

PROCESO DE GEOCODIFICACIÓN DE DIRECCIONES EN LA CIUDAD DE MEDELLÍN, UNA TÉCNICA DETERMINÍSTICA DE GEORREFERENCIACIÓN DE DIRECCIONES

Jaime Andrés Vargas Benjumea

Gobernación de Antioquia
Ingeniero Civil
cvjavar@gmail.com

Daniel Horfan Álvarez

HyG Consultores S.A.S.
Magister en Ingeniería de Sistemas
dhorfan@gmail.com

(Tipo de Artículo: **Investigación**. Recibido el 17/01/2013. Aprobado el 21/05/2013)

RESUMEN

El presente artículo es un estudio empírico comparativo de los diferentes procesos de geocodificación que se realiza en la Alcaldía de Medellín para la georreferenciación de información primaria en todos sus procesos de gobernabilidad y desarrollo territorial, los cual son el soporte para la toma de decisiones y la planeación de los recursos públicos. Dentro de los procesos a analizar se encuentra los métodos determinísticos y probabilísticos que permiten por medio de la estandarización y normalización de las direcciones un resultado enmarcado en un entorno espacial que cumpla con criterios de calidad tanto en el porcentaje de acierto de las direcciones encontradas como en la exactitud posicional del resultado, el cual es utilizado para los diferentes tipos de análisis espaciales en temas de seguridad, planeación territorial, asistencia social, infraestructura de vías, salud, catastro, medio ambiente, educación, atención a víctimas del conflicto y pqrs (peticiones, quejas, reclamos y solicitudes).

Palabras clave

Estandarización de Direcciones, Geocodificación, Normalización de Direcciones, Malla Vial, Predios, Ubicación de Direcciones.

GEOCODING PROCESS IN THE CITY OF MEDELLIN, A DETERMINISTIC TECHNIQUE FOR ADDRESS LOCATION

ABSTRACT

This article is a comparative empirical study of different geocoding processes performed in the Municipality of Medellin for the georeferencing of primary information in all processes of Governance and Territorial Development, which are the support for decision-making and planning of public resources. Among the processes to be analyzed we have deterministic and probabilistic methods which allow, through the standardization and regulation of addresses, a result framed in a space environment that meets quality standards for both the success percentage of found addresses and positional accuracy of the result, which are used for different types of spatial analysis in Security, Territorial Planning, Welfare, Road Infrastructure, Health, Land Registry, Environment, Education, Assistance to victims of conflict and RGCAs (requests, grievances, complaints and applications).

Keywords

Address Standardization, Geocoding, Address Normalization, Street Grid, Parcel, Address Location.

PROCESSUS DE GÉOCODIFICATION DES ADRESSES DANS LA VILLE DE MEDELLÍN-COLOMBIE, UNE TECHNIQUE DÉTERMINISTE DE GÉORÉFÉRENCIEMENT DES ADRESSES

RÉSUMÉ

Cet article présente une étude empirique comparative des différents processus de géocodification qui sont réalisés dans la mairie de Medellín pour la géoréférencement d'information primaire dans tous ses processus de gouvernabilité et développement territorial, lesquels sont le soutien pour la prise de décisions et le planning des fonds publics. Entre les processus à analyser on trouve les méthodes déterministes et probabilistes qui permettent au moyen de la standardisation et la normalisation des adresses un résultat encadré dans un environnement spatial qui accomplit avec des critères de qualité dans le pourcentage de réussite des adresses trouvés ainsi que dans l'exactitude positionnelle du résultat, qui est utilisée pour les différents types d'analyse spatiale par rapport a la sécurité, le planning territorial, assistance sociale, infrastructure des voies, santé, cadastre, environnement, éducation, assistance aux victimes du conflit armé et des pétitions, plaintes, appels et demandes.

Mots-clés

Standardisation des adresses, géocodification, normalisation des adresses, réseau routier, propriétés foncières, trouvaille des adresses.

1. INTRODUCCIÓN

Los procesos de geocodificación son fundamentales en el análisis espacial de información, existe una variedad de disciplinas y aplicaciones en las que se tienen gran aplicabilidad, es el caso del sector salud, aplicaciones en seguridad, aplicación en ciencias políticas [1] y planeación territorial. En el mundo existen diferentes procesos de geocodificación acompañados de algoritmos que ofrecen una solución a la ubicación de direcciones, pero todos estos procesos se cualifican de acuerdo al poder de acierto que ofrecen estos procesos en encontrar una posición determinada. Existen productos comerciales que ofrecen estos procesos como funcionalidad de sus productos, es el caso de ESRI, GeoMedia y MapInfo con sus productos de escritorio y/o servidor que permiten análisis de información espacial; con las últimas tendencias de servicios en la nube, grandes proveedores de servicios geoespaciales se dedicaron a desarrollar plataformas en la web que permitieran el acceso masivo a estas tecnologías, es el caso de empresas como NavTech y TeleAtlas que ofrecen tecnologías avanzadas de geocodificación de acceso restringido de tipo comercial; al igual que Google Maps, MapQuest, Bing y Yahoo Maps que ofrece de forma libre y gratuita.

La ubicación de direcciones es un proceso que las empresas tanto del sector público como privado requieren para referenciar espacialmente variables de su negocio. En el caso de la Alcaldía de Medellín, la cual es una entidad pública del estado, sus organismos dentro de los procesos internos del negocio se enfrentan a la necesidad de referenciar espacialmente los distintos tipos de información que manejan en su interior. La diversificación de información y de sistemas de información obligó a la administración a implementar un proceso de sistematización que fuera transversal a la organización y que tomara diferentes fuentes de información con el fin de alimentar los procesos de ubicación de direcciones, tomando como fuentes de información el sistema de Catastro Municipal, las bases de datos de instalaciones de energía de la empresa de servicios públicos de la ciudad y la malla vial.

Actualmente, El municipio de Medellín cuenta con el servicio de geocodificación determinístico a través de web services y en su portal geográfico corporativo Mapas.Medellin, desarrollo implementado por la empresa HyG Consultores SAS.

Anterior a este proceso de consolidación transversal para la entidad, en lo que concierne al proceso de referenciar información del negocio sobre un entorno enmarcado en un espacio, se intentaron varias iniciativas de diferentes tipos, en el año 2006 la empresa Procalculo Prosis S.A. implementó un aplicativo de escritorio y web para la geocodificación de direcciones basado en los "address locator" de ArcGIS conocido como "GEOCODING", con el cual se

pretendía potencializar el nivel de servicios de ubicación tanto para los usuarios internos como en el geoportal web de la entidad.

El propósito de este artículo es mostrar (1) el caso de éxito de la ciudad de Medellín en la sistematización de un proceso de ubicación de direcciones que toma como base la información generada por la entidad; (2) el proceso de ubicación de direcciones desde el modelo de datos y el algoritmo de estandarización de direcciones; (3) resultados comparativos con otros modelos de ubicación de direcciones tanto en la completitud, exactitud en la ubicación espacial como en el rendimiento de procesamiento de la solución.

2. PRINCIPIOS DE LA GEOCODIFICACIÓN

El proceso de geocodificación o ubicación de direcciones no es solo el proceso de referenciar espacialmente un punto sobre un mapa, este proceso involucra varias disciplinas como teoría de estructuras de información, teoría de decisión, teoría de la probabilidad y la semántica de las palabras.

2.1. Componentes en el proceso de geocodificación

El proceso de geocodificación requiere de tres componentes básicos: Datos de referencia, datos descriptivos y el software que realiza el proceso. Los datos de referencias están compuestos por las diferentes fuentes de información geográfica basada en archivos, en esta categoría de información se encuentra tanto archivos codificados con coordenadas como elementos más complejos como son las estructuras de información espacial y la capa de líneas que representan las vías de la ciudad. Los datos descriptivos son la información candidata al proceso de geocodificación, generalmente es la entrada del proceso. El software es la herramienta sistémica que permite procesar la entrada y dar un resultado de acuerdo a las reglas definidas en el proceso [2].

2.2. Proceso de geocodificación

La geocodificación es el proceso de asignar un par de coordenadas XY, ya sea en sistema de coordenadas plana o geográfica de la descripción de referencia de un sitio localizado en un entorno espacial. En general la geocodificación es el proceso que inicia con la entrada de una descripción (ej. Dirección), se identifica la semántica de la descripción para descomponerla por componentes definidos en el algoritmo de normalización, se estandariza estos componentes para unificar y homologar las estructura de los componentes y posteriormente se realiza la búsqueda en una fuente de datos para realizar la comparación por componentes y entrega el mejor candidato de la búsqueda para mostrarlo sobre un mapa.

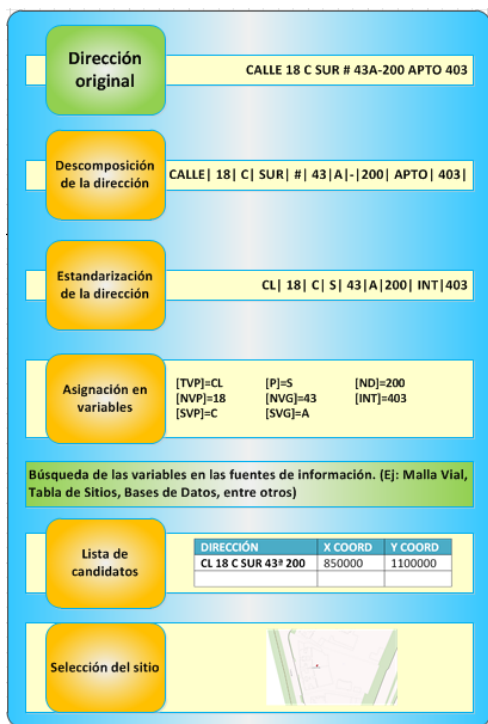


Fig. 1. Proceso Geocodificación para la ciudad de Medellín

El proceso de geocodificación se basa en el concepto de relacionamiento o vinculación de registros. Existen dos principales métodos de relacionamiento de datos, el determinístico y el probabilístico. El modelo de relacionamiento determinístico utiliza un identificador único que permite distinguir inequívocamente a la entidad (persona, familia, empresa, dirección, etc.), ordenando los registros y comparándolos con sus pares o no pares. Este modelo es simple de comprender e implementar, aunque en algunos casos, la solución de decisiones subjetivas, puede ser compleja en su solución. El relacionamiento probabilístico se basa en la teoría estadística desarrollada por Fellegi y Sunter [5], y es apropiado cuando las bases de datos a relacionar no tiene ningún identificador único en común.

2.3. Método determinístico o exacto (Deterministic record linkage)

Es la técnica o procedimiento más adecuado para ser utilizado cuando exista un identificador único y permita distinguir inequívocamente la entidad, siendo útil para relacionar conjuntos de bases de datos que contienen diferente información. Existiendo estos identificadores únicos, el problema de correlación de registros se vuelve trivial y la relación entre los datos es uno a uno (1-1) y así poder resolver el relacionamiento de datos. Para los autores que discuten el método determinístico, existe muy poca literatura sobre el tema, lo cual es un indicio de que el modelo es considerado simple de utilizar. Roos y Wajda [6] sugieren utilizar una medida llamada “Número medio de casos por bloque”, para estimar la cantidad de información relacionada en cualquier base de datos o archivo. Bous y e Scott presentaron una visión general de los dos métodos de

relacionamiento incluyendo una discusión o crítica al método determinístico [7]. En este tipo de relacionamiento, comparan dos registros, por ejemplo el nombre y el apellido, estos registros se consideran pares si los registros (nombre y apellidos) concuerdan en todos sus caracteres en las diferentes fuentes de información a relacionar. Un ejemplo claro del método determinístico o exacto es la vinculación de registros por medio de una clave primaria y foránea en una base de datos, sin importar el tipo de clave ya sea numérica o compuesta por caracteres; este es el caso de estructuras complejas como las direcciones.

El proceso consta de una serie de pasos los cuales permite determinar las coincidencias exactas de los registros, las parejas exactas formadas facilitan el relacionamiento y determinan el resultado de la coincidencia. En este método la vinculación de registros solo proporciona como resultado la coincidencia o no coincidencia de la vinculación, es por esto que se denomina determinístico ya que determina la relación exacta de la concordancia [8].

2.4. Método Probabilístico (Probabilistic record linkage)

La génesis de este concepto data del año 1946 donde Halbert Dunn, director de la Oficina Nacional de Estadísticas Vitales; dirección adscrita al servicio de salud pública de los Estados Unidos de América, publicó un artículo donde realizaba una abstracción de la importancia de la vinculación de registros como un mecanismo de asociación de información con otras agencias del estado y conocer todo lo relacionado a una persona como identificación del estado civil, enfermedades, edad, lugar de residencia y toda la información que el estado pudiera conocer de un individuo en particular, con el propósito de identificar la correlación de variables en eventos de salud pública [3]. Posteriormente Howard Border consolidó la base moderna de la teoría probabilística del relacionamiento o vinculación de registros [4], luego se formalizó esta teoría por Ivan Fellegi y Alan Sunter quienes demostraron que las reglas de decisión probabilística que ellos describieron eran óptimas para la comparación de atributos de cualquier fuente o condicionalmente independientes; el trabajo realizado por Ivan y Alan sigue siendo la base matemática del algoritmo de muchas aplicaciones de relacionamiento o vinculación de registro, incluso hoy en día [5]. El método probabilístico de vinculación de registros es un proceso de encontrar la similitud de dos datos bajo condiciones de incertidumbre. El objetivo es identificar y relacionar registros de entidades como personas, familias, eventos, instituciones o direcciones. En este proceso a diferencia del método determinístico, el registro que participa en la vinculación está sujeto a un error de comparación, el cual es determinado por la probabilidad de coincidencia del registro versus las posibles variantes del valor de este registro [9]. La asignación de tales probabilidades tiene la intención de

imitar al ser humano en la toma de decisiones.

2.5. Normalización y estandarización

La normalización hace parte de los métodos de vinculación de registros (Record Linkage). Sin embargo, es un componente tan especializado en los SIG, que permite normalizar nombres y direcciones especificando manualmente las reglas de transformación. Una cadena de texto es separada dentro de variables individuales, de esta manera cada variable es reconocida de acuerdo a las reglas de transformación, las cuales dependen de la semántica de las palabras a normalizar. El proceso descrito anteriormente, obedece a un enfoque determinista [10], esto quiere decir que cada uno de los componentes de una dirección son conocidos y normalizados, por ejemplo, tipo de vía principal, número de la vía, número de la casa, entre otros.

La estandarización es el proceso posterior a la normalización, el cual determina la estructura fonética de las palabras y las estandariza en textos simples según las equivalencias determinadas para asignación de variables, por ejemplo, la palabra "calle" se estandariza en "CL", la palabra "carrera" se estandariza en "CR" para posterior asignación a las variables definidas.

3. MODELO DE DATOS DE NOMENCLATURA

La ciudad de Medellín por medio del Acuerdo N°253 de 1° de diciembre de 1934, adopta el plan general de nomenclatura, el cual determina las bases del sistema de nomenclatura que actualmente es vigente, en este acuerdo se determina que la nomenclatura de la ciudad deja de ser descriptiva para adoptar un sistema exclusivamente numérico y secuencial, se define como origen la carrera Palacé (Carrera 50) y la calle Colombia (Calle 50) como orígenes del sistema y la definición de la nomenclatura domiciliaria de las casas y/o locales, "...se hará en la siguiente forma: cada puerta llevará una placa que tenga dos números separados por un guion y que indican: el primero, el número de la calle o carrera de numeración más baja entre las dos que la comprenden; y el segundo, la distancia en metros aproximada a la esquina de la misma calle. Las placas de lado derecho, siguiendo la numeración, llevarán esta distancia en números pares; y las placas del lado izquierdo, siguiendo también la numeración, llevarán estas distancias en números impares." [11].

Uno de los grandes desafíos de un buen localizador de direcciones es interpretar correctamente la nomenclatura domiciliaria de la ciudad, la cual se encuentra disponible en datos de referencia, los cuales deben ser de buena calidad. Los datos de referencia en la mayoría de los casos están compuestos por elementos geográficos con su componente atributivo asociado a cada elemento; como también se pueden

tener fuentes de información descriptiva alfanuméricamente que esté representada por una coordenada X y Y como atributo de la fuente de información; este es el caso de las bases de datos que poseen atributos espaciales producto de un proceso de georreferenciación previa. Estos dos tipos de información son muy comunes en las organizaciones y los modelos de localización dependen de la fuente de información. Para el caso de estudio, se referenciarán los modelos adoptados por la ciudad de Medellín para la georreferenciación de direcciones el cual consiste en una solución integrada de las diferentes fuentes de información tanto alfanumérica proveniente de bases de datos previamente georreferenciada como de información geográfica como la malla vial y los predios.

3.1. Estructura Nomenclatura Medellín

La nomenclatura de la ciudad está definida por el Acuerdo 46 de 2006 "Plan de Ordenamiento Territorial", como un sistema a través del cual se identifica en una malla urbana o rural las vías vehiculares, peatonales, los predios y construcciones. En el artículo 455 al 459 se definen todos los elementos estructurales de la nomenclatura como son:

3.1.1. Calles

Se denominan calles aquellas vías que poseen orientación oriente - occidente aproximadamente, y su numeración aumentará a partir de la calle 50 (Colombia) hacia el norte y disminuirá en sentido contrario a partir de esa misma vía.

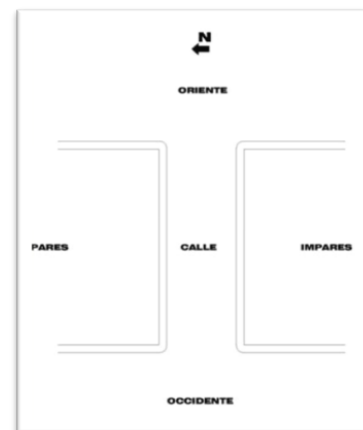


Fig. 2. Representación de Calles.

A partir del punto en que se agota la numeración de las calles se utiliza el apéndice sur. De acuerdo con lo anterior la calle 1a. es común con punto cero (0) en el eje de la vía así: Calle 1a. su costado norte y calle 1a. Sur (1S) su costado sur.

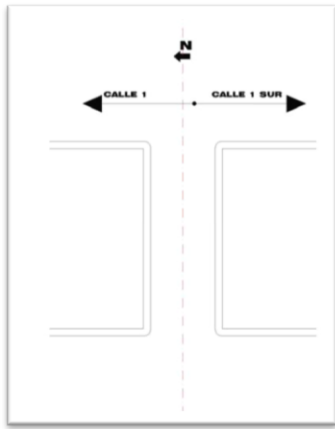


Fig. 3. Representación de Calles. Límite para el prefijo SUR.

3.1.2. Carreras

Se denominan carreras aquellas vías que poseen orientación sur - norte aproximadamente y su numeración aumentará a partir de la carrera 50 (Palacé) hacia el occidente y disminuirá en sentido contrario a partir de esta misma vía hacia el oriente.



Fig. 4. Representación de Carreras.

A partir del punto en que se agota la numeración de las carreras se utiliza el apéndice oriente. De acuerdo con lo anterior la carrera 1a. es común con punto cero (0) en el eje de la vía así: Carrera 1a. (1) su costado occidental y carrera 1a. oriente (1 O) su costado oriental.

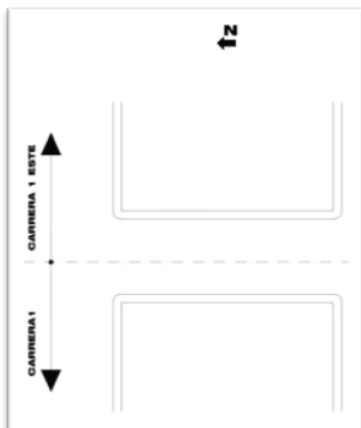


Fig. 5. Representación de Carreras. Límite para prefijo ESTE.

3.1.3. Transversales

Se denominan transversales aquellas vías cuya orientación no es la misma de las calles en el sector pero se asemeja a la de éstas. (Orientación sur - norte). Las transversales recibirán numeración de calle de acuerdo a las del sector.

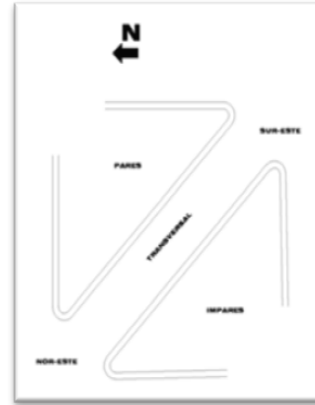


Fig. 6. Representación de Transversales.

3.1.4. Diagonales

Se denominan diagonales aquellas vías cuya orientación no es la misma de las carreras en el sector, pero se asemeja a la de ellas (orientación sur - norte). Las diagonales recibirán numeración de carrera de acuerdo con las correspondientes a éstas en el sector.



Fig. 7. Representación de Diagonales.

3.1.5. Circulares

Se denominan circulares aquellas vías cuyo alineamiento horizontal se asemeja a secciones de círculo. En la ciudad esta denominación únicamente ha sido empleada en el sector de Laureles - Bolivariana.



Fig. 8. Representación de Circulares.

3.1.6. Avenidas

Se emplea el término avenida acompañada del nombre tradicional de una vía para resaltar su importancia. El uso de la denominación de avenida únicamente posee el carácter anterior; su numeración se basará en la calle o carrera a la que se asemeja más.

3.1.7. Pasajes

Espacio público o privado destinado a la circulación peatonal y a través del cual se permite acceder a lotes internos de un predio o cruzarlo de una vía a otra.

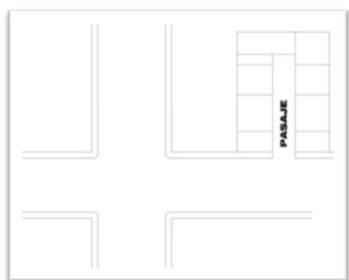


Fig. 9. Representación de Pasajes.

3.2. Modelo 1: Malla vial de Nomenclatura

La Malla Vial es la modelación vectorial de las vías del Municipio de Medellín, por medio del cual se pueden localizar las direcciones de la ciudad con la finalidad de ubicarlas en un entorno espacial. Contiene información del inventario de vías, senderos, servidumbres y pasos peatonales de la ciudad. Sirve de soporte a los procesos de localización de direcciones. Su estructura es compatible con los requerimientos del software ARCGIS de tal forma que se puede fácilmente realizar la conversión de direcciones estandarizadas en coordenadas dentro de la cartografía de la ciudad.

Tabla No 1. Estructura de datos Malla Vial

Nombre del Campo	Tipo de Datos	Descripción
OBJECTID	Object ID	Identificador único del Arcgis
TIPO_VIA	Texto	Un dominio con los valores posibles de tipo de vía que existen en la ciudad. Para la

Nombre del Campo	Tipo de Datos	Descripción
		Geodatabase Catastral se han establecido los siguientes valores: <ul style="list-style-type: none"> • Carrera (CR): Aquellas vías que poseen orientación sur - norte aproximadamente. • Calle (CL): Aquellas vías que poseen orientación oriente - occidente aproximadamente. • Circular (CQ): Aquellas vías cuyo alineamiento horizontal se asemeja a secciones de círculo. • Diagonales (DG): Aquellas vías cuya orientación no es la misma de las carreras en el sector, pero se asemeja a la de ellas. • Servidumbre-Peatonal: Espacio público o privado destinado a la circulación peatonal y a través del cual se permite acceder a lotes internos de un predio o cruzarlo de una vía a otra. • Transversales (TV): Aquellas vías cuya orientación no es la misma de las calles en el sector pero se asemeja a la de éstas. (Orientación sur - norte). • Vía Rural: Aquellas vías que permiten la comunicación de la zona urbana con los corregimientos.
NAME	Texto	Corresponde a la identificación de la vía a la cual representa. Su estructura es la siguiente: NAME= [VIA PRINCIPAL] “-“ [VIA GENERADORA] Número y apéndice de la Vía Principal seguido de un guion y el Número y apéndice de la Vía Generadora (Cuando tiene orientación Sur/ Este se omite este campo) Su función es la de determinar el cruce al cual representa.
NOMBRE_COMUN	Texto	Nombre Popular que tradicionalmente identifica las vías, se emplea para resaltar su importancia.
MUNICIPIO	Texto	Campo que determina el municipio al cual pertenece la vía permitiendo la posibilidad de ubicar la malla vial de varios municipios, por defecto para este caso se coloca “Medellín”.
PREFIJO	Texto	Campo que denota la orientación de la vía para el caso de Calle la orientación es SUR y para las carreras es ESTE. Se debe ingresar solo la inicial: S o E.
FROMLEFT	Entero Largo	Campo numérico que determina el rango inicial en el

Nombre del Campo	Tipo de Datos	Descripción
		costado izquierdo sobre el que se van a calcular las distancias sobre la longitud del arco. En caso de no tener un dato se tomara como 1.
TOLEFT	Entero Largo	Campo numérico que determina el rango final en el costado izquierdo sobre el que se van a calcular las distancias sobre la longitud del arco. En caso de no tener un dato
FROMRIGHT	Entero Largo	Campo numérico que determina el rango inicial en el costado derecho sobre el que se van a calcular las distancias sobre la longitud del arco. En caso de no tener un dato se tomara como 2.
TORIGHT	Entero Largo	Campo numérico que determina el rango final en el costado derecho sobre el que se van a calcular las distancias sobre la longitud del arco. En caso de no tener un dato
VERSION	Entero Largo	Campo numérico que hace referencia a un dominio con valores que indican la existencia de la vía o que solo es un proyecto vial. VIA EXISTENTE: Cuando se trata de una vía ya construida. PROYECTO VIAL: Cuando se trata de una vía en proyecto.
LABEL	Texto	Tipo Vía + espacio + Número de Vía (Vía Principal) + Prefijo(orientación SUR - ESTE si la tiene) [TIPO_VIA] & " " & [VIA_PRINCIPAL] & " " & [PREFIJO]
VIA_PRINCIPAL	Texto	Número de la Vía+ apéndice (Cuando tiene orientación Sur/ Este se omite para este campo)
VIA_GENERADORA	Texto	Es la vía menor que cruza la vía principal y es punto de referencia sobre el la cual se genera la placa domiciliaria: Número de la Vía+ apéndice (Cuando tiene orientación Sur/ Este se omite para este campo)
COMUNA	Texto	Número de la Comuna en la cual se encuentra el tramo de la vía. EJ:01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 50, 60, 70, 80, 90.
COD_BAR_IZ	Texto	Código del Barrio hacia la izquierda del predio en el mapa
COD_BAR_DE	Texto	Código del Barrio hacia la derecha del predio en el mapa
BARRIO_IZQ	Texto	Nombre del Barrio hacia la izquierda del predio en el mapa
BARRIO_DER	Texto	Nombre del Barrio hacia la derecha del predio en el mapa
IMPORTANCIA	Entero Corto	Relevancia o Importancia, ya sea alta, media o baja, para vías rurales.
SHAPE	Geometría	Polyline
SHAPE.LEN	Doble	Longitud de la línea

La localización de direcciones en este modelo se realiza por medio de la interpolación en base a las distancias acumuladas de los arcos que conforman el segmento, la figura 10 muestra los parámetros utilizados para el algoritmo de interpolación, d y θ determinan la ubicación del punto por medio de geometría ángulo-distancia, q es la distancia de la intersección de las vías al paramento de manzana, y v es la interpolación de la distancia con base a la nomenclatura domiciliaria.

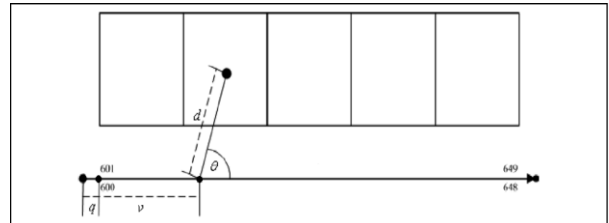


Fig. 10. Algoritmo de geocodificación para manzana vial.

3.3. Modelo 2: Predios

El modelo de predios catastrales es el dato con mayor calidad espacial que se puede encontrar, ya que es la unidad mínima para el planeamiento territorial de una ciudad o de un territorio. Ubicar direcciones sobre predios permite la representación espacial con base al centroide del polígono diferente a la interpolación de distancias del modelo de manzana vial. La principal diferencia de la localización de direcciones entre el modelo de manzana vial y el modelo de predios es que la ubicación del modelo de predios es exacto a la dirección, a diferencia del modelo de manzana vial el cual se puede localizar una dirección siempre y cuando se encuentre en el rango de distancia correspondiente a la longitud del segmento. El modelo de manzana vial no garantiza la existencia de una dirección en la ciudad, en cambio el modelo de predios garantiza que la dirección que se pretende ubicar existe y está registrada en el catastro de la ciudad.

Para que el modelo de predios funcione correctamente se requiere como fuente de información la entidad geográfica predio la cual representa los polígonos con los linderos y dimensiones de los predios de la ciudad en su parte urbana y rural. Se compone de atributos (Campos) que identifican espacialmente la ubicación del predio dentro de la división político administrativa de la ciudad y que permite, con el código CBML distinguir explícitamente entre un lote o parcela y otro. El atributo geométrico área representa el área del terreno catastral. El código CBML refiere a la unión de los códigos de COMUNA, BARRIO, MANZANA Y LOTE, el cual representa la numeración única del lote en la ciudad de Medellín; también se requiere la entidad direcciones de la base de datos jurídica catastral de la ciudad, con el fin de cruzar estas dos entidades por medio del CBML y obtener la correspondencia geográfica del lote asociada a la dirección oficial de la ciudad.

La figura 11 muestra la entidad predio con la representación al centroide del lote, esto con el fin de realizar una búsqueda por dirección y conocer las coordenadas x, y del centroide.

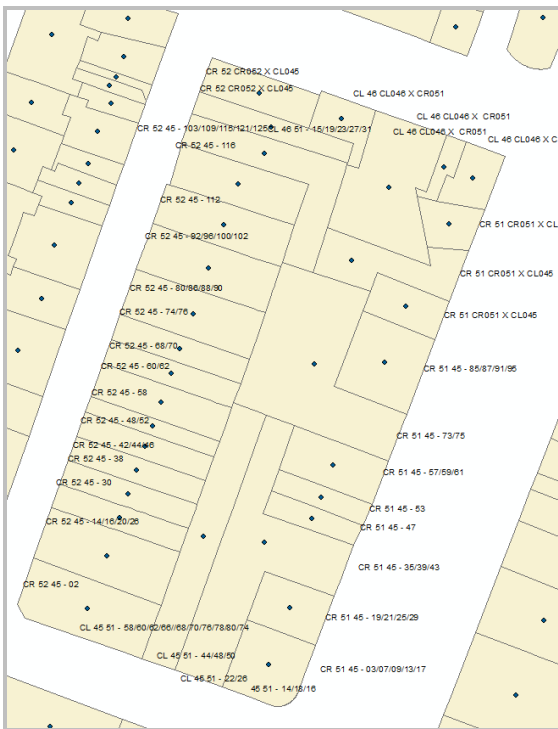


Fig. 11. Ubicación x, y del centroide del predio.

3.4. Modelo 3: Bases de datos de direcciones

El modelo de bases de datos de direcciones, se deriva de tablas maestras donde es conocida las direcciones y los atributos de espacialización (x, y o latitud, longitud), estas tablas maestras, para el caso de la ciudad de Medellín, provienen de entidades externas como Empresas Públicas de Medellín, la cual tiene localizada todas las direcciones de las instalaciones donde se ofrece el servicio de energía. El proceso de geocodificación se realiza con estas fuentes de información el cual determina las coordenadas x, y por medio de una búsqueda al campo de dirección estandarizado y normalizado.

4. CALIDAD EN LA GEOCODIFICACIÓN

Para todos los procesos de geocodificación es necesario cuantificar la calidad de los resultados, a pesar del uso generalizado de estos procesos en diferentes disciplinas, la calidad de la información no ha tomado la importancia o relevancia que esta requiere. Varias de las investigaciones que usan la geocodificación como método de georreferenciación para el análisis espacial de la información, no mencionan o referencian la calidad de la geocodificación; aquellas investigaciones que sí referencian la calidad de la geocodificación usualmente lo realizan con el método de acierto (match rate). Los geocodificadores comerciales ofrecen un alto porcentaje de acierto en sus productos o servicios, sin

tener en cuenta que existen otras variables en la calidad de la geocodificación. Investigaciones recientes recomiendan poner un poco de atención al método de acierto (match rate) ya que este puede ser engañoso en términos de resultados reales en la localización de direcciones [12].

La calidad de los resultados de la geocodificación pueden ser caracterizados en los siguientes componentes: completitud y exactitud posicional. La completitud es el porcentaje de registros localizados. Exactitud posicional indica cual es la cercanía de la ubicación en comparación con la realidad.

4.1. Completitud (match rate)

Es la medida de la calidad de la geocodificación en cuanto acierto de direcciones ubicadas, también se define como el porcentaje de direcciones ubicadas. Muchos se preguntarán cuál es el porcentaje aceptable?; uno de los pocos estudios que habla sobre el tema, fue realizado en Monte Carlo, el cual consistía en simular los incidentes delictivos ubicados a nivel de bloques censales para determinar la tasa de coincidencia y así obtener un patrón fiable de delincuencia [13]. El resultado indicó que el porcentaje necesario para aceptar estadísticamente un resultado de geocodificación debe ser mayor del 85%.

4.2. Exactitud posicional

La exactitud posicional mide el absoluto de la distancia del punto localizado por medio de la geocodificación respecto a la ubicación real del mismo, varios estudios han determinado una estimación cuantitativa sobre la exactitud posicional en la geocodificación. Una estimación dentro de los umbrales normales, determina que para un área residencial el rango de error posicional aceptable está entre 25-168 metros estudiados en [14] – [24] basados en los valores medios de la distribución del error. Los resultados en las áreas urbanas son generalmente más exactos que en las zonas rurales [15], [16] y [18]. También es de anotar que la ocurrencia de grandes errores de posición es relativamente común. Por ejemplo, en uno de los estudios más exhaustivo por Cayo & Talbot (2003), el 10% de una muestra de direcciones urbanas geocodificadas tuvo un error de posición mayor de 96 m y un 5% de la muestra tuvo un error de posición mayor de 152 m. Para direcciones rurales la exactitud posicional estuvo entre 1.5 -2.9 Km.

5. MÉTODO

5.1. Área de estudio

El área de estudio para efectos de este artículo, se concentró en la zona urbana del Municipio de Medellín, la cual está constituida por 16 comunas y 249 barrios, en esta zona existe información que permite la geocodificación de direcciones tales como información catastral de los predios, malla vial que es trazada como línea central de las vías de la ciudad y que contiene

información vial como de nomenclatura domiciliaria y bases de datos de terceros.

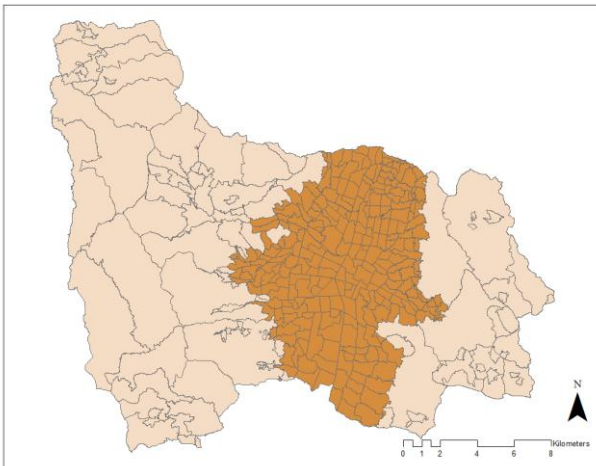


Fig. 12. Área de estudio: Zona Urbana de Medellín.

5.2. Bases de datos de estudio

Las bases de datos utilizadas para este estudio, provienen de diferentes fuentes de información, tales como sisben, educación e industria y comercio del año 2006, con esto se quiere demostrar que la fuente de información no es relevante ya que el proceso de geocodificación debe garantizar la heterogeneidad de las estructuras de las fuentes de información y dar un resultado que cumpla con las expectativas en la ubicación de direcciones. En total se va realizar un proceso de geocodificación a (n=20744) direcciones distribuidas en 3 bases de datos.

Dentro de las fuentes de información utilizada como punto de referencia para la exactitud posicional, es la capa de nomenclatura domiciliaria, capa geográfica construida por catastro Medellín con el fin de tener un punto de referencia de la nomenclatura domiciliaria, se define como los elementos que tienen asociados la nomenclatura domiciliaria de cada uno de los CBML de la ciudad de Medellín; también se define como el nivel de información que provee la ubicación espacial de las placas de las nomenclaturas domiciliarias de los predios del Municipio de Medellín en el SIG Catastral. La estructura y representación gráfica del elemento es la siguiente:

Tabla No 2. Estructura de datos Malla Vial

Nombre del campo	Tipo De Datos	Definición
OBJECTID	ObjectID	Identificador del elemento geográfico
COBAMA	Texto	Código COBAMA del predio
CBML	Texto	Código CBML del predio
VIA	Texto	Vía sobre la cual se ubica el acceso a la vivienda
PLACA	Texto	Placa de la nomenclatura domiciliaria
ANGULO	Double	Posición en que queda ubicada la nomenclatura del predio

Nombre del campo	Tipo De Datos	Definición
SHAPE	Geometría	Punto

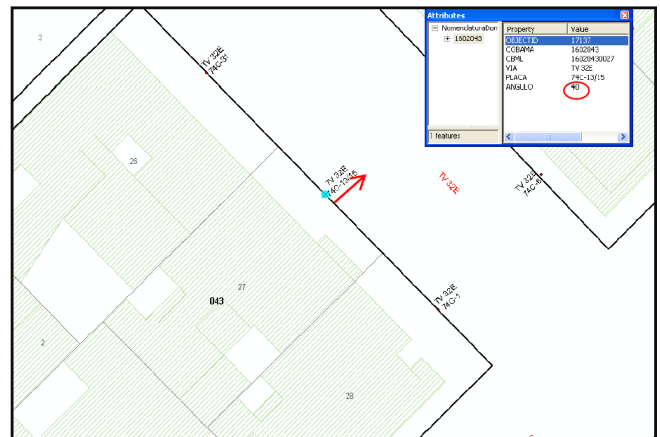


Figura. 13. Representación gráfica de la capa Nomenclatura

5.3. Proceso de Geocodificación

El proceso de geocodificación se realizará con 2 herramientas que permitirán la geocodificación de direcciones y su calidad de geocodificación tanto de completitud como de exactitud espacial; las herramientas a utilizar son: Geocoding de ArcGIS configurado para la versión 10.1 y el Geocoding desarrollado para la ciudad de Medellín (GEOCOD) el cual dentro de su proceso involucra a diferentes fuentes de información definidas en el ítem 3. El geocoding de la ciudad de Medellín (GEOCOD) es una aplicación transversal desarrollada en Oracle (PL/SQL) y Python al lado del servidor que permite la ubicación de direcciones de una forma determinística utilizando la capacidad de procesamiento de los RDBMS. El geocoding de ArcGIS permite la ubicación de direcciones en un modelo probabilístico.

5.4. Análisis

Para cada base de datos se realizará el proceso de geocodificación tanto con GEOCODING-ArcGIS (Probabilístico) como con el GEOCOD-Medellín (Determinístico), a los resultados obtenidos se calculará la calidad del proceso tanto en completitud como en la exactitud posicional, la velocidad del procesamiento será cuantificada con el fin de determinar la eficiencia en el procesamiento de la información; con estos resultados determinaremos para la ciudad de Medellín cual es el mejor proceso de geocodificación dentro de las herramientas que tiene a disposición.

Para efectos de este estudio la completitud se tomará como referencia las direcciones perfectamente encontradas (puntaje=100) en el proceso de geocodificación probabilístico esto con el fin de homogenizar los resultados tanto para el método determinístico como para el probabilístico.

6. RESULTADOS

6.1. Completitud

En la tabla 3 se observa el resultado del acierto en la ubicación de direcciones por medio de los procesos de geocodificación, tomando como referencia las tres fuentes de información de direcciones, la realizada con GEOCODING-ArcGIS (método probabilístico) y con GEOCOD-Medellín (método determinístico), teniendo como referencia la zona urbana de Medellín, la cual está compuesta por 16 comunas. Para el proceso de geocodificación por medio del GEOCODING-ArcGIS se tuvo un porcentaje medio de direcciones ubicadas para la ciudad de Medellín de un 78% en comparación al 86% del proceso de geocodificación por medio del GEOCOD-Medellín.

De acuerdo al estudio realizado por Ratcliffe [13], donde establece un valor de referencia de aceptación mínimo del 85% que garantiza un patrón confiable para los procesos de geocodificación, se utilizó este valor de referencia para definir un análisis de cobertura temático donde nos indicara el comportamiento de la ciudad de Medellín con los 2 procedimientos de geocodificación. En la figura 14, 15 y 16 se observa el comportamiento del acierto en la ubicación de direcciones de la ciudad de Medellín, determinando que el proceso de geocodificación del GEOCOD-Medellín ofrece una mayor confiabilidad en la completitud de ubicación de direcciones para las bases de datos que intervinieron en el estudio.

La figura 14, muestra el resultado de cobertura del patrón de confiabilidad aceptado en el acierto para la fuente de información "Base de Datos1", donde el proceso de geocodificación GEOCODING-ArcGIS tiene una proporción de cobertura del territorio de un 37.5% frente a un 87.5% del proceso de geocodificación con GEOCOD-Medellín.

La figura 15, muestra el resultado de cobertura del patrón de confiabilidad aceptado en el acierto para la fuente de información "Base de Datos2", donde el proceso de geocodificación GEOCODING-ArcGIS tiene una proporción de cobertura del territorio de un 6.25% frente a un 81.25% del proceso de geocodificación con GEOCOD-Medellín.

La figura 16, muestra el resultado de cobertura del patrón de confiabilidad aceptado en el acierto para la fuente de información "Base de Datos3", donde el proceso de geocodificación GEOCODING-ArcGIS tiene una proporción de cobertura del territorio de un 37.5% frente a un 87.5% del proceso de geocodificación con GEOCOD-Medellín.

6.2. Exactitud Posicional

En la tabla 4 muestra el resultado de la exactitud posicional por medio de los procesos de geocodificación, tomando como referencia la tres

fuentes de información de direcciones, el primer proceso realizado con GEOCODING-ArcGIS (método probabilístico) y el segundo con el GEOCOD-Medellín (método determinístico), teniendo como referencia la zona urbana de Medellín, la cual está compuesta por 16 comunas. Para el proceso de geocodificación por medio del GEOCODING-ArcGIS se tuvo una exactitud posicional media para la ciudad de Medellín de 42.6 m en comparación al 34.6 m del proceso de geocodificación por medio del GEOCOD-Medellín.

Los resultados obtenidos de la exactitud posicional con los dos procesos de geocodificación, muestran que se encuentra entre los valores medios normales referenciado en los estudios de diferentes autores [14]-[24], garantizando las precisiones normales para estudios con información espacial.

Las gráficas 1, 2, 3, muestran la exactitud posicional media de las direcciones ubicadas en cada comuna, entre más bajo sea el valor mayor es la precisión de las direcciones encontradas con el proceso de geocodificación en relación a la posición real de la dirección. El comportamiento de los dos procesos de geocodificación es similar, se resalta en la gráfica, donde la geocodificación con GEOCOD-Medellín tiene una tendencia generalizada inferior a los 20 m.

6.3. Rendimiento de procesamiento

A medida que la tecnología informática avanza a pasos agigantados, las capacidades de cómputo y procesamiento lo hacen paralelamente, en la actualidad en el mercado de los procesadores se encuentra unidades de cómputo para el hogar y trabajo de 16 núcleos y servidores comerciales superiores a 64 núcleos de procesamiento, que le dan una capacidad de análisis de información a las empresas que en años pasados era inimaginable. Para los procesos de geocodificación se requiere de este tipo de característica tecnológica ya que los algoritmos de los procesos de geocodificación involucran gran cantidad de cálculos y consultas a base de datos.

Los procesos de geocodificación utilizados en el presente artículo están implementados en diferentes plataformas tecnológicas, en el caso del GEOCODING-ArcGIS, éste es soportado para ejecutarse a nivel de equipo de cómputo (portátil y/o escritorio) y para el GEOCOD-Medellín, éste está soportado para correrse como un web services o dentro de una aplicación propietaria de la Alcaldía de Medellín. Este último se encuentra implementado en una plataforma tecnológica con 16 núcleos de procesamiento.

Los resultados muestran en la tabla 5, que el proceso de geocodificación a nivel de servidor es más eficiente en comparación al proceso de geocodificación que se corre en equipos de escritorio y/o portátil, la relación es aproximadamente 16x más rápido a nivel servidor.

Tabla No 3. Completitud procesos de Geocodificación

FUENTE DE INFORMACIÓN	COMUNA	ARCGIS-GEOCODING				GEOCOD				TOTALES
		ENCONTRADA		NO ENCONTRADA		ENCONTRADA		NO ENCONTRADA		
		Número	Porcentaje	Número	Porcentaje	Número	Porcentaje	Número	Porcentaje	
Base de Datos 1	01	753	77%	224	23%	915	94%	62	6%	977
	02	428	82%	97	18%	485	92%	40	8%	525
	03	737	78%	213	22%	891	94%	59	6%	950
	04	485	86%	76	14%	491	88%	70	12%	561
	05	370	87%	57	13%	391	92%	36	8%	427
	06	613	80%	158	20%	596	77%	175	23%	771
	07	837	76%	265	24%	998	91%	104	9%	1102
	08	715	57%	545	43%	581	46%	679	54%	1260
	09	681	78%	193	22%	787	90%	87	10%	874
	10	420	93%	30	7%	434	96%	16	4%	450
	11	365	86%	59	14%	377	89%	47	11%	424
	12	509	91%	53	9%	529	94%	33	6%	562
	13	698	75%	230	25%	837	90%	91	10%	928
	14	286	77%	87	23%	323	87%	50	13%	373
	15	447	85%	77	15%	490	94%	34	6%	524
	16	824	80%	203	20%	945	92%	82	8%	1027
SUBTOTAL- BD1		9168	78%	2567	22%	10070	86%	1665	14%	11735
Base de Datos 2	01	200	77%	60	23%	240	92%	20	8%	260
	02	108	85%	19	15%	118	93%	9	7%	127
	03	178	77%	53	23%	217	94%	14	6%	231
	04	135	81%	31	19%	141	85%	25	15%	166
	05	97	82%	22	18%	110	92%	9	8%	119
	06	178	84%	35	16%	171	80%	42	20%	213
	07	223	75%	76	25%	260	87%	39	13%	299
	08	185	57%	140	43%	164	50%	161	50%	325
	09	182	78%	50	22%	206	89%	26	11%	232
	10	95	86%	16	14%	101	91%	10	9%	111
	11	85	82%	19	18%	91	88%	13	13%	104
	12	116	82%	25	18%	121	86%	20	14%	141
	13	195	73%	73	27%	233	87%	35	13%	268
	14	102	81%	24	19%	111	88%	15	12%	126
	15	128	83%	27	17%	141	91%	14	9%	155
	16	220	79%	59	21%	256	92%	23	8%	279
SUBTOTAL-BD2		2427	77%	729	23%	2681	85%	475	15%	3156
Base de Datos 3	01	352	74%	126	26%	444	93%	34	7%	478
	02	209	87%	31	13%	228	95%	12	5%	240
	03	366	80%	94	20%	433	94%	27	6%	460
	04	270	86%	45	14%	279	89%	36	11%	315
	05	182	91%	18	9%	191	96%	9	5%	200
	06	322	80%	79	20%	333	83%	68	17%	401
	07	447	78%	129	22%	536	93%	40	7%	576
	08	355	57%	272	43%	305	49%	322	51%	627
	09	353	79%	92	21%	407	91%	38	9%	445
	10	213	92%	19	8%	220	95%	12	5%	232
	11	193	89%	24	11%	204	94%	13	6%	217
	12	223	89%	28	11%	229	91%	22	9%	251
	13	327	71%	132	29%	406	88%	53	12%	459
	14	116	77%	35	23%	133	88%	18	12%	151
	15	220	83%	46	17%	243	91%	23	9%	266
	16	435	81%	100	19%	489	91%	46	9%	535
SUBTOTAL-BD3		4583	78%	1270	22%	5080	87%	773	13%	5853
TOTAL-MEDELLÍN		16178	78%	4566	22%	17831	86%	2913	14%	20744

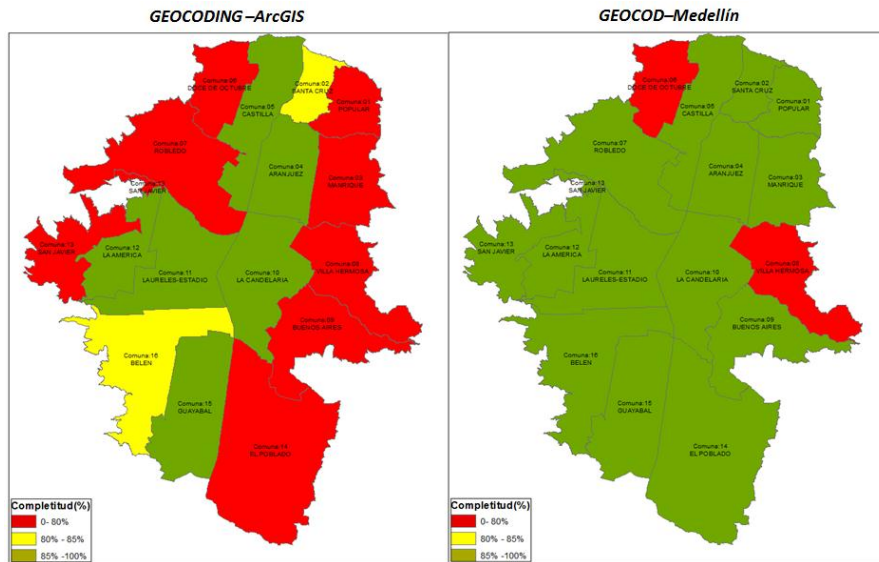


Figura. 14. Completitud, procesos de Geocodificación Base de Datos 1.

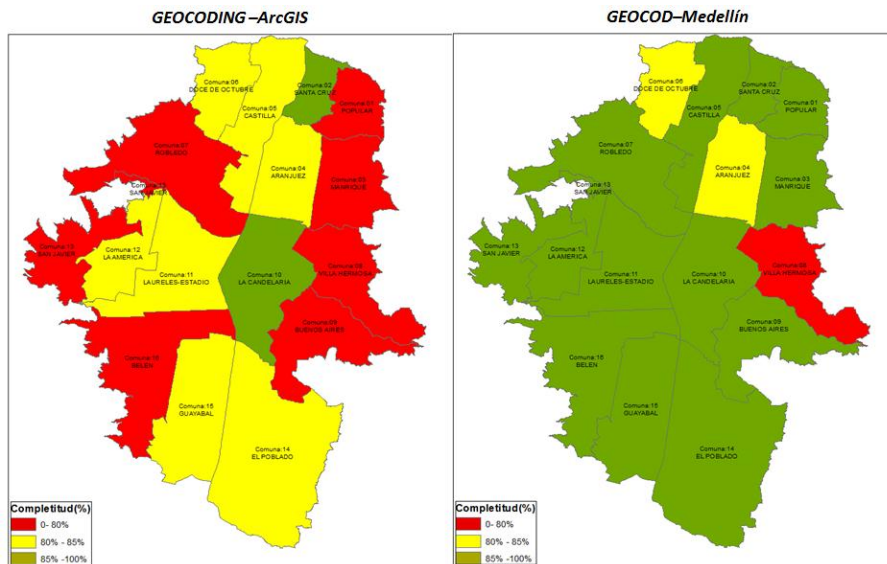


Figura. 15. Completitud, procesos de Geocodificación Base de Datos 2.

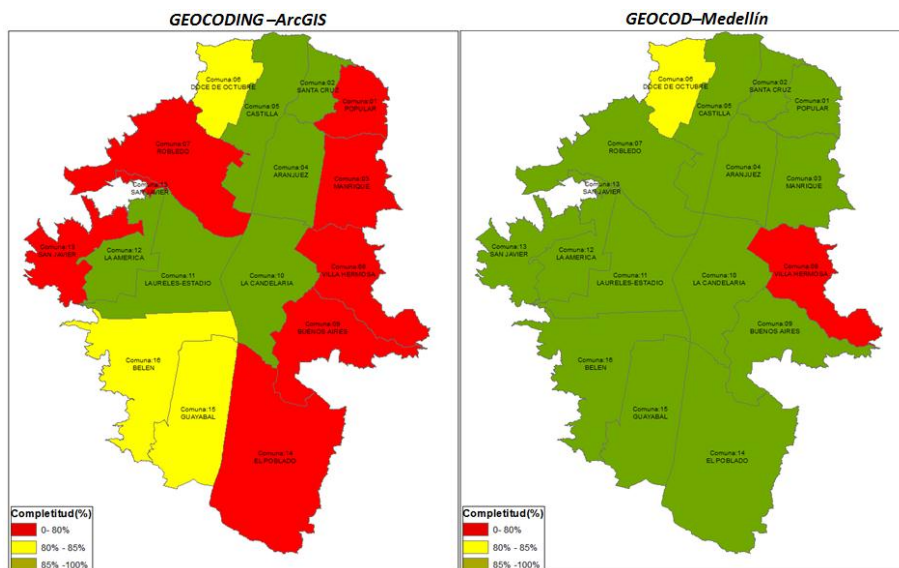
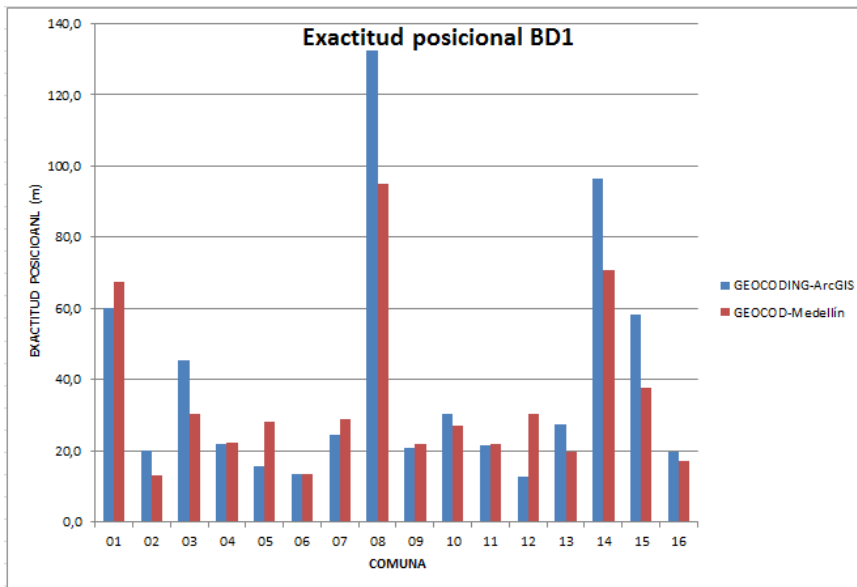


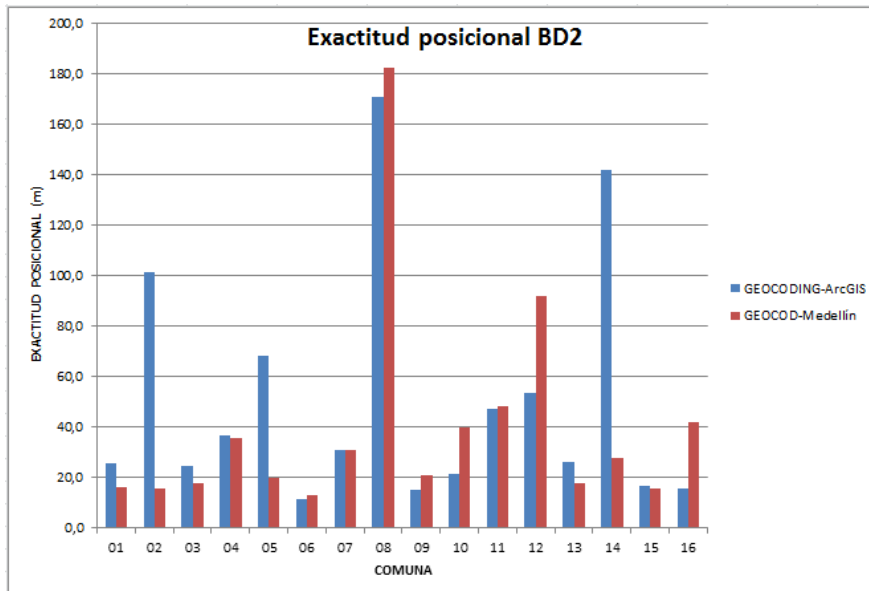
Figura. 16. Completitud, procesos de Geocodificación Base de Datos 3.

Tabla No 4. Exactitud posicional de los procesos de geocodificación.

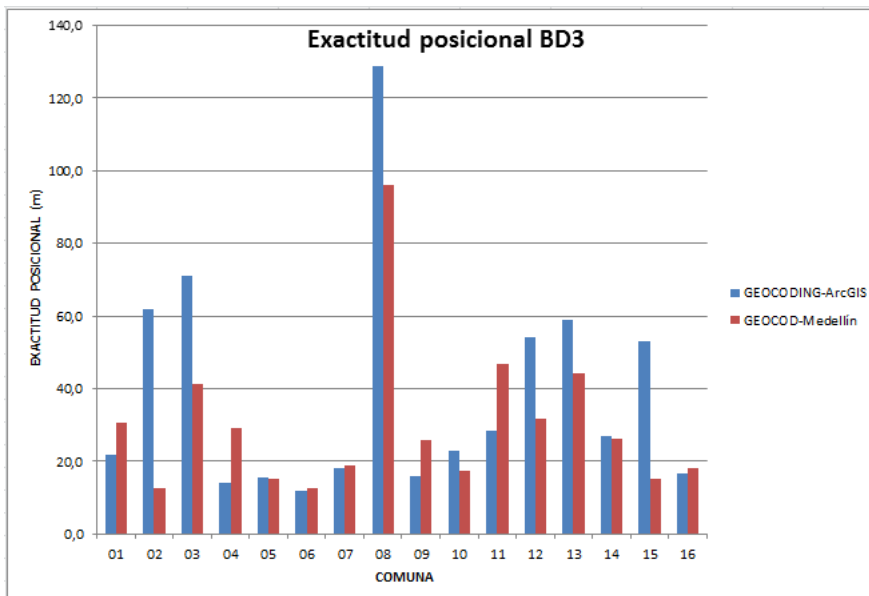
FUENTE DE INFORMACIÓN	COMUNA	GEOCODING-ArcGIS	GEOCOD-Medellín
		ABS DISTANCIA (m)	ABS DISTANCIA (m)
Base de Datos 1	01	59,9	67,3
	02	20,1	12,9
	03	45,3	30,4
	04	21,8	22,1
	05	15,5	28,3
	06	13,6	13,5
	07	24,6	29,0
	08	132,4	95,1
	09	20,9	21,8
	10	30,1	27,0
	11	21,4	21,9
	12	12,8	30,2
	13	27,5	19,5
	14	96,6	70,7
	15	58,4	37,5
	16	19,5	17,1
MEDIA-BD1		38,8	34,0
Base de Datos 2	01	25,4	16,0
	02	101,2	15,7
	03	24,6	17,6
	04	36,4	35,3
	05	68,3	19,9
	06	11,4	13,1
	07	30,7	31,0
	08	171,0	182,2
	09	15,1	20,7
	10	21,5	39,9
	11	47,2	48,1
	12	53,6	91,8
	13	25,8	17,6
	14	141,9	27,7
	15	16,8	15,5
	16	15,3	42,0
MEDIA-BD2		50,4	39,6
Base de Datos 3	01	21,9	30,5
	02	61,8	12,6
	03	71,0	41,3
	04	14,0	29,1
	05	15,7	15,2
	06	11,9	12,6
	07	18,3	19,1
	08	128,9	96,2
	09	16,0	25,9
	10	22,9	17,6
	11	28,5	47,0
	12	54,3	31,7
	13	58,9	44,4
	14	27,1	26,4
	15	53,0	15,3
	16	16,5	18,3
MEDIA-BD3		38,8	30,2
TOTAL MEDIA		42,6	34,6



Gráfica. 1. Exactitud posicional Base de Datos 1.



Gráfica. 2. Exactitud posicional Base de Datos 2.



Gráfica. 3. Exactitud posicional Base de Datos 3.

Tabla No 5. Tiempo de procesamiento procesos de geocodificación.

FUENTE DE INFORMACIÓN	Tiempo de ejecución (Seg)	
	Servidor	Escritorio
Base de Datos 1	11	187
Base de Datos 2	30	347
Base de Datos 3	36	696



Gráfica. 4. Tiempo de procesamiento procesos de geocodificación.

7. CONCLUSIONES

Este estudio tiene como finalidad mostrar una comparación empírica de los diferentes procesos de geocodificación que se utiliza en la Alcaldía de Medellín. En general, estos procesos ofrecen una alternativa de georreferenciación de direcciones para los organismos del ente municipal como para los ciudadanos del común que visita el geoportal de la ciudad, las herramientas o procesos estudiados, determina que los métodos determinísticos para la geocodificación de direcciones son más efectivos que los métodos probabilísticos, ya que no hay que olvidar que en este tipo de métodos la incertidumbre es una variable compleja de controlar que altera el resultado y depende de diferentes elementos como la estructura de direcciones de la ciudad, la formación de los predios y vías en el territorio, la calidad de la información que soporta el proceso, entre otros. La determinística es una estrategia de buenos resultados, siempre que se cuente con información completa y de calidad.

En la actualidad la Alcaldía de Medellín a determinado utilizar el GEOCOD-Medellín para los procesos transversales de la organización por la confiabilidad en los resultados, la capacidad de procesamiento y la integración que tiene con fuentes externas e internas de información en tiempo real; es el caso del proceso PQRS (Peticiónes, Quejas, Reclamos y Solicitudes) de la entidad quien integra la georreferenciación y la involucra para el análisis espacial en pro de toma de decisiones.

REFERENCIAS

- [1] D. W. Goldberg, J. P. Wilson and C. A. Knoclock "From text to geographic coordinates: The current state of geocoding" *URISA Journal*, Vol. 19, pp. 33-46, 2007.
- [2] ESRI. "Geocoding Rule Base Developer Guide", Redlands, CA, 2003. Online [May. 2013].
- [3] H. L. Dunn. "Record Linkage". *American Journal of Public Health* 36 (12): pp. 1412-1416, 1946.
- [4] H. B. Newcombe, J.M. Kennedy, S.J. Axford and A. P. James. "Automatic Linkage of Vital Records". *Science* 130(3381), pp. 954-959, 1959.
- [5] I. P. Fellegi and A. B. Sunter. "A Theory for Record Linkage". *Journal of the American Statistical Association* 64 (328): pp. 1183-1210, 1969.
- [6] LL. Roos and A. Wajda. "Record linkage strategies. *Methods of Information in Medicine*", Silver Spring, v. 30, n. 2, pp. 117-123, 1991.
- [7] C. A. Boussy, and K. G. Scott, "Use of data base linkage methodology in epidemiological studies of mental retardation". *International Review of Research in Mental Retardation*, San Diego, v. 19, pp. 135-161, 1993.
- [8] S. Gomatam, and R. A. Carter, "Computerized stepwise deterministic strategy for linkage", Gainesville: University of Florida, Department of Statistics. Technical Report, 1999.
- [9] P. A. Zandbergen, "A comparison of address point, parcel and street geocoding techniques". *Computers, Environment and Urban System*, Vol. 32, Issue3, pages 214-232, Copyright © Elsevier Ltd. All rights reserved, 2008.

- [10] P. Christen, T. Churches and J. X. Zhu, "Probabilistic name and address cleaning and standardization". The Australasian Data Mining Conference, Canberra, Australia, December 3, 2002.
- [11] A. Morales P., "Medellín en la memoria de Ricardo Olano" Colección Biblioteca Básicas de Medellín Tomo 23, Medellín, Colombia. p. 40, 2006.
- [12] E. A. Whitsel, K. M. Rose, J. L. Wood, A. C. Henley, D. Liao and G. Heiss, "Accuracy and repeatability of commercial geocoding". American Journal of Epidemiology, 160(10), pp. 1023–1029, 2004.
- [13] J. H. Ratcliffe, "Geocoding crime and a first estimate of a minimum acceptable hit rate". International Journal of Geographical Information Science, 18(1), pp. 61–72, 2004.
- [14] S. M. Dearwent, R. J. Jacobs and J. B. Halbert, "Locational uncertainty in georeferencing public health datasets". Journal of Exposure Analysis and Environmental Epidemiology, 11, pp. 329-334, 2001.
- [15] M. R. Cayo and T. O. Talbot, "Positional error in automated geocoding of residential addresses". International Journal of Health Geographics, 2(10), 2003.
- [16] M. R. Bonner, D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim, "Positional accuracy of geocoded addresses in epidemiologic research". Epidemiology, 14(4), pp. 408–412, 2003.
- [17] H. A. Karimi and M. Durcik, "Evaluation of uncertainties associated with geocoding techniques". Computer-Aided Civil and Infrastructure Engineering, 19, pp. 170-185, 2004.
- [18] M. H. Ward, J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, JS. Colt and P. Hartge, "Positional accuracy of two methods of geocoding". Epidemiology, 16(4), pp. 542-547, 2005.
- [19] E. A. Whitsel, P. M. Quibrera, R. L. Smith, D. J. Catellier, D. Liao, A. C. Henley and G. Heiss, "Accuracy of commercial geocoding: Assessment and implications". Epidemiological Perspectives and Innovations, 3(8), 2006.
- [20] F. B. Zhan, J. D. Brender, I. De Lima, L. Suarez and P. H. Langlois, "Match rate and positional accuracy of two geocoding methods for epidemiologic research". Annals of Epidemiology, 16(11), pp. 842–849, 2006.
- [21] M. Schootman, D. A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo and G. Higgs, "Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research". Annals of Epidemiology, 17(6), pp. 464–470, 2007.
- [22] M. J. Strickland, C. Siffel, B. R. Gardner, A. K. Berzen and A. Correa, "Quantifying geocode location error using GIS methods". Environmental Health, 6(10), 2007.
- [23] P. A. Zandbergen, "Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads". BMC Public Health, 7(37), 2007.
- [24] D. L. Zimmerman, X. Fang, S. Mazumdar, and G. Rushton, "Modeling the probability distribution of positional errors incurred by residential address geocoding". International Journal of Health Geographics, 6(1), 2007.