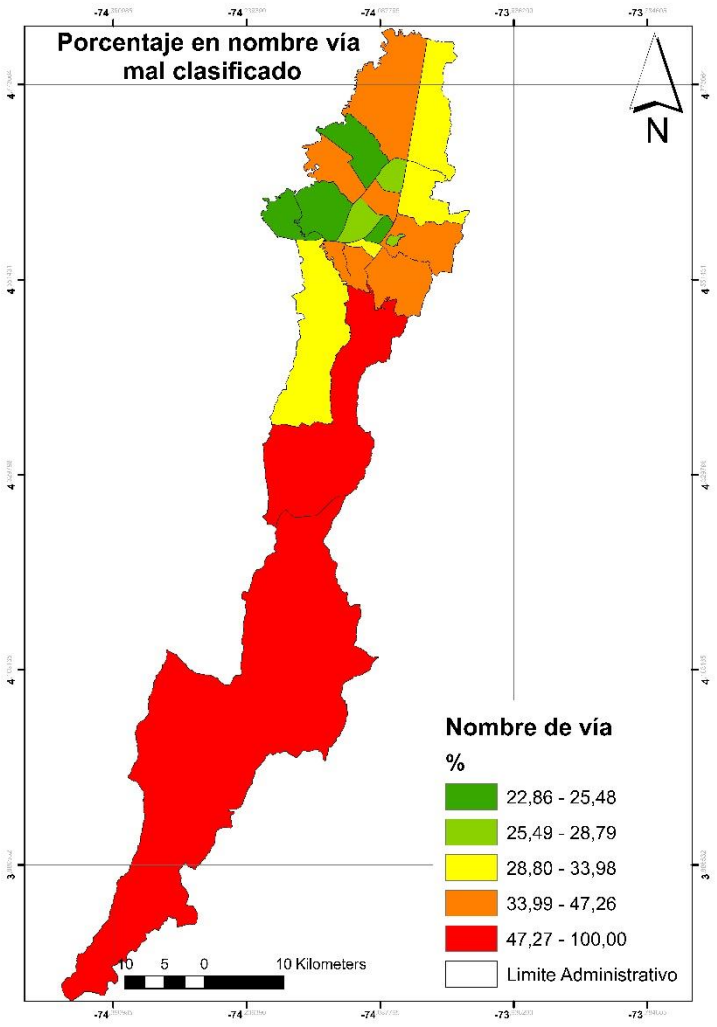


Figura 5-21: Porcentaje Nombre vial mal clasificado.

Para generar un resultado total de la calidad VGI usando las medidas anteriormente expuestas, se tomaron cada uno de los resultados de las **Tablas** (5.3 [Resultados Completitud nivel Bogotá], 5.5 [Exactitud posicional 95%], 5.6[Exactitud temática-Jerarquía vial], 5.7[Exactitud temática sentido vial] y 5.8[Exactitud del nombre vial]) y se calcularon los promedios cuando fue necesario. Estos valores son expuestos a continuación

La completitud para la malla vial de Bogotá obtuvo 12.6 km omitidos y 36.42 km en exceso, lo que quiere decir que 36.42 km se encontraban codificado en *OSM,* pero en realidad no existían en la malla vial de IDECA. En términos generales la malla vial de Bogotá evaluada a través de la medida completitud encontró que hacían falta un total de 12.6 km de un total de 8.195,90 Km proveniente de IDECA y que OSM tenía codificado 0.44 km en exceso. El total de elementos final fue de 116 964 datos.

Respecto a la exactitud posicional, el promedio del error en exactitud posicional para TODA la malla vial de Bogotá fue de 3.89m con un error medio cuadrático promedio de 2.16 metros.

La evaluación de la calidad VGI promedio para la exactitud temática-Jerarquía vial encontrada fue del 40.2% de error. Lo que indica que aproximadamente el 60% de los datos contaban con una buena clasificación de Jerarquía vial.

La evaluación promedio para toda la malla vial de Bogotá en relación a la exactitud temática-sentido vial fue de 17.2% de error. Lo que indica que aproximadamente el 82% de los datos contaban con una buena clasificación en sentido vial. En cuanto exactitud temática para el nombramiento vial se encontró que el promedio de error fue de 38.16%, indicando que en promedio el 62% de los datos tenían este dato correctamente codificado.

Con todos estos promedios se creó el siguiente resultado para TODA la malla vial de Bogotá en relación a su calidad VGI:

Tabla 5-9: Resumen de la evaluación VGI.

Evaluación de la calidad VGI calculada de forma univariada					
Malla vial	%Error Jerarquía Vial	%Error Sentido vial	%error Nombramiento	Promedio% error	Exactitud posicional. m
% error	40,2	17,2	38,16	**31,85**	3.98
% sin error	59,8	82,8	61,84	68,15	3.98

Como resultado general se obtuvo que la calidad para la malla vial de Bogotá usando los promedios de las medidas VGI fueron 68.15% con una exactitud posicional promedio de 3.98 m. Lo que indica que solo el 38.16% de los datos contenía algún tipo de error.

La completitud se añade por separado a este resultado pues todos los cálculos parten de la ausencia o no de geometría, donde ya se sabe que OSM contenía -14.6% elementos que IDECA.

5.4 Muestreo simple por asignación proporcional a la localidad y aplicación del análisis multivariado (ACM)

Por cuestión de rendimiento de maquina fue necesario trabajar con una parte de los datos, Por ello se creó un muestreo simple por asignación proporcional a la localidad. Esto permitió trabajar con 6424 registros (menos del 5% de la data original). Los resultados de la asignación proporcional por localidad se muestran a continuación:

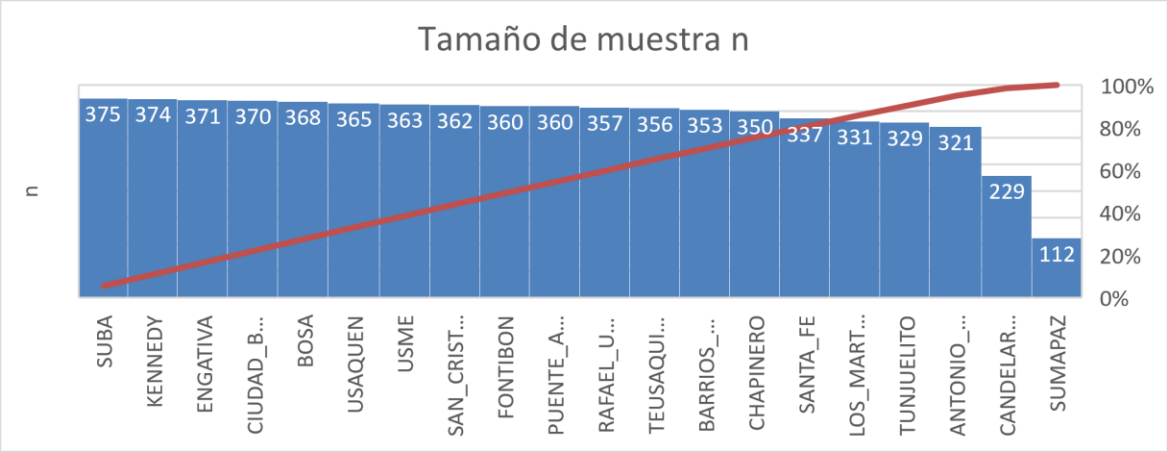


Figura 5-22: Muestreo simple por asignación proporcional a la localidad [5.18].

Se tuvo que trabajar de esta forma ya que el algoritmo ACM no era capaz de procesar todos los datos. La ventaja de haber trabajado con un muestreo simple por asignación proporcional a la localidad consistió en que hizo mucho más fácil el análisis visual ya que todos los datos debían ser proyectados en un mismo plano coordenado. Y al tener una cantidad reducida de estos, las relaciones podían ser analizadas de mejor forma.

El análisis de correspondencia preliminar se generó a partir de 6 variables, una de ellas complementaria. Con estos datos se obtuvieron los porcentajes de inercia acumulada (ver **Figura** (5.23). Para las dos primeras dimensiones se obtuvo un valor de 16.63%. a partir de este valor y con la ayuda de los individuos variables y categorías proyectadas en el espacio coordenado (Dim 1 vs Dim2) se empezó a analizar la relación entre los datos.

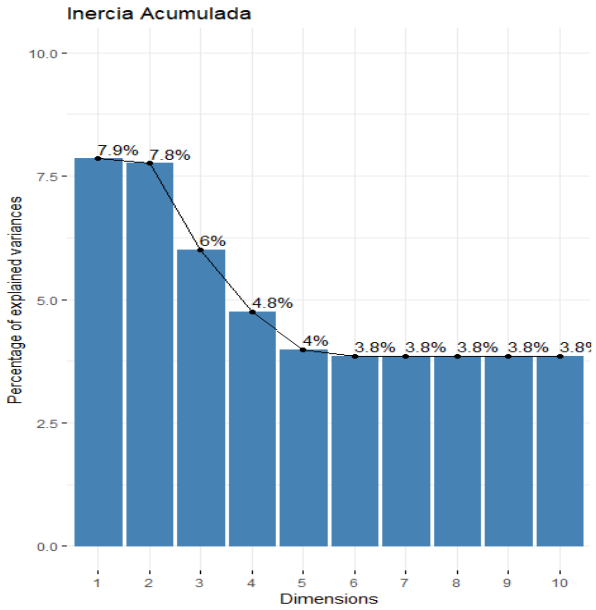


Figura 5-23: Inercia Acumulada.

como resultado inicial se obtuvo que la variable exactitud posicional estaba fuertemente asociada a la variable localidad, como se puede apreciar en la **Figura** 5.24, la variable localidad se encuentra próxima a la variable exactitud posicional y ambas contribuyen fuertemente a la creación de los ejes coordenados. Si hubiésemos querido mostrar cómo está asociada la exactitud posicional con las localidades, este sería el escenario perfecto. Pero lo que aquí se quiso estudiar fue la posible relación que existe entre errores para poder medir su calidad de manera conjunta. Por este motivo la variable localidad fue retirada del estudio. Esto permitió correr de nuevo el ACM con 5 variables, sabiendo que 1 de estas 5 variables era suplementaria, la cual no participa en los cálculos de inercia ni en la creación de los ejes coordenados. Esta variable permitió entender hacía que tipo de errores estaban tendiendo las observaciones.

En la **Figura** 5.24, la cual hace referencia al primer plano coordenado creado con 6 variables, se puede observar en la parte inferior izquierda, que los errores de jerarquía vial, sentido vial y nombramiento se encontraban próximos a la variable suplementaria (circulo verde).

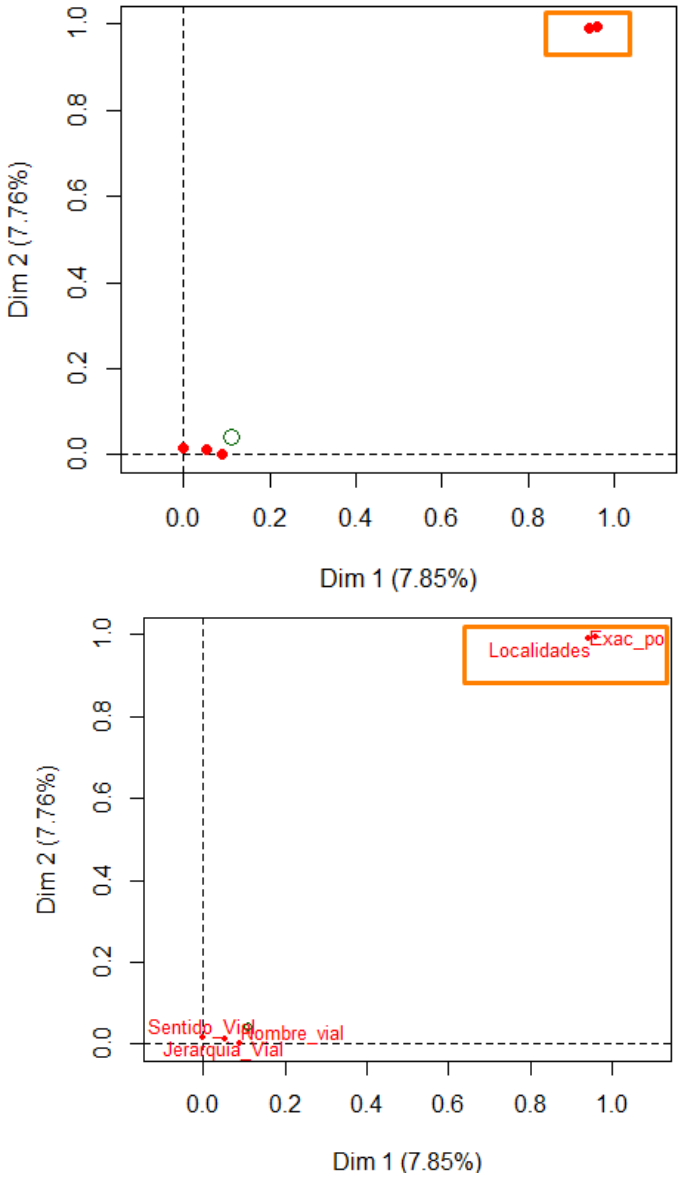


Figura 5-24: Plano factorial construido usando 6 variables.

Eliminada la variable localidad, se recalculó el *ACM*. Obteniendo valores de inercia cercanos al 48% en las dos primeras dimensiones (Ver (**Figura** (5.25)), esto en contraste con el primer ACM que retenía únicamente el 16% de la varianza acumulada en los dos primeros ejes coordenados. Graficando las variables en el primer plano factorial se pudo observar que la exactitud posicional y sentido vial se encontraron próximas, mientras que la jerarquía y nombramiento vial poseían el mismo comportamiento (Ver **Figura** (5.25)).

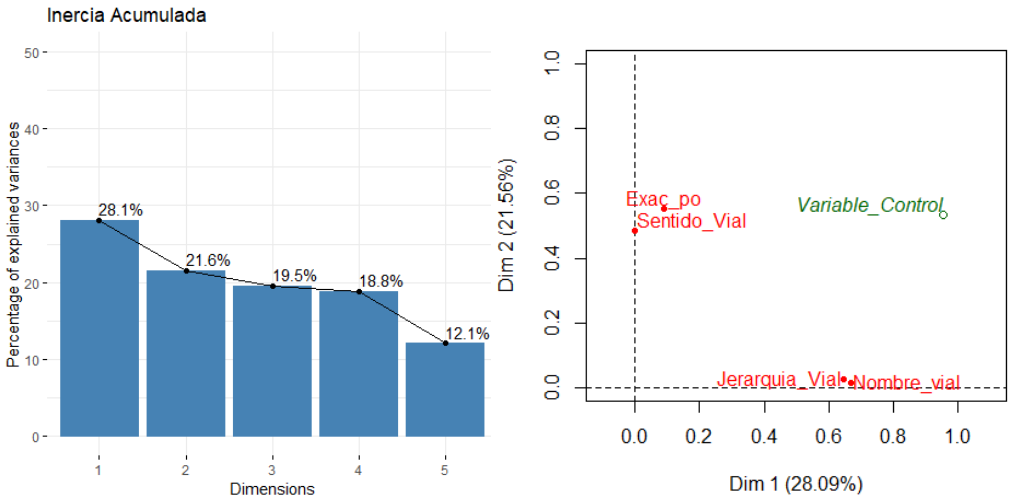


Figura 5-25: Inercia acumulada primer plano factorial.

Pero esta proximidad también depende de los ejes x, y en donde los datos fueron representados. Por esta razón se proyectaron las sub categorías de las variables en varias dimensiones, con el fin de seleccionar la proyección que mejor represente a los datos. Como resultado los planos coordenados proyectados fueron: Dim 1vs Dim2, Dim2 vs Dim3, Dim 3vs Dim5 y finalmente dim4vsDim5. Los resultados se muestran en la **Figura** (5.26). Nótese que la proyección de las categorías cambia según los diferentes planos.

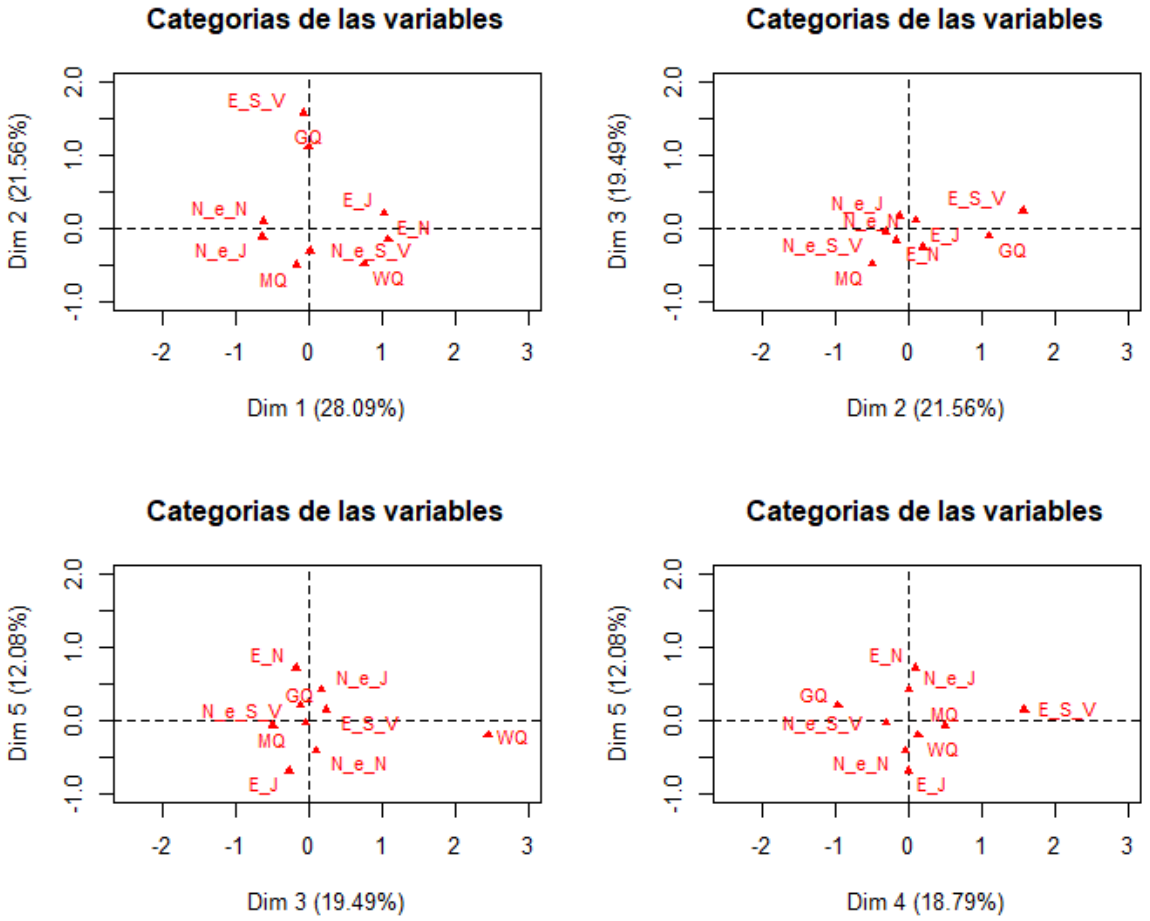


Figura 5-26: Planos factoriales variables.

Las coordenadas generadas para las variables en los planos 1…5, se muestran a continuación:

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
GQ	-0.008764173	1.10462191	-0.11387442	-0.9652441406	0.20557044
MQ	-0.166468566	-0.50133944	-0.49110603	0.5067247469	-0.07043693
WQ	0.756953207	-0.49660417	2.44917613	0.1306878688	-0.19342212
E_J	1.020892125	0.19997070	-0.26859218	-0.0006462312	-0.67761031
N_e_J	-0.632716299	-0.12393545	0.16646485	0.0004005134	0.41996120
E_S_V	-0.075172604	1.56469812	0.23455561	1.5847355441	0.15145809
N_e_S_V	0.014905517	-0.31025445	-0.04650860	-0.3142275477	-0.03003170
E_N	1.085925816	-0.16116388	-0.17573078	0.0911829375	0.72014335
N_e_N	-0.615949627	0.09141401	0.09967652	-0.0517200121	-0.40847360

Las coordenadas de los individuos pueden ser consultadas en el **Anexo F** “Coordenadas de individuos”.

En principio se escogió el plano que tuviera mayor cantidad de inercia retenida, en este caso, el plano conformado por la dimensión 1 y 2. Los resultados de las contribuciones realizadas por las categorías a la construcción de los ejes coordenados se aprecian en la **Figura** (5.27). Allí se puede observar que la contribución en la dimensión 1 fue realizada en su mayoría por la sub categoría E_N: Error en nombramiento, pues el aporte a la creación de estas dimensiones está representado por al menos el 30% de esta variable. Le sigue la variable E_J (Error en Jerarquía vial) con 28% del aporte sobre esta dimensión.

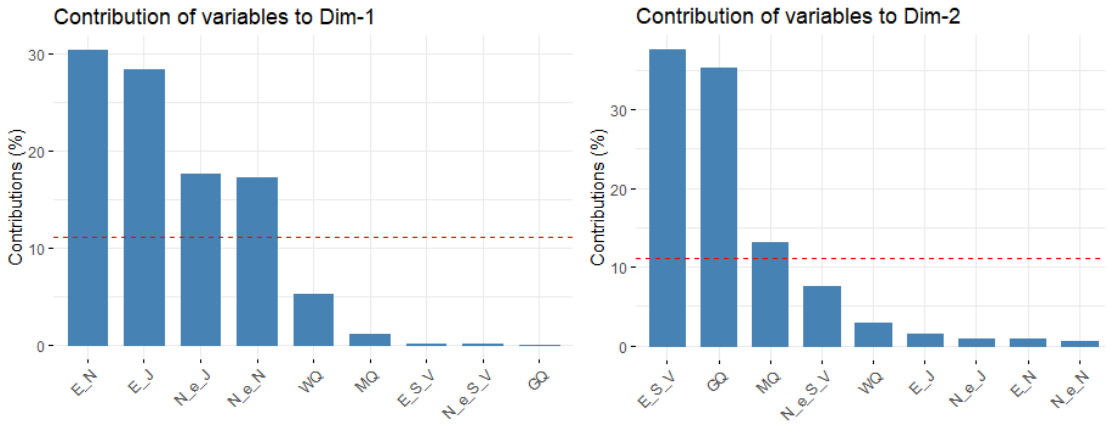


Figura 5-27: Contribuciones por categorías.

Las mayores contribuciones en la dimensión 1 fueron aportadas por *E_S_V y GQ* que son respectivamente error en sentido vial y exactitud posicional con categoría buena, la cual hace referencia a un *RMSE* menor a 1.88 m. En conjunto, estas dos variables aportaron más del 65% a la creación del eje coordenado, por ello el desplazamiento sobre el eje Y estará relacionado a la exactitud posicional y al error en sentidos viales. De manera similar, el eje de las X estará relacionado a tipos de error en jerarquía vial y nombramiento.

La **Figura**: (5.28), permite analizar las contribuciones de manera más cómoda, las variables con colores fríos son aquellas que poco tuvieron que ver con la construcción de los ejes, mientras que las variables resaltadas con colores naranjas y rojizos fueron aquellas que participaron activamente en la creación de los ejes coordenados. La lectura de este grafico se hace de la siguiente manera, de izquierda a derecha en el eje de las x se encuentran aquellos valores que pasan de no tener errores en nombramiento vial,

jerarquía y nombramiento de vias (*N_e_N, N_e_J , N_e_S_V*) a tener aunque sea uno de ellos. Entre más nos desplacemos a la izquierda del eje *X* encontraremos menos datos con error (*N_E: No error*). La calidad de estos datos varia también de izquierda a derecha, donde el cuadrante III aparentemente contendrá los valores de exactitud posicional media MQ, este símbolo indica errores promedio de *RMSE* entre los 2 y los 2.34 metros. El eje de las Y explica la parición de errores en nombramiento vial y un aumento en la exactitud posicional en el sentido +Y.

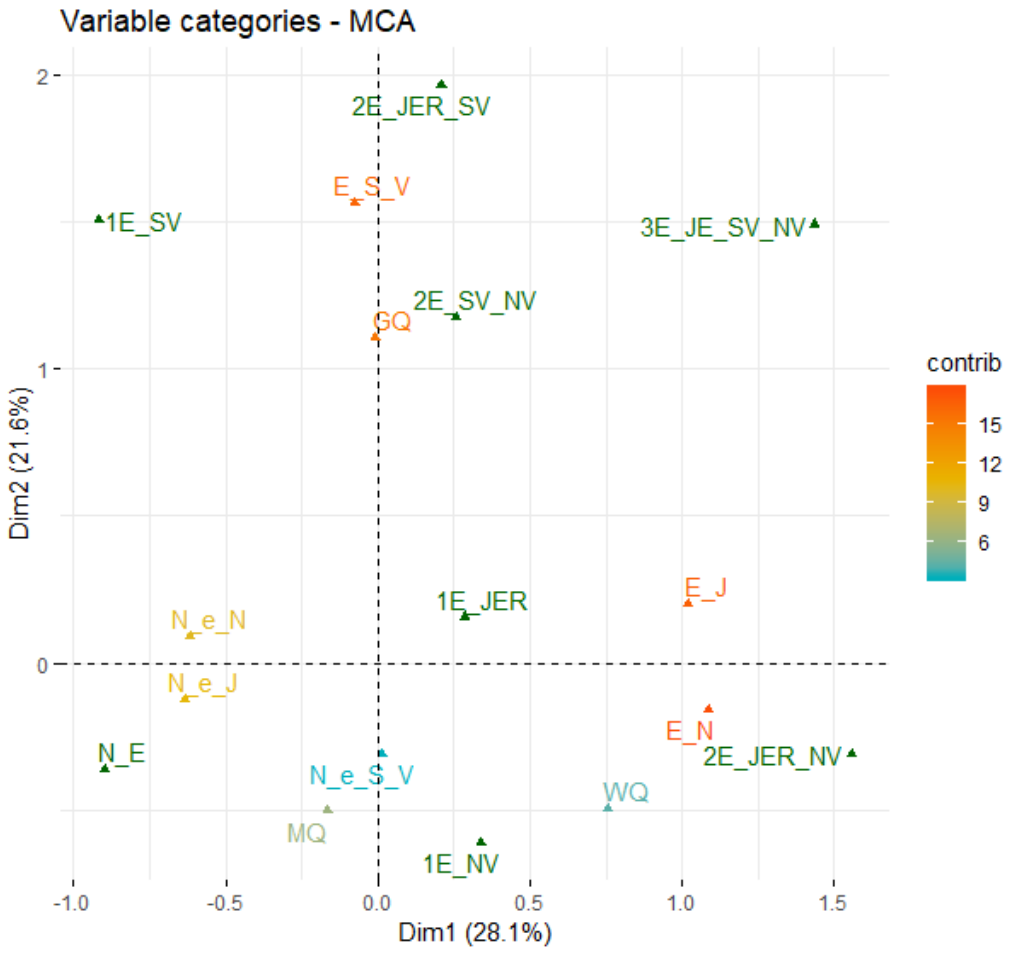


Figura 5-28: Proyección de las variables categóricas y suplementarias.

Finalmente, detectadas las relaciones entre variables y categorías haciendo uso de las dimisiones 1,2. Se procedió a proyectar los individuos con las coordenadas que calculó el ACM para cada uno de ellos.

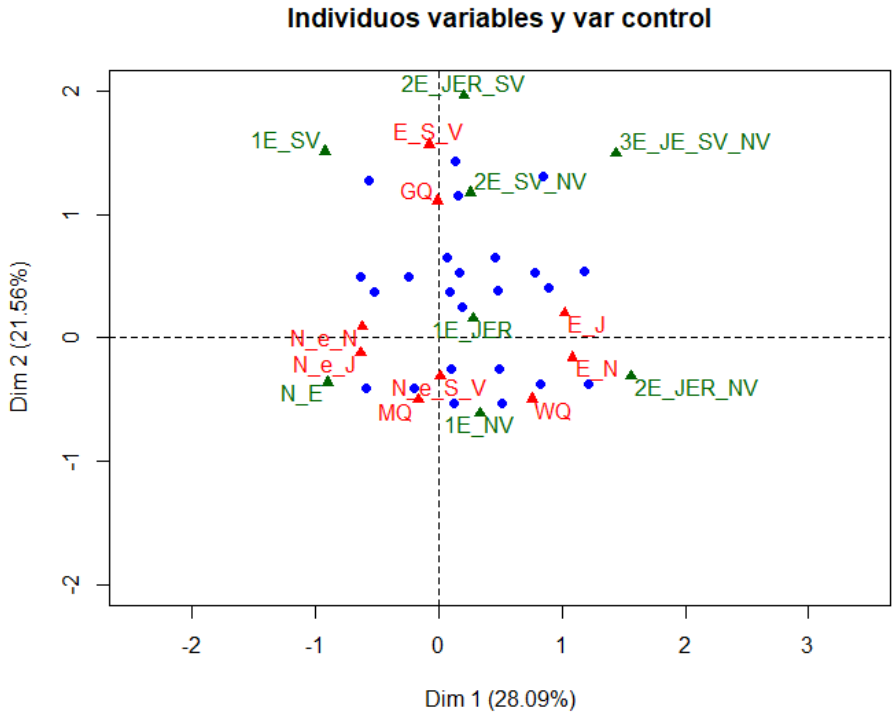


Figura 5-29: Proyección de individuos.

Tabla 5-10: Variables objeto de estudio

Variables	Descripción	RMSE m
GQ	Buena calidad	<= 1.88
MQ	Calidad media	2-2.34m
WQ	Mala calidad	>2.47
E_J	Error en jerarquía	na
N_e_J	Ningun error en jerarquía	na
E_S_V	Error en sentido vial	na
N_e_S_V	Ningun error en sentido vial	na
E_N	Error en nombre de vía	na
N_e_N	Ningun error encontrado	na

Para hacer legible el grafico, solo se proyectó un pequeño porcentaje de individuos (Puntos en color azul). Las variables verdes corresponden a la variable suplementaria, la cual permite entender la relación que hay entre tipo de errores e individuos. Por ejemplo en la **Figura** (5.29), se puede ver que existen pocos individuos que poseen 3 tipos de errores (3E_JE_SV_NV), Sabemos que son pocos por 2 razones, la primera es que no hay individuos cercanos a esta etiqueta, la segunda razón tiene que ver con su posición en los ejes proyectados, pues variables categóricas que se encuentren lejos del centro de masas indican un conteo de frecuencias bajo, por lo que el análisis ACM los considera eventos poco frecuentes. En contra partida existen muchos individuos (nodos) que poseen el error 1E_JER, el cual indica errores en la jerarquización vial, estos individuos solo poseen este tipo de error y a demás siguen la tendencia de exactitud posicional GQ, que significa buena calidad. En el cuadrante IV, se puede observar que los datos poseen la peor exactitud posicional calculada, esta hace referencia a un error medio cuadrático superior a 2.47 metros. Este valor aparentemente está relacionado con errores en Jerarquía y nombre vial y poco relacionados con los errores de sentido vial ubicados en el cuadrante II.

5.5 Agrupación y visualización de resultados por medio de análisis de conglomerados (Método Jerárquico)

El análisis visual mostrado en el ACM por medio de la **Figura** (5.28) y (5.29) permite entender como los errores referentes a la calidad VGI aparentemente se encuentran relacionados. Ahora con la ayuda de los clústeres, agrupamos de manera sistemática aquellas relaciones encontradas. Como recordará el lector, el método aglomerativo aquí empleado parte de las inercias y coordenadas calculadas en el ACM para generar los conglomerados, esto quiere decir que los conglomerados calculados solo son una representación de agrupamiento encontrado en el ACM (ver **Figura** (5.30)) usando, como se mencionó anteriormente las coordenadas calculadas de los individuos.

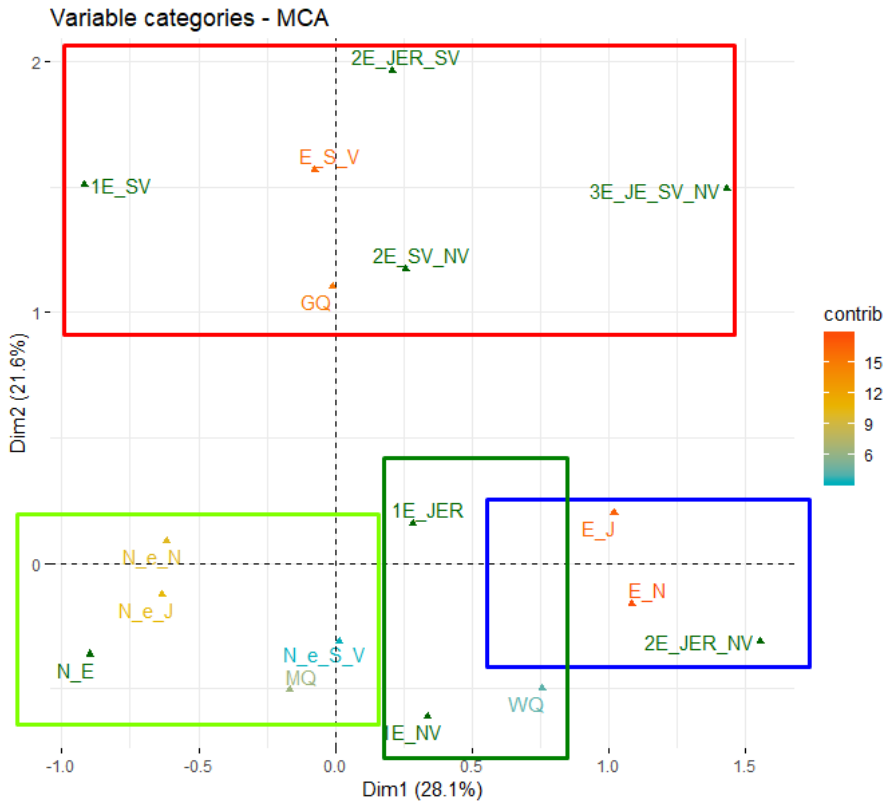


Figura 5-30: Esquema ilustrativo de aglomeración de categorías.

La aglomeración se realizó haciendo uso del método de Ward y la métrica de similitud euclídea, pues fue el algoritmo y métrica de distancia que agrupo de mejor forma los resultados. A continuación, se muestra el dendograma usado para seleccionar el número de clústeres final.

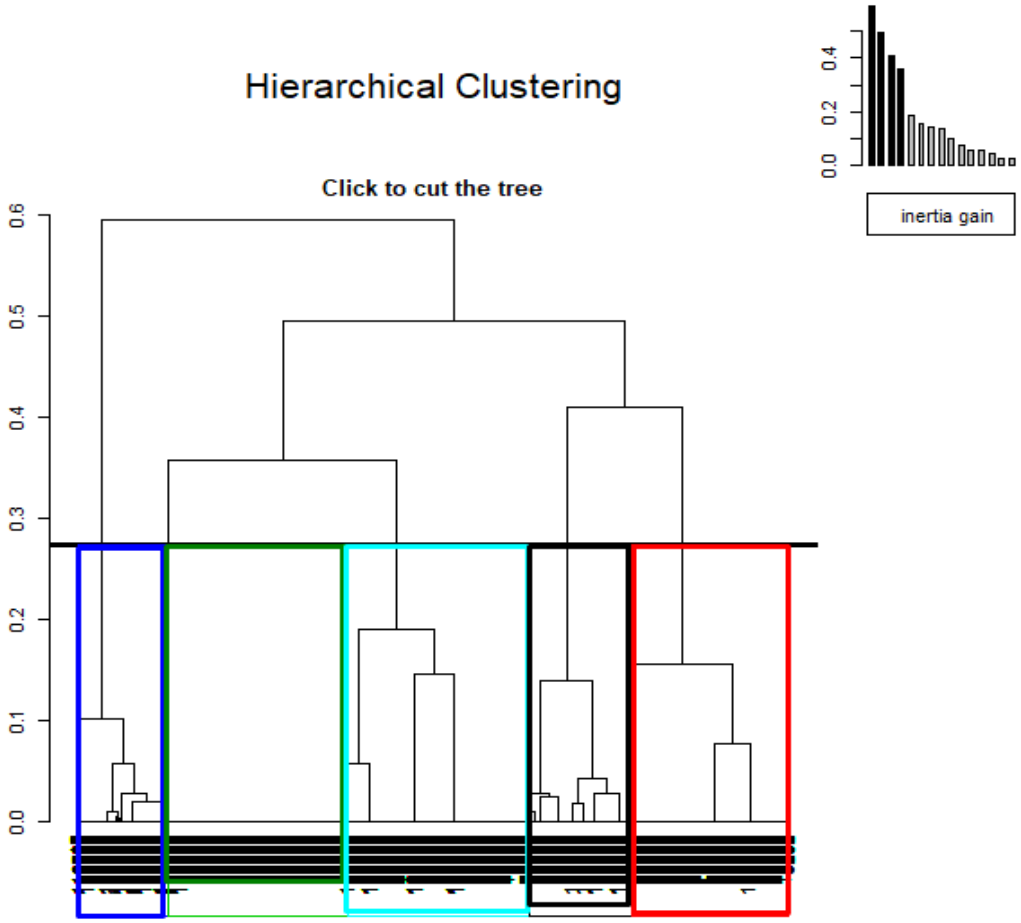


Figura 5-31: Dendograma de agrupación por nodos comparados.

La clasificación de los individuos se puede observar en los recuadros de color azul, verde, cian, negro y rojo, resaltadas dentro del dendograma. En la parte superior derecha de la **Figura** (5.31) se encuentran las inercias de los grupos generados, el algoritmo usa la inercia calculada en cada una de las dimensiones para extraer la mayor cantidad de individuos proyectados. La cantidad de dimensiones usadas fue 5, pues en ellas (ver **Figura** (5.27)) se acumuló más del 80% de la inercia explicada.

Cada uno de estos conglomerados contiene la aparición o no de errores encontrados por individuo, cálculo de exactitud posicional y demás medidas de calidad, esta información se extrajo de cada clúster y fue unida a los datos de localidad para poder comparar los resultados multivariados vs univariados expuestos en la sección 5.3. Cada uno de estos clústeres fue clasificado de la siguiente manera:

Los clústeres de color **negro** hacen referencia a todos los elementos que no contenían ningún tipo de error y su exactitud posicional contenía el dato de MQ, el cual significa Exactitud posicional media, haciendo referencia a que los errores medios cuadráticos en este clúster se encontraban dentro del rango de los 2- 2.34 metros. Los clústeres **azules** hacen referencia a datos de tipo Jerarquía y nombramiento vial, donde existen combinación de errores para estas dos categorías y además su calidad posicional es considerada como la peor dentro del análisis "WQ" pues su error medio cuadrático se encuentra sobre los 2.47 metros.

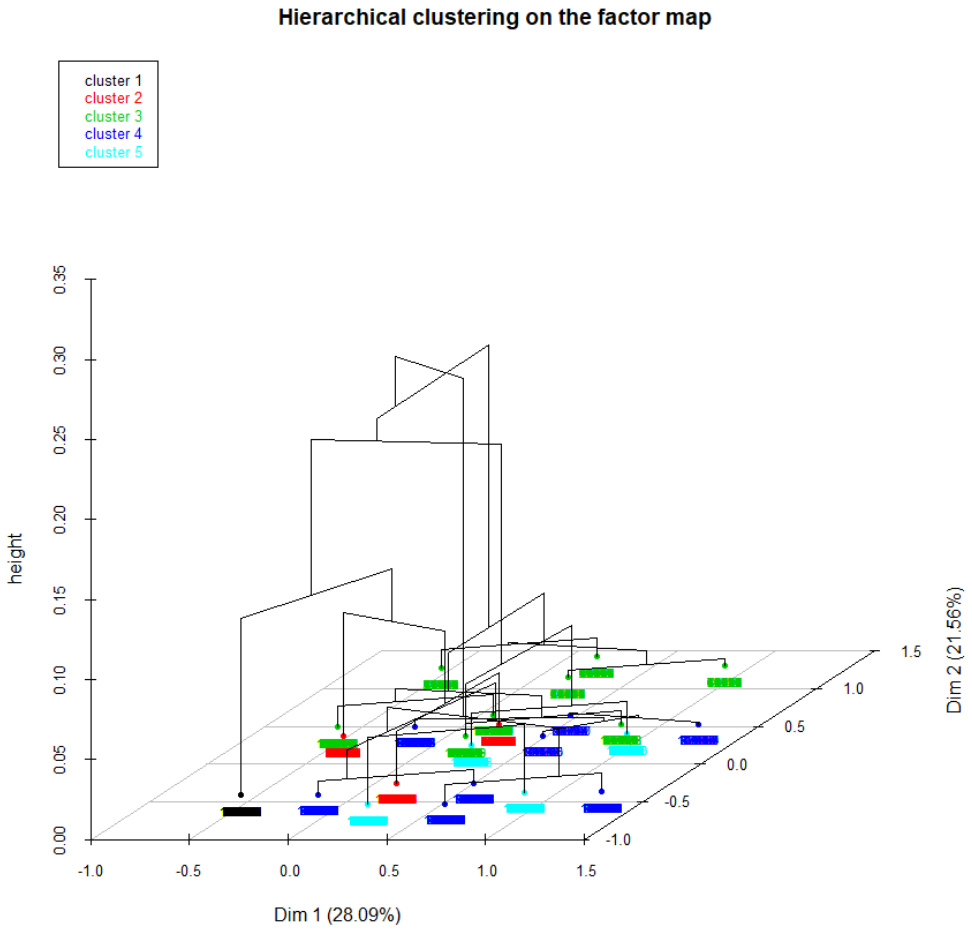
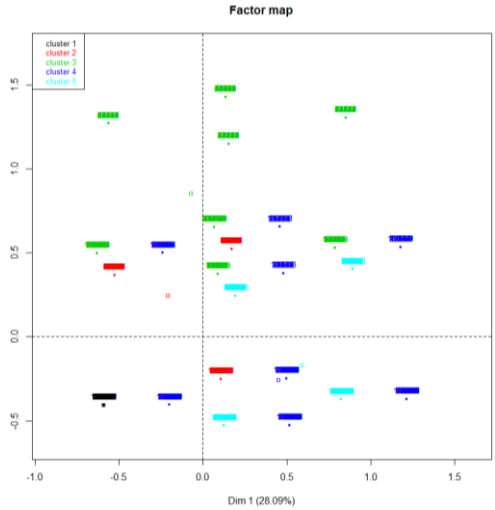


Figura 5-32: Agrupación de errores por clústeres.

El clúster verde, contiene los datos que poseen la mejor calidad en cuanto a exactitud posicional, pues se ubican por debajo de los 1.98 metros. También contiene errores combinados que dependen considerablemente del sentido vial. En este clúster la exactitud temática se ve fuertemente impactada, aunque su exactitud posicional es muy buena. El clúster **rojo** recoge individuos con exactitud posicional media MQ a calidad GQ, pero la exactitud temática varía de acuerdo a la cantidad de errores conjuntos que se encontraron. Errores en sentido vial y nombramiento aparecen de manera conjunta mientras que los errores de jerarquía vial aparecen, pero de manera separada, esto quiere decir que algunos errores aparecían hasta dos veces en un mismo individuo y en otras ocasiones un individuo solo parecía tener un único error. Finalmente, el clúster **cian** contiene un compilado de errores individuales relacionados a errores tipo jerárquico y de nombramiento vial que pueden aparecer de manera conjunta (dentro de un mismo nodo) o de manera separada. La exactitud posicional en este clúster varía desde muy buena GQ a mala WQ. Toda la información referente a clúster puede ser revisada en el **Anexo E**.

El resultado de esta aglomeración en los planos factoriales escogidos se presenta en las **Figuras**: 5.31 y 5.32. La figura 5.33 contiene una breve descripción de las distancias a las cuales se formar los clústeres.




```
Cluster: 1
      704       786       931      1636      1678
0.9674664 0.9674664 0.9674664 0.9674664 0.9674664
------------------------------------------------------------
Cluster: 2
     6155      6242      6892      6921      6930
1.334224 1.334224 1.334224 1.334224 1.334224
------------------------------------------------------------
Cluster: 3
    12498     18614     18932     20123     22373
1.673801 1.673801 1.673801 1.673801 1.673801
------------------------------------------------------------
Cluster: 4
   100064    102379     89215     90885    101597
1.762786 1.762786 1.762786 1.762786 1.762786
------------------------------------------------------------
Cluster: 5
      166       198       292       335       341
1.440061 1.440061 1.440061 1.440061 1.440061
```

Figura 5-33: Agrupamiento de resultados y datos de clúster.

Para cada clúster la información fue extraída y combinada con las localidades, con el fin de poder realizar la comparación numérica. Como resultado se obtuvo la siguiente figura:

Localidad	Cluster 1E_JER_GQ	Cluster 1E_JER_MQ	Cluster 1E_JER_WQ	Cluster 1E_NV_GQ	Cluster 1E_NV_MQ	Cluster 1E_NV_WQ	Cluster 1E_SV_GQ	Cluster 1E_SV_MQ	Cluster 1E_SV_WQ	Cluster 2E_JER_NV_GQ	Cluster 2E_JER_NV_MQ	Cluster 2E_JER_NV_WQ	Cluster 2E_JER_SV_GQ	Cluster 2E_JER_SV_MQ	Cluster 2E_JER_SV_WQ	Cluster 2E_SV_NV_GQ	Cluster 2E_SV_NV_MQ	Cluster 2E_SV_NV_WQ	Cluster 3E_JE_SV_NV_GQ	Cluster 3E_JE_SV_NV_MQ	Cluster 3E_JE_SV_NV_WQ	Cluster N_E_GQ	Cluster N_E_MQ	Cluster N_E_WQ
ANTONIO_NA	16,8%	0,0%	0,0%	7,6%	0,0%	0,0%	10,7%	0,0%	0,0%	13,0%	0,0%	0,0%	4,6%	0,0%	0,0%	4,6%	0,0%	0,0%	3,1%	0,0%	0,0%	39,7%	0,0%	0,0%
BARRIOS_UNI	10,8%	0,0%	0,0%	7,8%	0,0%	0,0%	8,3%	0,0%	0,0%	13,2%	0,0%	0,0%	5,9%	0,0%	0,0%	2,0%	0,0%	0,0%	3,4%	0,0%	0,0%	48,5%	0,0%	0,0%
BOSA	0,0%	8,4%	0,0%	0,0%	8,4%	0,0%	0,0%	3,0%	0,0%	0,0%	12,0%	0,0%	0,0%	0,8%	0,0%	0,0%	0,3%	0,0%	0,0%	1,4%	0,0%	0,0%	65,8%	0,0%
CANDELARIA	14,0%	0,0%	0,0%	5,2%	0,0%	0,0%	17,5%	0,0%	0,0%	12,2%	0,0%	0,0%	12,2%	0,0%	0,0%	0,9%	0,0%	0,0%	2,2%	0,0%	0,0%	35,8%	0,0%	0,0%
CHAPINERO	0,0%	15,4%	0,0%	0,0%	5,1%	0,0%	0,0%	9,1%	0,0%	0,0%	20,6%	0,0%	0,0%	7,7%	0,0%	0,0%	2,3%	0,0%	0,0%	4,6%	0,0%	0,0%	35,1%	0,0%
CIUDAD_BOLI	28,1%	0,0%	0,0%	9,7%	0,0%	0,0%	1,4%	0,0%	0,0%	19,5%	0,0%	0,0%	0,5%	0,0%	0,0%	0,3%	0,0%	0,0%	0,5%	0,0%	0,0%	40,0%	0,0%	0,0%
ENGATIVA	0,0%	12,7%	0,0%	0,0%	4,9%	0,0%	0,0%	7,3%	0,0%	0,0%	12,4%	0,0%	0,0%	4,0%	0,0%	0,0%	4,0%	0,0%	0,0%	2,2%	0,0%	0,0%	52,6%	0,0%
FONTIBON	0,0%	9,7%	0,0%	0,0%	9,7%	0,0%	0,0%	5,8%	0,0%	0,0%	25,8%	0,0%	0,0%	5,0%	0,0%	0,0%	5,0%	0,0%	0,0%	2,5%	0,0%	0,0%	36,6%	0,0%
KENNEDY	0,0%	8,0%	0,0%	0,0%	9,4%	0,0%	0,0%	6,4%	0,0%	0,0%	12,3%	0,0%	0,0%	1,9%	0,0%	0,0%	2,1%	0,0%	0,0%	1,1%	0,0%	0,0%	58,8%	0,0%
LOS_MARTIRE	14,3%	0,0%	0,0%	6,1%	0,0%	0,0%	14,3%	0,0%	0,0%	10,7%	0,0%	0,0%	13,1%	0,0%	0,0%	4,0%	0,0%	0,0%	3,4%	0,0%	0,0%	34,1%	0,0%	0,0%
PUENTE_ARA	0,0%	8,3%	0,0%	0,0%	6,1%	0,0%	0,0%	10,0%	0,0%	0,0%	10,0%	0,0%	0,0%	2,2%	0,0%	0,0%	3,1%	0,0%	0,0%	6,1%	0,0%	0,0%	54,2%	0,0%
RAFAEL_URIB	0,0%	11,5%	0,0%	0,0%	26,1%	0,0%	0,0%	4,2%	0,0%	0,0%	13,4%	0,0%	0,0%	3,9%	0,0%	0,0%	4,5%	0,0%	0,0%	2,2%	0,0%	0,0%	34,2%	0,0%
SAN_CRISTOB	13,3%	0,0%	0,0%	19,6%	0,0%	0,0%	2,2%	0,0%	0,0%	23,5%	0,0%	0,0%	1,9%	0,0%	0,0%	2,2%	0,0%	0,0%	0,6%	0,0%	0,0%	36,7%	0,0%	0,0%
SANTA_FE	0,0%	12,8%	0,0%	0,0%	9,2%	0,0%	0,0%	8,9%	0,0%	0,0%	15,2%	0,0%	0,0%	5,1%	0,0%	0,0%	3,3%	0,0%	0,0%	5,1%	0,0%	0,0%	40,5%	0,0%
SUBA	0,0%	0,0%	11,2%	0,0%	0,0%	10,9%	0,0%	0,0%	4,8%	0,0%	0,0%	23,7%	0,0%	0,0%	1,1%	0,0%	0,0%	1,3%	0,0%	0,0%	2,1%	0,0%	0,0%	44,8%
SUMAPAZ	0,0%	0,0%	0,0%	0,0%	0,0%	1,8%	0,0%	0,0%	0,0%	0,0%	0,0%	98,2%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
TEUSAQUILLO	10,7%	0,0%	0,0%	6,2%	0,0%	0,0%	10,4%	0,0%	0,0%	19,9%	0,0%	0,0%	8,1%	0,0%	0,0%	5,1%	0,0%	0,0%	5,9%	0,0%	0,0%	33,7%	0,0%	0,0%
TUNJUELITO	0,0%	0,0%	6,1%	0,0%	0,0%	16,7%	0,0%	0,0%	16,1%	0,0%	0,0%	19,1%	0,0%	0,0%	3,6%	0,0%	0,0%	4,3%	0,0%	0,0%	3,0%	0,0%	0,0%	31,0%
USAQUEN	0,0%	8,5%	0,0%	0,0%	11,8%	0,0%	0,0%	5,2%	0,0%	0,0%	17,8%	0,0%	0,0%	4,7%	0,0%	0,0%	1,9%	0,0%	0,0%	1,4%	0,0%	0,0%	48,8%	0,0%
USME	0,0%	4,1%	0,0%	0,0%	22,9%	0,0%	0,0%	0,6%	0,0%	0,0%	47,9%	0,0%	0,0%	0,0%	0,0%	0,0%	1,7%	0,0%	0,0%	0,6%	0,0%	0,0%	22,3%	0,0%

Figura 5-34: Datos extraídos de los Clústeres.

El siguiente resultado es uno de los más importantes de esta tesis, pues da solución a el objetivo principal de este trabajo.

Esta tabla contiene todas las combinaciones de errores y no errores (en %) encontrados. Al final de esta tabla se puede observar todos los valores que no tuvieron error en ninguno de las variables analizadas NE_**GQ**, NE_**MQ** y NE**WQ.** En cuanto a la exactitud posicional, cada uno de estos hallazgos está relacionado con este valor. Recuerde que GQ, MQ y WQ hacen referencia al RMCE calculado, donde GC comprende valores por debajo de los 1.98 m. MQ se encuentra entre los 2,4 a 2,35m y WQ, que se encuentra por encima de los 2.47m. Entonces usando la cantidad de errores encontrados se pudo establecer que:

Tabla 5-11: Resumen de la evaluación VGI bajo el análisis Multivariado VGI.

Malla vial	N_E_GQ	N_E_MQ	N_E_WQ	Promedio % error	Exactitud posicional
% sin error	13,4%	22,4%	3,8%	39,7%	3.97
% error	86,6%	77,6%	96,2%	**60,3%**	3.97

La calidad VGI promedio, evaluada de manera multivariada dio como resultado que el 60.3% de los datos poseían al menos un error y que en promedio su exactitud posicional

era de 3.97m. En contraste, los datos obtenidos midiendo la calidad VGI de manera univariada (ver **Tabla** (5.9)), mostraron que el promedio de error hallado para las 20 localidades fue de 31.85% con una exactitud posicional promedio de 3.97 m. Casi 40% menos error que el encontrado en el análisis multivariado. Haciendo el mismo ejercicio, pero con los datos que no poseen ningún error, podemos observar que para el método multivariado este valor fue de 39.7% mientras que, en contraparte para el cálculo univariado el valor hallado fue de 61.84% con una exactitud posicional de 3.97 m. En términos generales los valores encontrados de manera multivariada son mucho más pesimistas que los encontrados de manera univariada. Note que la tabla anterior tiene embebidos los cálculos de exactitud posicional y temática, por ello tiene sentidos la comparación.

La completitud al igual que los resultados encontrados de manera univariada no se vieron afectados, pues todos los cálculos parten de la existencia de la geometría. A continuación, se comparten los resultados de la completitud para la malla vial de Bogotá:

La completitud se evaluó en términos generales, sobre toda la ciudad de Bogotá y particularmente sobre cada una de las localidades que la conforman. Como se puede observar en la **Tabla** (5.11), se encontró que *OSM* ha omitido 1183,29 Km, lo que equivale a un 12,6% menos respecto a los *Km* reportados por *IDECA*. Por otra parte, el conteo derivado de la comparación automática, mostró que *OSM* contiene -14.6% elementos lineales que los encontrados en *IDECA*.

Tabla 5-12: Tabla datos Omisión.

Resultados Completitud nivel Bogotá				
Variables	IDECA	OSM	Delta	% Omisión
KM	9379,21	8195,91	1183,29	12,6%
Elementos	136958	116964	199,94	14,6%

Analizando los datos por comisión se encontró un exceso en valor de comisión de 36.42 *Km* correspondientes a 1266 objetos lineales en OSM. En términos de porcentaje esto correspondió respectivamente al 0,44 y 1,11%. El resumen de los valores de comisión es

expuesto en la **Tabla** (5.11). Se resalta que el valor real de datos contenidos en OSM fue de 114183, debido a que se le eliminaron los valores de comisión.

Tabla 5-13: Tabla datos Comisión.

Resultados Completitud nivel Bogotá				
Variables	IDECA	OSM	Delta	% Comisión
KM	9379,21	8195,91	36.42	0,44%
Elementos	136958	116964	1266	1,11%

Los resultados por localidad serán expuestos y explicados a continuación. Para poder realizar el análisis de los resultados por localidad se muestra la tabla 5.14 que hace referencia a la combinación de errores encontrados:

Tabla 5-14: Combinación de errores por localidad

Combinación de errores encontrados en los clústeres	
1E_JER_GQ	1 error: Jerarquia vial y buena calidad
1E_NV_GQ	1 error: nombramiento vial y buena calidad
1E_SV_GQ	1 error: sentido vial y buena calidad
2E_JER_NV_GQ	2 errores: jerarquía vial -sentido vial y buena calidad
2E_JER_SV_GQ	2 errores: jerarquía vial - sentido vial y buena calidad
2E_SV_NV_GQ	2 errores: sentido vial -nombramiento vial y buena calidad
3E_JE_SV_NV_GQ	3 errores: jerarquía vial sentido vial- nombramiento vial y buena calidad
N_E_GQ	Ningún error y buena calidad
…	Todas las combinaciones de error siguen el mismo patrón

La localidad Antonio Nariño mostró que el 60.3% de sus datos tenían un error en cuando a jerarquía, sentido y nombramiento vial. Este error, como se muestra en la **Figura** (5.29) se descompone de la siguiente manera. El 3% corresponde a un error múltiple dado por

sentido vial, jerarquía, nombramiento y exactitud posiciona media. El 16% de los errores encontrados estaban relacionados con la jerarquía vial, 13% se encontraban relacionados con Jerarquía vial y nombramiento, 10.7% a sentido vial, 4.6% relacionado a Jerarquía y Sentido vial y 4.6% a errores en sentido vial y nombramiento.

Datos de calidad por localidad

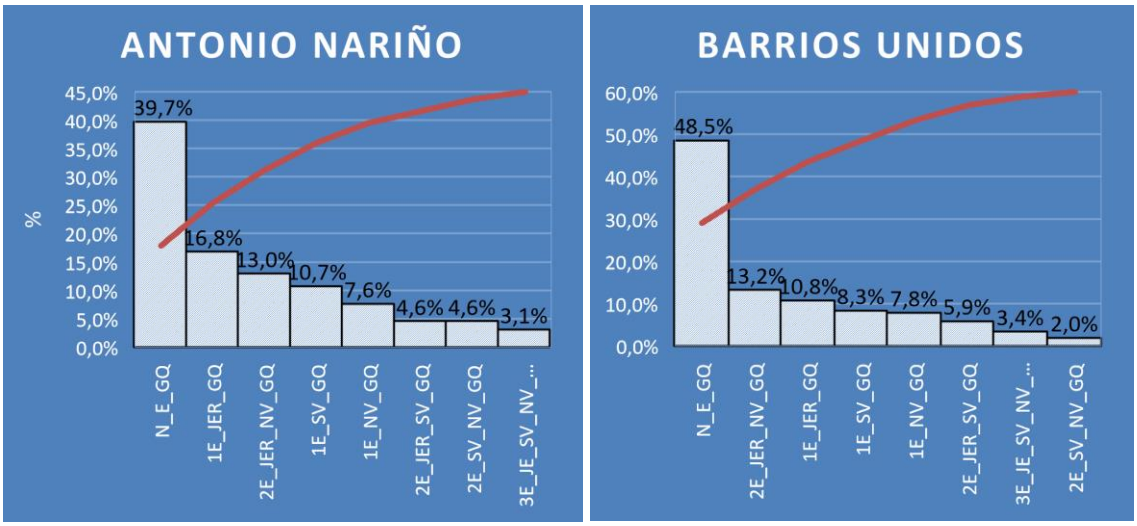


Figura 5-35: Errores por localidad.

Note que estos valores no son excluyentes, es decir, si existe 16.8% de error en jerarquía vial, entonces los errores encontrados que hacen referencia a Jerarquía + sentido vial(2E_JER_SV_GQ) deberían restar al valor de error en Jerarquía total (1E_JER_GQ), puesto que un porcentaje de 16.8% se encuentra representado en los errores combinados de jerarquía + nombramiento vial. Para ser exactos el error real de jerarquía (Error no combinado con otros) es de 13.0% mientras que el error de jerarquía + nombramiento vial es de 13%, donde 9.2 corresponde únicamente a errores de nombramiento y el 3.8 hace referencia a errores de jerarquía. Se hace énfasis en que los errores calculados no son excluyentes.

Barrios unidos mostró un porcentaje de error de 51.5%, donde el 13.2% representaba errores de exactitud temática- jerarquía y Nombramiento vial.

Datos de calidad por localidad

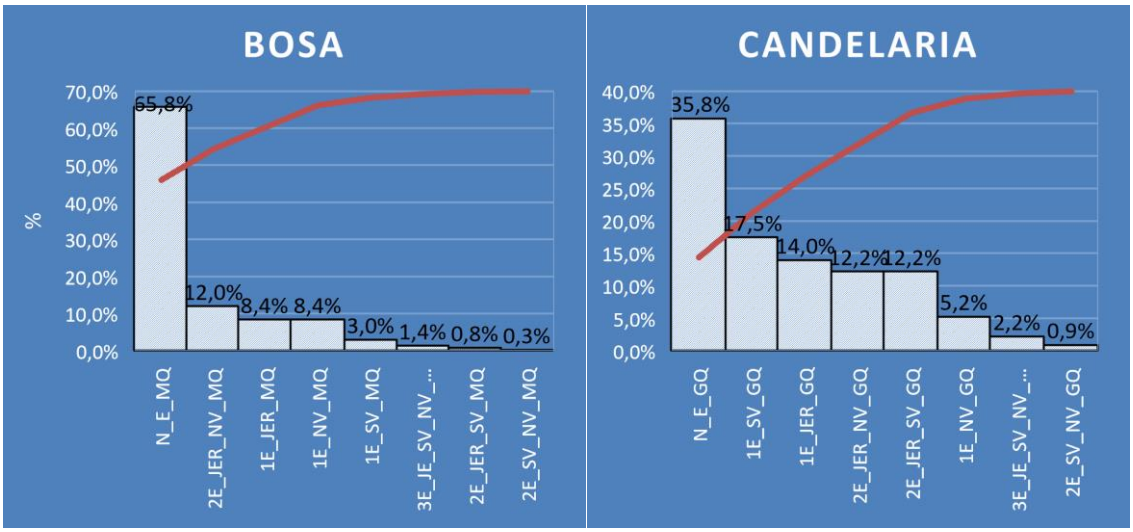


Figura 5-36: Errores por localidad.

Respecto a Bosa y Candelaria se puede observar que el 34.2% y 64.2% de errores fueron encontrados, donde para Bosa el mayor aporte en % de error fue dado por jerarquía y nombramiento vial, con una exactitud posicional catalogada como media, mientras que para Candelaria fue Sentido vial con una buena exactitud posicional.

Datos de calidad por localidad

Figura 5-37: Errores por localidad.

Chapinero mostró un error del 64.9%, el porcentaje que mayor aporto a este valor fue dado por errores combinados de jerarquía y nombramiento vial (20.6%) seguido de errores individuales en jerarquía 15.4%. Ciudad Bolívar contó con errores alrededor de 60%, una de las localidades con mayor porcentaje de error, Jerarquía vial y nombramiento fueron las variables que más aportaron a este porcentaje- aquí el porcentaje de error relacionado a sentido vial fue del 5%, mientras que en el cálculo univariado fue de solo un 3.4%.

Para el resto de las localidades, Engativá, Fontibón, Kennedy, Los Mártires, Puente Aranda, Rafael Uribe Uribe, San Cristóbal, santa fe, Sumapaz, suba, Teusaquillo y Tunjuelito, los porcentajes de error encontrados fueron: 47.4%, 63.4, %, 41.2%,65,8%, 63,3%,59,5%, 98,2% (Sumapaz), 55,2%, 66,3%, 69%, 51,2%, 52,1%. Las gráficas junto a todos los porcentajes de error que componen los valores expuestos, se encuentran en las figuras siguientes:

Datos de calidad por localidad

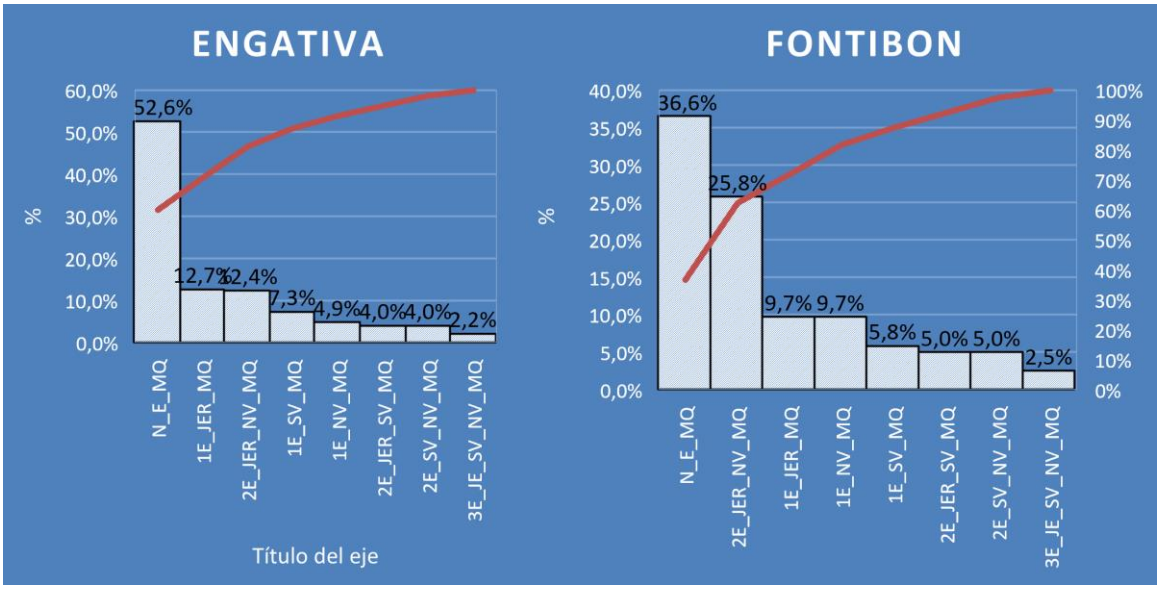


Figura 5-38: Errores por localidad.

La localidad de Engativá muestra que el 56% de los datos no poseía ningún error (N_E) y en términos generales la calidad relacionada con la exactitud posicional fue buena.Para esta localidad se encontraron erros combinados de jerarquía vial y nombramiento (12.3%)

La localidad de Fontibón presenta más errores en jerarquía vial y sentido vial (14.3%)

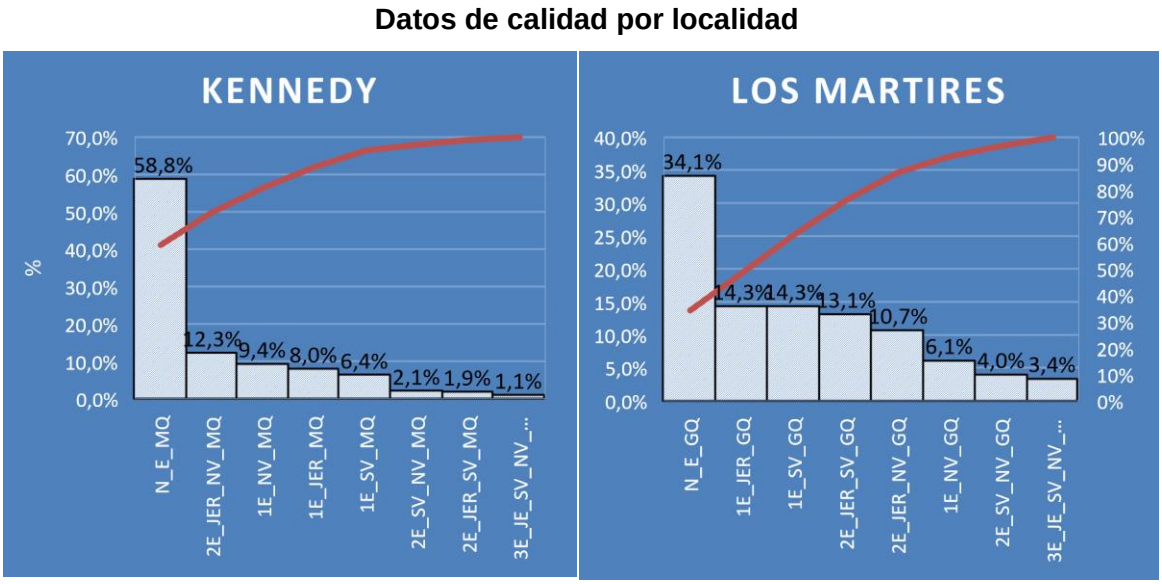


Figura 5-39: Errores por localidad.

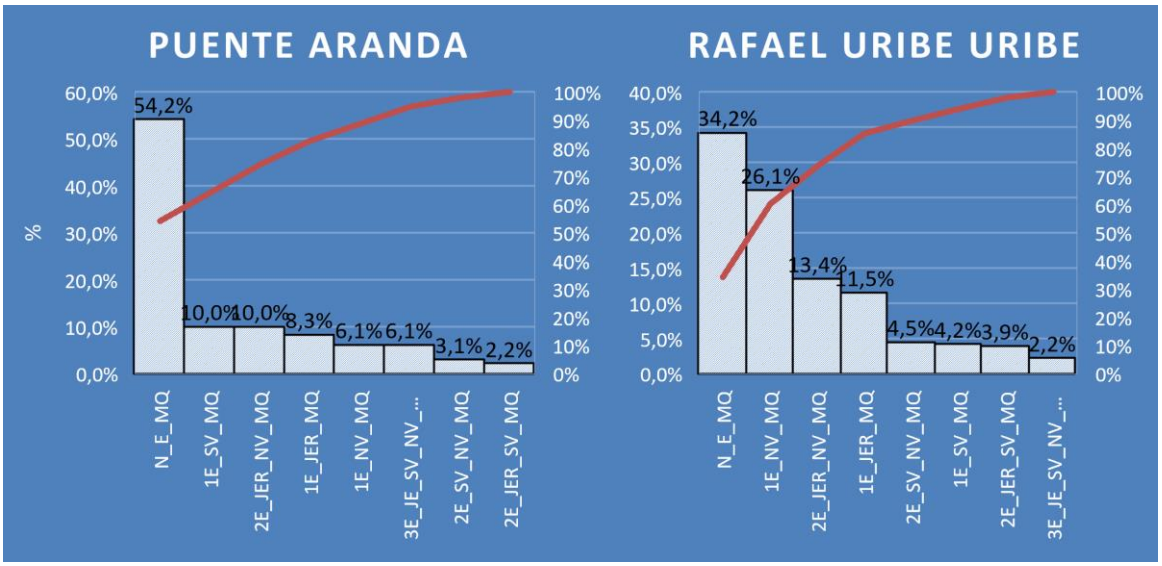


Figura 5-40: Errores por localidad.

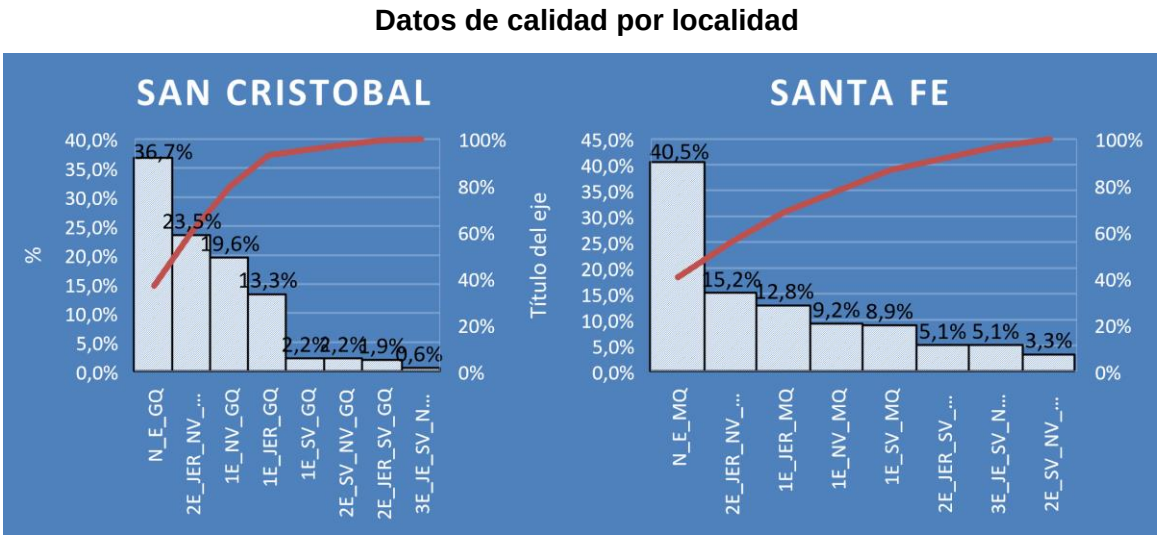


Figura 5-41: Errores por localidad.

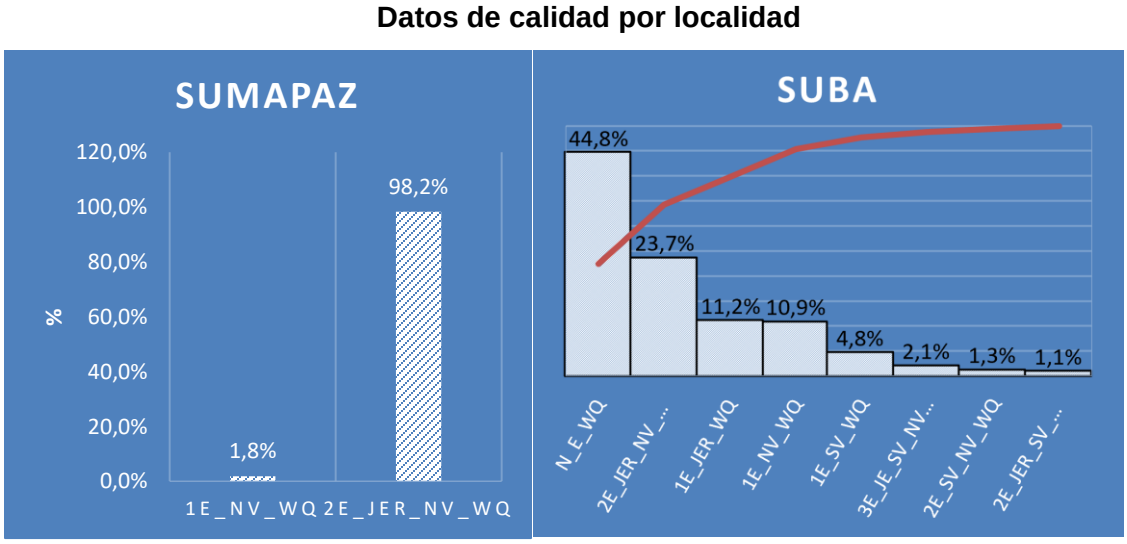


Figura 5-42: Errores por localidad.

Los resultados de calidad VGI combinando la aparición de errores muestran que la localidad de Sumapaz contiene el 98.2% de error en jerarquía vial JER + Nombre vial NV donde la calidad de exactitud posicional es una de las peores encontradas WQ. Este valor concuerda con las conclusiones de (Haklay, 2010) y(Koukoletsos, Haklay, & Ellul, 2012) donde las zonas rurales son aquellas con mayor déficit y calidad en la información VGI.

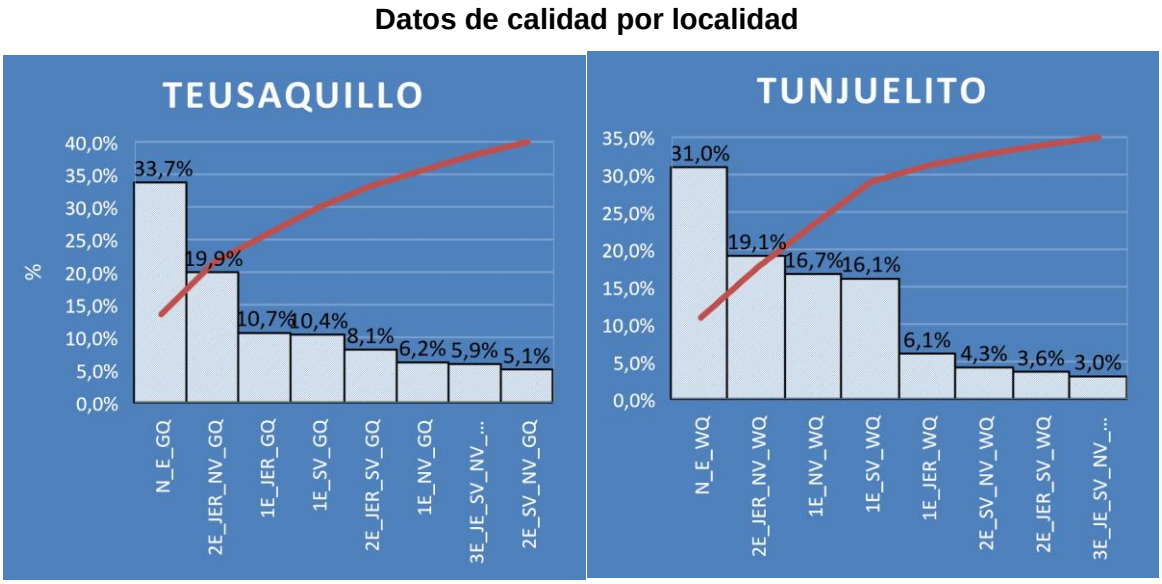


Figura 5-43: Errores por localidad.

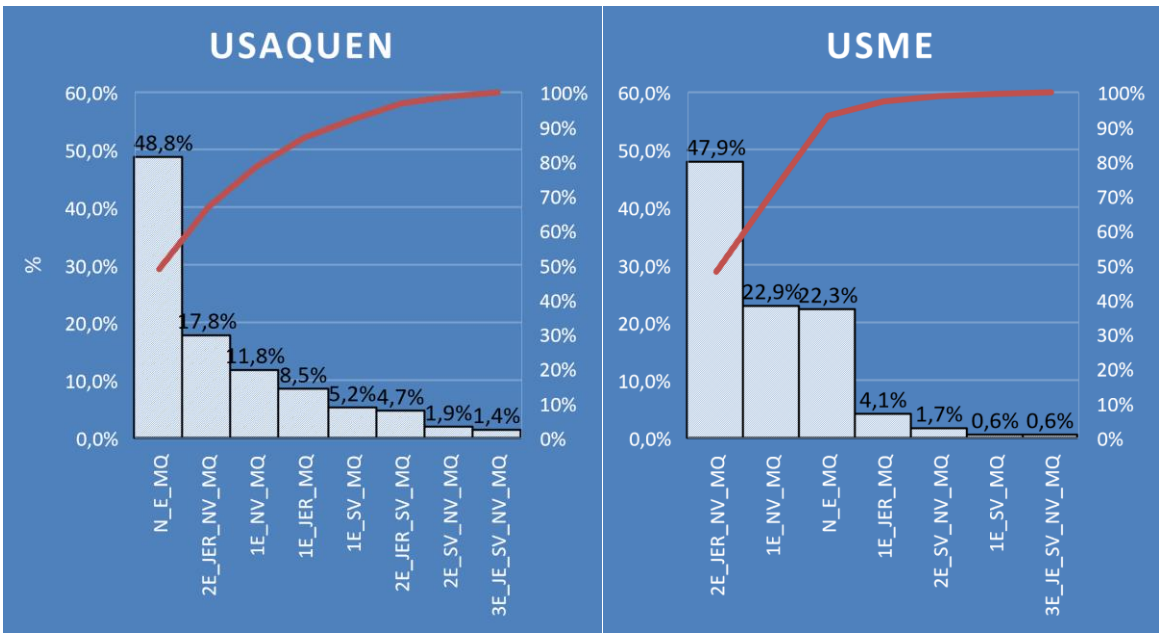


Figura 5-44: Errores por localidad.

La siguiente tabla, representa el promedio de los errores encontrados en función de la exactitud temática jerárquica, sentido vial y nombramiento analizados univariadamente, si se comparan con los errores promedio expuestos en esta sección anterior para cada una de las localidades, se puede concluir que los % de error multivariados exceden a casi todos los valores de error promedio univariados. Por lo que en términos de calidad VGI, el análisis realizado es mucho más pesimista según los resultados encontrados.

Tabla 5-15: Promedio porcentaje de errores análisis univariados [25].

Localidad	ANTONIO NARIÑO	TUNJUELITO	RAFAEL URIBE URIBE	CANDELARIA	BARRIOS UNIDOS
% Error promedio	30,1	31,9	31,8	37,2	31,1
Localidad	TEUSAQUILLO	PUENTE ARANDA	LOS MARTIRES	SUMAPAZ	USAQUEN
% Error promedio	34,8	25,6	33,5	66,2	26,5
Localidad	CHAPINERO	SANTA FE	SAN CRISTOBAL	USME	CIUDAD BOLIVAR
% Error promedio	34,4	33,4	28,5	43,6	28,5
Localidad	BOSA	KENNEDY	FONTIBON	ENGATIVA	SUBA
% Error promedio	16,0	19,1	34,5	23,2	26,7

6. Discusión

Respecto al objetivo #1, el cual consistió en:

Analizar y estandarizar las relaciones existentes entre los atributos OSM y IDECA.

Para estandarizar y analizar los datos además de realizar la exploración directa de la fuente, fue necesario comprender bajo que reglas fueron codificados, para ello este proyecto intento analizar la mayor cantidad información posible respecto a metadatos y guías de mapeo usadas como han sugerido algunos autores(Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2016b). gracias a esto se logró determinar que la clase *highway de OSM*, la clasificación *motorway* no debe ser codificada en Colombia, ya que OSM estableció que estas vías no existen en el país. Esto efectivamente es correcto pues no contamos con vías de acceso controlado(Alcaldía Mayor de bogotá, 2018), por lo que toda codificación encontrada con esos valores fue contada como un error. También se encontró que las carreteras nacionales troncales son asignadas a la etiqueta highway=trunk. Según OSM las vías NQS, Calle 80, Autopista Norte, Avenida Eldorado representan esta categoría lo cual no es totalmente cierto sí se examina y compra con la clasificación vial que tiene IDECA(C et al., 2012). En términos generales se encontró que OSM aún tiene problemas en la documentación para la codificación de datos de jerarquía vial, tal como afirma(Loai Ali, 2016) :*OSM aún maneja términos ambiguos y poco claros a la hora de realizar recomendaciones de codificación a los usuarios.*

Una vez analizados los datos se procedió a crear una estructura que permitiera la estandarización de estos, esta estandarización dio como resultado un modelo entidad relación que permitió entender qué tipo de variables y que clases debían ser estandarizadas de manera similar a como fue realizado por (Nowak Da Costa, 2016b) en *Towards Building Data Semantic Similarity Analysis:OpenStreetMap and the Polish Database of Topographic Objects pag.. 273*. Aquí también se propuso una alteración en la clasificación de datos mediante la fusión de algunas de las calases existentes para

minimizar y optimizar la comparación entre las fuentes de datos obteniendo buenos resultados.

Como resultado se muestra el modelo generado donde se ilustra las relaciones entre tablas de atributos.

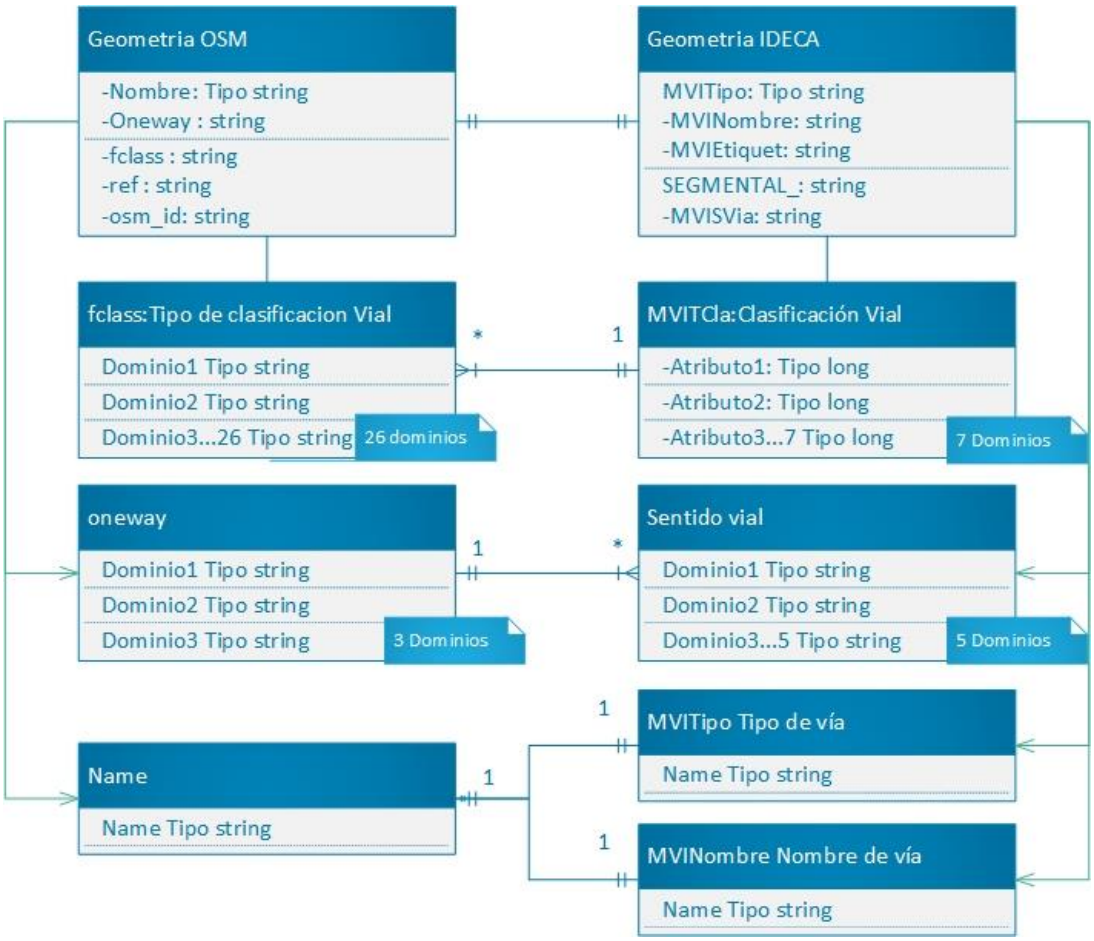


Figura 6-1: Modelo entidad relación.

Objetivo #2, el cual consistía en:

Examinar y completar los datos de texto VGI para la estandarización.

Muchos de los datos alojados en OSM referentes al nombramiento de las vías tenían problemas asociados con espacios, texto incompleto etc., esto puede ser causado por múltiples factores entre ellos el descuido de los usuarios, la diversificación con que fueron colectados los datos o incluso ediciones deliberadamente codificadas de manera errónea para afectar las bases de datos (Vandalismo informático) OSM (Neis & Zielstra, 2014). Para no perder esta información y lograr una mejor comparación se usaron expresiones regulares donde se seleccionan y se estandarizar los datos por medio de patrones. Trabajos similares fueron desarrollados por (Ballatore et al., 2013) donde en lugar de texto se trabajaron etiquetas para reclasificar elementos de tipo geográfico ayudando a reducir el ruido y la ambigüedad de la información OSM. Al igual que en el trabajo de Ballatore, aquí también se consiguieron buenos resultados pues casi 85.125 registros fueron estandarizados, lo que corresponde a un 73% de del total de los datos.

Respecto al objetivo #3, el cual consistía en:

Comparar y analizar semi automáticamente la información usando muestras estratificadas para clasificar de manera multivariada su calidad.

Por medio del matching por objeto más cercano, el cual consistía en un buffer móvil que dependía de la distancia al nodo más próximo, se lograron comparar 114.183, registros, lo que significó un match de 85.42%. Muchos nodos no se encontraron cerca de sus homólogos, pues alguna geometría en *OSM* tenía desplazamientos. Solo el 14.58% de los datos no pudo ser unido debido a problemas con geometría. Comparando este pareo semi automático con el de otros autores como(Zielstra & Zipf, 2010), aparentemente el método empleado permite obtener resultados más precisos pues Ziliestra realizaba la comparación usando una grilla con un determinado tamaño mientras que aquí la comparación se realizó uno a uno con un porcentaje de fracaso relativamente bajo (Aprox 15%). Por otro lado, comparando este trabajo con el de autores que realizaron pareo automático como lo son

(Steffen Volz, 2006) y (Abdolmajidi 2015) se puede observar que el método empleado aquí es de menor calidad que del autor previamente citado pues aún no se superan los errores de unión en geometrías complejas, sin embargo es mucho más rápido en términos de procesamiento y es de mejor calidad que la metodología de buffer estático usada con tradicionalmente(M. F. Goodchild & Hunter, 1997; Janelle, D.G. and Goodchild, 2011). Este tipo de errores aparecen sobre aquellos elementos complejos como deprimidos, glorietas y en términos generales sobre vías con geometría complejas. Este tipo de problemas ya habían sido reportados en la literatura(Michael F. Goodchild & Hunter, 1997). El problema del match de geometría usando el nodo central radico en que no todos los nodos pudieron ser localizados cerca a su par.En las siguientes figuras se puede observar en color negro, los nodos que lograron hacer match y por ende pudieron comprar los atributos objeto de estudio **Figuras** (5.13), en color verde se puede observar la geometría IDECA y en color rojo OSM. Nótese que, para esta región, el match fue muy bueno en las vías principales pero algunas vías internas no consiguieron rastrear el objeto.

En la **figura** 5.13, en el cuadro color azul, se muestra una coincidencia perfecta, pero tiene una particularidad. El algoritmo no evalúa topología. Aunque en este ejemplo en particular se pudo comparar el nodo central con su homónimo, esto no quiso decir que la topología este correcta. Los cuadros color naranja muestran dos estados, primero resaltando que OSM tenía geometría que IDECA no, y que por ende no podía tener ningún match y segundo, muestra buenos resultados en segmentos de línea cortos.

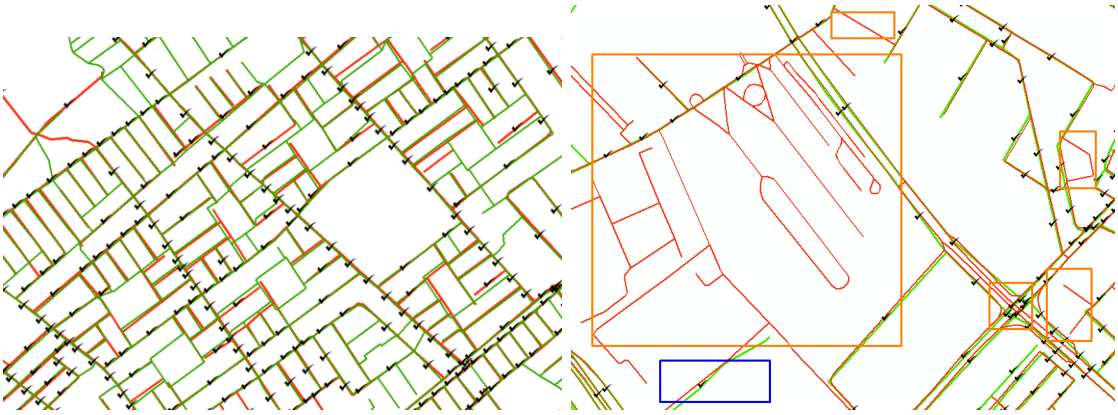


Figura 6-2: Proceso de comparación usando nodos centrales

En ALGUNAS geometrías complejas donde no había superposición de nodos la comparación fue exitosa:

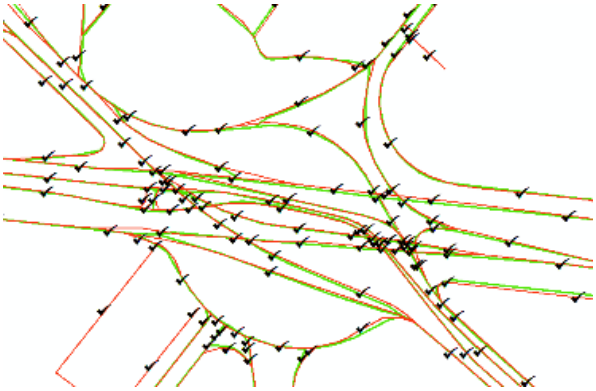


Figura 6-3: Proceso de comparación automática de datos.

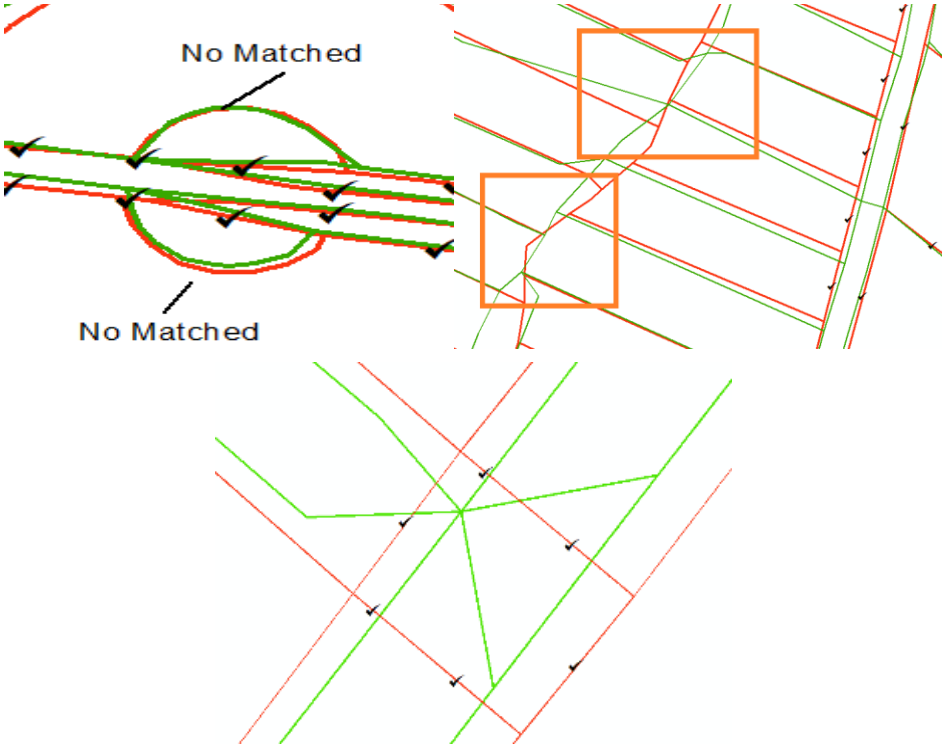


Figura 6-4: Problemas con el match en algunas geometrías.

Los problemas en el match como los mostrados en la figura 5.15 NO pudieron ser resueltos automáticamente por ello el 14.58 % de la información tubo que ser analizada manualmente.

En términos de rendimiento, se invirtieron 1.5 horas para realizar el mach geométrico y 30 minutos por atributo para realizar la comparación de campos, lo que significó un total de 3 horas de procesamiento, otros algoritmos invierten hasta 5 hora en la ejecución de esta tarea, como lo reporta (Abdolmajidi et al., 2015).

Completitud

Los valores de completitud encontrados sugieren que los datos VGI de OSM tenían pocas diferencias respecto a la fuente IDECA pues la discrepancia que se encontró en valores de omisión fue de apenas el 12% mientras que el de geometría por exceso equivalió a 1.11%. Este resultado es insignificante pues las diferencias siguen siendo relativamente pequeñas, efectos similares fueron encontrados por (Zielstra & Zipf, 2010) donde se encontraron diferencias alrededor de 7%.

Exactitud posicional y temática

Comparando el resultado obtenido (3.98m) con una investigación realizada en Inglaterra donde se obtuvo un error aproximado de 7.9 metros (Antoniou, 2011) y se concluyó que esta exactitud permitía comenzar a pensar que los datos OSM podrían ser usados por algunas agencias en su país lleva a plantear la misma inquietud para Colombia pues la exactitud encontrada fue mucho mejor que la mencionada en el estudio citado.

Respecto a la exactitud temática de la ciudad de Bogotá, se encontró que: el 31.85% de los datos se estaban mal clasificados. La revisión de literatura ha mostrado que estos valores fluctúan, algunas veces se encuentran buenos resultados (Correcta clasificación) de hasta un 95%(Dorn et al., 2015) y otros con hasta el 52% de error(Mooney, Corcoran, & Winstanley, 2010b). La variabilidad de estos resultados puede deberse a la heterogeneidad de los contribuidores, la diversa cantidad de metodologías usadas para

codificar los datos y hasta puede depender de recomendaciones ambiguas realizadas por la misma plataforma.

Tabla 6-1: Resumen de la evaluación VGI univariada.

Evaluación de la calidad VGI calculada de forma univariada					
Malla vial	%Error Jerarquía Vial	%Error Sentido vial	%error Nombramiento	Promedio% error	Exactitud posicional. m
% error	40,2	17,2	38,16	**31,85**	3.74
% sin error	59,8	82,8	61,84	68,15	3.74

Respecto al objetivo #4, el cual consistía en:

Agrupar y clasificar los resultados mediante técnicas multivariadas.

Tabla 6-2: Resumen de la evaluación VGI bajo el análisis Multivariado VGI.

Malla vial	N_E_GQ	N_E_MQ	N_E_WQ	Promedio % error	Exactitud posicional
% sin error	13,4%	22,4%	3,8%	39,7%	3.97
% error	86,6%	77,6%	96,2%	**60,3%**	3.97

La calidad VGI promedio, evaluada de manera multivariada dio como resultado que el **60.3%** de los datos poseían al menos un error y que en promedio su exactitud posicional era de 3.97m. En contraste, los datos obtenidos midiendo la calidad VGI de manera univariada (ver **Tabla** (5.9)), mostraron que el promedio de error hallado para las 20 localidades fue de **31.85%** con una exactitud posicional promedio de 3.74 m. Casi 40% menos error que el reportado en el análisis multivariado. En términos generales los valores encontrados de manera multivariada son mucho más pesimistas que los encontrados de manera univariada. La gran diferencia entre estos dos resultados se debe a que un nodo con más de un error tiene cuenta con un mayor peso en la aglomeración del cluster que los demás.

Los resultados también muestran una relación de error entre una mala clasificación vial y un nombramiento errado encontrando muchas veces esto dos errores en un mismo nodo. Este resultado es difícil de comparar con la literatura pues aparentemente no existen estudios que hayan realizado de este tipo.

La completitud al igual que los resultados encontrados de manera univariada no se vieron afectados, pues todos los cálculos parten de la existencia de la geometría.

7. Conclusiones y recomendaciones

La evaluación de la calidad VGI a través del enfoque multivariado dio como resultado un error promedio de 60.3%, con una exactitud posicional de 3.98 m y una completitud invariante de 13% respecto a los valores de omisión y 1% respecto a los valores de comisión.

Para robustecer los resultados, la calidad VGI fue evaluada de forma univariada y multivariada. Esto permitió concluir que, en todos los escenarios, los valores multivariados fueron más pesimistas en cuanto el porcentaje de error encontrado. Pues para el caso univariado se halló un porcentaje de error promedio de 31.85 vinculado con una exactitud posicional promedio de 3.98m, mientras que el valor encontrado multivariadamente fue de 60.3 % y 3.97m. En consecuencia, la calidad VGI de manera multivariada es considerada deficiente respecto a los niveles de calidad mínima establecidos en el modelo de datos de IDECA.

Los resultados sugieren que los atributos Jerarquía y Nombre vial aparentemente se encuentran fuertemente relacionados, pues casi siempre sus errores aparecen de manera conjunta. Se sugiere tener presente solo uno de ellos al evaluar la calidad conjunta de los datos VGI.

Con solo el 50% de la información extraída de las 2 dimensiones analizadas en el ACM se pudo encontrar un error del más del 60% en los datos VGI. Para posteriores estudios se recomienda usar más dimensiones donde posiblemente se puedan ver más interacciones entre las variables.

La estandarización de los datos usando el modelo entidad relación fue exitoso, sin embargo, para mejores resultados la categoría servicios de OSM identificada en la en el análisis de exploración de los datos debería ser analizada por separado, debido a que la geometría asociada a este tipo de jerarquía genera problema en el mach de los nodos.

El uso de expresiones regulares para mejorar las cadenas de caracteres relacionadas al nombramiento vial permitió ajustar 85125 registros de acuerdo a su contexto dentro de los datos. Esto permitió generar una comparación automática más confiable. Cadenas de caracteres mayores a 20 elementos NO pudieron ser reparadas debido a la complejidad para encontrar un patrón, por ello se recomienda continuar con el estudio de este tipo de elementos.

La comparación semi automática mostro un porcentaje de acierto cercano al 85% de los datos, permitiendo concluir que la metodología planteada usando un buffer móvil es viable para la comparación de grandes bancos de datos.

El proceso de comparación semiautomático empleado aquí, permitió cotejar las fuentes y generar los cálculos para medir la calidad VGI, sin embargo, por no ser el objeto pleno de esta investigación aún queda por desarrollar las mejoras algorítmicas, tales como match sobre geometrías complejas y la escalabilidad de la herramienta. Sin embargo, se resalta como se mencionó en el capítulo de resultados, que el algoritmo de buffer móvil empleado aquí invirtió 3h de procesamiento lo cual implica un nivel de procesamiento más rápido que el de algunas metodologías citadas en este documento las cuales invertían hasta 5h con una cantidad similar de datos.

Se recomienda analizar los datos relacionados a líneas de servicio (Transmilenio) por separado, pues en términos de pareo automático, fue una de los atributos más difíciles de monitorear debido a la cercanía de estas vías con otras de mayor jerarquía.

El muestreo generado permitió evaluar cómodamente los resultados de calidad en el análisis multivariado pues permitió analizar de mejor manera las coordenadas de los individuos con sus respectivas categorías y variables, sin embargo, se recomienda trabajar con todos los datos de la población con el fin de obtener mayor poder de interpretación respecto a la calidad VGI. En cuanto a la agrupación y visualización de resultados usando análisis de conglomerados podemos concluir que son técnicas muy buenas para ver la asociación de elementos, pero cuando se trata de exponer los resultados numéricamente se quedan cortas pues no permiten añadir al análisis visual elementos que permitan concluir de mejor forma los resultados, fue por este motivo que en el capítulo de resultados,

la información de los clústeres creados fueran extraída para la creación de graficas un poco más comprensibles.

En términos generales se encontró que OSM aún tiene problemas en la documentación para la codificación de datos de jerarquía vial, tal como afirma (Loai Ali, 2016) :OSM aún maneja términos ambiguos y poco claros a la hora de realizar recomendaciones de codificación a los usuarios.

Se recomienda hacer uso de los metadatos como variable de control para verificar la calidad VGI.

Se recomienda realizar este tipo de estudios utilizando la variable temporal pues estos datos van cambiando constantemente en el tiempo, producto de mejoras en los procesos de colección y validación de los datos. Esto hace necesario entender cómo está cambiando la calidad VGI a través del tiempo.

La exactitud posicional encontrada estuvo por debajo de los 4 metros de error, por ello es posible que la información VGI referida a calidad posicional sea usada por entidades gubernamentales como complemento a la información oficial.

OpenStreetMap es una fuente gratuita, que a menudo es cuestionada por muchos investigadores debido a los problemas relacionados con la heterogeneidad de los datos. Algunos autores sugieren integrar los datos VGI a fuentes oficiales. Los resultados aquí encontrados mostraron que algunas medidas de calidad en los datos VGI como lo son exactitud posicional y completitud permiten apoyar esta idea. La evaluación por exactitud posicional estimada y completitud en este documento permite afirmar que los datos VGI correspondientes a la malla vial de Bogotá pueden ser usados como datos de referencia para entidades oficiales.

Al conocer la calidad de datos VGI para la malla vía de Bogotá se incrementa la posibilidad de poder incorporar de manera confiable aquella información que posiblemente cumpla con los requisitos de calidad aceptable para entidades gubernamentales que requieran información geográfica libre de pago.

Al igual que (Haklay, 2010) en este trabajo también se concluye que una comparación por conteo no es suficiente para describir las diferencias entre dos fuentes de datos.

La investigación relacionada a calidad VGI no es una tarea trivial, debido a la diversidad con que los datos son colectados pues cada vez existen más contribuyentes y fuentes que alimentan este tipo de datos. Todos los investigadores citados en este documento concuerdan con ello, por ello; se plantea la necesidad de encontrar mejores y nuevas metodologías que permitan evaluar cuantificar y calificar la calidad VGI.

A. Anexo: Patrones creados a partir de expresiones regulares.

B. Anexo: Tabla comparación de atributos.

C. Anexo: Muestreo por estrato.

D. Anexo: Aportes individuales de las variables y Eigenvalores.

E. Anexo: Clústeres

F. Anexo: Coordenadas de individuos.

G. Anexo: Código R

H. Anexo F “Graficas ACM”

Bibliografía

Abdolmajidi, E., Mansourian, A., Will, J., & Harrie, L. (2015). Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-Spatial Information Science*, *18*(2–3), 65–80. https://doi.org/10.1080/10095020.2015.1071065

Alcaldía Mayor de bogotá. (2017). *Resumen Del Diagnóstico General Plan De Ordenamiento Territorial De Bogotá*. Bogotá. Retrieved from http://www.sdp.gov.co/portal/page/portal/PortalSDP/POT_2016/diagnostico_general/201708 RD V3.0.pdf

Alcaldía Mayor de bogotá. (2018). Vías, Transporte y Servicios Públicos. Retrieved from http://www.sdp.gov.co/gestion-territorial/vias-transporte-y-servicios-publicos/vias

Antoniou, V. (2011). *User Generated Spatial Content: An Analysis of the Phenomenon and its Challenges for Mapping Agencies*. *Doctoral thesis*. Retrieved from http://discovery.ucl.ac.uk/1318053/

Antoniou, V., & Skopeliti, A. (2015). MEASURES AND INDICATORS OF VGI QUALITY: AN OVERVIEW. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, *37*(1), 61–81. https://doi.org/10.1007/s10115-012-0571-0

Ballatore, A., & Zipf, A. (2015). A Conceptual Quality Framework for Volunteered Geographic Information. In *Spatial Information Theory* (Vol. 9368, pp. 89–107). https://doi.org/10.1007/978-3-319-23374-1

Beh, E. J. (2012). Exploratory multivariate analysis by example using R. *Journal of*

Applied Statistics. https://doi.org/10.1080/02664763.2012.657409

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2014). A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Information Sciences*, *258*, 312–327. https://doi.org/10.1016/j.ins.2013.07.013 10.3390/ijgi2030680, ISPRS; De Caluwe, R., De Tré, G., Van Der Cruyssen, B., Devos, F., Maesfranckx, P., Time management in fuzzy and uncertain object-oriented databases (2000) Knowledge Management in Fuzzy Databases, , O. Pons, A. Vila, J. Kacprzyk, Physica-Verlag Heidelberg pp. 67-88; De Tré, G., Consistently Handling Geographical User Data Context-Dependent Detection of Co-located POIs (2010) Proceedings of the Int. Conf. on Information Processing and Management of U

Bridges, C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports*. https://doi.org/10.2466/pr0.1966.18.3.851

c, O. N., & Greenacre, M. (2007). Correspondence Analysis in *R* , with Two- and Three-dimensional Graphics: The **ca** Package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v020.i03

C, R. D. E. B. D., Distrital, S., Sdm, D. M., Álvarez, M. H., Benavides, P., Carlos, J., … Morales, Y. (2012). *Especificación Técnica para el Mapa de Referencia de Bogotá D.C.* Bogotá.

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., & Rau, R. (2011). Osmonto - An Ontology of OpenStreetMap Tags. *State of the Map*. Retrieved from http://www.informatik.uni-bremen.de/~okutz/osmonto.pdf

Degrossi, L. C., de Albuquerque, J. P., Fan, H., & Zipf, A. (2016). A conceptual model for quality assessment of VGI for the purpose of flood management. Retrieved from http://www.cs.nuim.ie/~pmooney/LinkVGI2016/AGILE_2016_LiviaCastroDegrossi_CameraReady.pdf

Dorn, H., Törnros, T., & Zipf, A. (2015). Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information*, *4*(3). https://doi.org/10.3390/ijgi4031657

Edition, T. (2003). Third Edition. *New York*. https://doi.org/10.1080/00049538408255324

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of*

the Association of American Geographers, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657

Esmaili, R., Naseri, F., & Esmaili, A. (2013). Quality Assessment of Volunteered Geographic Information. *American Journal of Geographic Information System*, *2*(2), 19–26. https://doi.org/10.5923/j.ajgis.20130202.01

Fairbairn, D., & Al-Bakri, M. (2013). Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, *2*(2), 349–370. https://doi.org/10.3390/ijgi2020349

FGDC. (1998). Geospatial Positioning Accuracy Standards Part 3 : National Standard for Spatial Data Accuracy. *Geospatial Positioning Accuracy Standards*, 28. Retrieved from http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*. https://doi.org/10.1007/s10708-008-9188-y

Fonte, C. C.; Antoniou, V.; Bastin, L.; Bayas, L.; See, L.; Vatseva, R. (2017). Assessing VGI data quality. *Citizen Sensor*, 137–163. https://doi.org/10.5334/bbf.g

Fonte, C. C., Bastin, L., Foody, G., Kellenberger, T., Kerle, N., Mooney, P., … See, L. (2015). VGI Quality Control. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 317–324. https://doi.org/10.5194/isprsannals-II-3-W5-317-2015

Fonte, Cidália C., Bastin, L., See, L., Foody, G., & Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2015.1018266

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., & Boyd, D. S. (2013). Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS*, *17*(6), 847–860. https://doi.org/10.1111/tgis.12033

Garcia, Juan; Segovia, C. (2014). Análisis de conglomerados (ll): El procedimiento Conglomerados jerárquicos. *Guia SPSS*.

Girres, J. F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Gondar, J. E. (2000). Análisis de cluster. *Artículos Estadísticos*.

Goodchild, M. F., & Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, *11*(3), 299–306. https://doi.org/10.1080/136588197242419

Goodchild, Michael F. (1996). Directions in GIS. In *Third International Conference/Workshop on Integrating GIS and Environmental Modeling* (pp. 1–15). Retrieved from http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/goodchild_michael/good.html

Goodchild, Michael F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*. https://doi.org/10.1007/s10708-007-9111-y

Goodchild, Michael F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, *3*(3), 231–241. https://doi.org/10.1080/17538941003759255

Goodchild, Michael F., & Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, *11*(3), 299–306. https://doi.org/10.1080/136588197242419

Goodchild, Michael F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Goodchild, Michael F. (1992). Geographical information science. *International Journal of Geographical Information Systems*, *6*(1), 31–45. https://doi.org/10.1080/02693799208901893

Graser, A., Straub, M., & Dragaschnig, M. (n.d.). Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph. *Transactions in GIS*, *18*(4), 510–526. https://doi.org/10.1111/tgis.12061

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. https://doi.org/10.1068/b35097

Haklay, M., & Weber, P. (2008). OpenStreet map: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Hudson-Smith, A., Batty, M., Crooks, A., & Milton, R. (2009). Mapping for the Masses

Accessing Web 2.0 Through Crowdsourcing. *Social Science Computer Review*, *27*(4), 524–538. https://doi.org/10.1177/0894439309332299

Ipeirotis, P. G., Provost, F., Sheng, V. S., & Wang, J. (2014). Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, *28*(2), 402–441. https://doi.org/10.1007/s10618-013-0306-1

ISO 19157:2013. (2013). Geographic information -- Data quality. *International Standard*. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=32575

ITAM. (2015). *Análisis Multivariado Un Manual para Investigadores*. *Diplomado en Estadística aplicada*.

Jackson, S., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, *2*(2), 507–530. https://doi.org/10.3390/ijgi2020507

Janelle, D.G. and Goodchild, M. F. (2011). Concepts, principles, tools, and challenges in spatially integrated social science. *The SAGE Handbook of GIS and Society*, 27–45. https://doi.org/10.4135/9781446201046.n2

Jaume, M. J. R., & Catalá, R. M. (2001). *Estadística informática: casos y ejemplos con el SPSS*. Universidad de Alicante. Retrieved from https://books.google.com.co/books?id=BIIOAgAACAAJ

Johnson, R. A., & Wichern, D. W. (1998). Applied Multivariate Statistical Analysis. *Pearson Education International*, 226–235. https://doi.org/10.1198/tech.2005.s319

Jokar Arsanjani, J., Mooney, P., Zipf, A., & Schauss, A. (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. *Lecture Notes in Geoinformation and Cartography*, (9783319142791), 37–58. https://doi.org/10.1007/978-3-319-14280-7_3

Jonietz, D., & Zipf, A. (2016). Defining Fitness-for-Use for Crowdsourced Points of Interest (POI). *ISPRS International Journal of Geo-Information*, *5*(9), 149. https://doi.org/10.3390/ijgi5090149

Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, *16*(4), 477–498. https://doi.org/10.1111/j.1467-9671.2012.01304.x

Kresse, W., & Danko, D. (2011). *Springer handbook of geographic information*. https://doi.org/10.1007/978-3-540-72680-7

Lewis, C. S. (n.d.). Análisis de correspondecia simple.

Loai Ali, A. (2016). Tackling the thematic accuracy of areal features in OpenStreetMap. *European Handbook of Crowdsourced Geographic Information.*, 113–129. https://doi.org/10.5334/bax

Longley, P. A., Goodchi, Ld, M. F., Maguire, D. J., & Rhind, D. W. (2011). *Geographical Information Systems and Science*. *City*. https://doi.org/10.2307/215736

Lukacs, M., & Bhadra, D. (2003). *Mastering Python Regular Expressions*. *Schriften des Forschungszentrum J*[illegible]*lich Reihe Energietechnik*. https://doi.org/10.1002/ejoc.201200111

Mahabir, R., Stefanidis, A., Croitoru, A., Crooks, A., & Agouris, P. (2017). Authoritative and Volunteered Geographical Information in a Developing Country: A Comparative Case Study of Road Datasets in Nairobi, Kenya. *ISPRS International Journal of Geo-Information*, *6*(1), 24. https://doi.org/10.3390/ijgi6010024

Marín, J. M. (2009). Análisis de Cluster y Arboles de Clasificación. *Springer*. https://doi.org/10.1007/BF01882344

McKone, K. E., Schroeder, R. G., & Cua, K. O. (2001). Impact of total productive maintenance practices on manufacturing performance. *Journal of Operations Management*. https://doi.org/10.1016/S0272-6963(00)00030-9

Montenegro Alvaro, P. C. E. (2005). Análisis de correspondencias de tablas de contingencia estructuradas. Bogotá. Retrieved from Análisis de correspondencias de tablas de contingencia estructuradas

Mooney, P., & Corcoran, P. (2012). The Annotation Process in OpenStreetMap. *Transactions in GIS*. https://doi.org/10.1111/j.1467-9671.2012.01306.x

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010a). Towards quality metrics for OpenStreetMap. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems GIS 10*, 514–517. https://doi.org/10.1145/1869790.1869875

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010b). Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS ’10* (p. 514).

https://doi.org/10.1145/1869790.1869875

Moss, K. (2012). The Entity-Relationship model. In *IEEE Global Engineering Education Conference, EDUCON*. https://doi.org/10.1109/EDUCON.2012.6201182

Murtagh, F. (2007). Multiple correspondence analysis and related methods. *Psychometrika*. https://doi.org/10.1007/s11336-006-1579-x

Neis, P., & Zielstra, D. (2014). Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet*. https://doi.org/10.3390/fi6010076

Nelli, F. (2015). *Python Data Analytics: Data Analysis and Science using pandas, matplotlib and the Python Programming Language*. *The expert's voice in Python*.

Nowak Da Costa, J. (2016a). Novel tool for examination of data completeness based on a comparative study of VGI data and official building datasets. *Geodetski Vestnik*, *60*(03), 495–508. https://doi.org/10.15292/geodetski-vestnik.2016.03.495-508

Nowak Da Costa, J. (2016b). Towards Building Data Semantic Similarity Analysis: OpenStreetMap and the Polish Database of Topographic Objects. In *Proceedings - 2016 Baltic Geodetic Congress (Geomatics), BGC Geomatics 2016* (pp. 269–275). https://doi.org/10.1109/BGC.Geomatics.2016.55

NSAI standards. (2013). EN ISO 19157:2013/A1. *Geographic Information - Data Quality - Amendment 1*, 1–146.

OSM. (2017). Introduction to the OpenStreetMap Project. Retrieved February 3, 2017, from https://github.com/mapbox/mapping/wiki/Introduction-to-the-OpenStreetMap-Project

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. *International Journal of Morphology*. https://doi.org/10.4067/S0717-95022017000100037

Peña, D. (2002). Análisis de datos multivariantes. *Book*, 515. https://doi.org/8448136101

R, J. G. (n.d.). Analisis de conglomerados, 1–21.

See, L., Comber, A., Salk, C., Fritz, S., Velde, M. Van Der, Perger, C., … Obersteiner, M. (2013). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts, *8*(7), 1–11. https://doi.org/10.1371/journal.pone.0069958

Senaratne, H., Mobasheri, A., Ali, A., Capineri, C., & Haklay, M. (2016a). A review of volunteered geographic information quality assessment methods. *Int J Geogr Inf Sci*,

1–29. https://doi.org/10.1080/13658816.2016.1189556

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2016b). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *8816*(June), 1–29. https://doi.org/10.1080/13658816.2016.1189556

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2016.1189556

Shock, D. E. (2003). Data vs . Information. *CAB Corner on Quality*.

Silberschatz, A. (Bell L., Korth, H. F. (Bell L., & Sudarshan, S. (Instituto Indio de Tecnología, B. (2002). *Fundamentos de bases de datos*. *Victoria*. https://doi.org/10.1017/CBO9781107415324.004

Stark, H.-J. (2010). Quality assessment of volunteered geographic information (vgi) based on open web map services and iso/tc 211 19100-family standards. *Geoinformatics*, *13*(7), 28–31. https://doi.org/10.1126/science.1107075

Steffen Volz. (2006). An Iterative Approach for Matching Multiple Representations of Street Data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Taylor, D. R. F. (2014). Some recent developments in the theory and practice of cybercartography: Applications in indigenous mapping: An introduction. *Modern Cartography Series*. https://doi.org/10.1016/B978-0-444-62713-1.00001-5

Unidad Administrativa Especial de Catastro Distrital -Infraestructura de Datos Espaciales para el Distrito Capital -Gerencia IDECA. (2013). *Procedimiento para evaluar y reportar la calidad de los datos espaciales*. BOGOTA.

van Oort, P. A. J. (2006). *Spatial data quality: from description to application*. *Production*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Spatial+data+quality+:+from+description+to+application#0

Walpole, R., Myers, R., Myers, S., & Keying, Y. (2012). *Probabilidad y estadística para ingeniería y ciencias*. *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Yanai, H., & Ichikawa, M. (2006). Factor Analysis. *Handbook of Statistics*. https://doi.org/10.1016/S0169-7161(06)26009-7

Yang, B., Zhang, Y., & Lu, F. (2014). Geometric-based approach for integrating VGI POIs and road networks. *International Journal of Geographical Information Science*, *28*(1), 126–147. https://doi.org/10.1080/13658816.2013.830728

Yoshida, H. (2013). Muestreo estratificado, 2–3.

Zhang, M., & Meng, L. (2008). Delimited stroke oriented algorithm-working principle and implementation for the matching of road networks. *Geographic Information Sciences*. https://doi.org/10.1080/10824000809480638

Zielstra, D., & Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *13th AGILE International Conference on Geographic Information Science* (Vol. 1, pp. 1–15). https://doi.org/10.1119/1.1736005