UNIVERSITAT POLITÈCNICA DE VALÈNCIA

IDENTIFICACIÓN DE PATRONES DE MOVILIDAD URBANA

Trabajo Final del

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor:

Lucas M. Rodríguez

Directores:

Miguel Rebollo

Javier Palanca

Valencia, España

Julio de 2018

Resumen

Detectar patrones de movilidad urbana puede resultar de interés para los gobiernos y las instituciones a la hora de gestionar servicios relacionados a los desplazamientos de las personas. Esto puede contribuir también con el desarrollo sustentable de la ciudad en cuestión. En este trabajo se generaron modelos predictores de caminos entre un punto inicial y un punto final, utilizando *tweets* con información geoposicionada como fuente de datos. Se analizaron los resultados de las predicciones de estos modelos contrastándolos con caminos reales, generados a partir de los mismos *tweets.* Luego, se hizo una comparativa de la predicción basada en el género. Se arribó a la conclusión de que el mejor modelo predictor es el modelo gravitacional combinado con el de atracción, y que existen diferencias en los desplazamientos de hombres y mujeres, siendo los caminos del género femenino más sinuosos y cortos.

Resum

Detectar patrons de mobilitat urbana pot resultar d'interès per als governs i les institucions a l'hora de gestionar serveis relacionats als desplaçaments de les persones. Açò pot contribuir també amb el desenvolupament sustentable de la ciutat en qüestió. En aquest treball es van generar models per a predir de camins entre un punt inicial i un punt final, utilitzant *tweets* amb informació geoposicionada com a font de dades. Es van analitzar els resultats de les prediccions d'aquests models contrastant-los amb camins rals, generats a partir dels mateixos *tweets*. Després, es va fer una comparativa de la predicció basada en el gènere. Es va arribar a la conclusió que el millor model per a predir és el model gravitacional combinat amb el d'atracció, i que existeixen diferències en els desplaçaments d'homes i dones, sent els camins del gènere femení més sinuosos i curts.

Abstract

Detecting human mobility patterns may be of interest to governments and institutions when it comes to managing services related to the movement of people. This can also contribute to the sustainable development of the specific city. In this work, models were generated in order to predict paths between a starting point and an ending point, using tweets with geo-localization as a data source. The results of the predictions of these models were analyzed by contrasting them with real paths, also generated from these tweets. Then, a gender-based comparison of the prediction was made. It was concluded that the best predictive model is the gravitational one combined with the attraction one, and that there are differences in the movements of men and women, with the female paths being more sinuous and short.

Índice

1. Introducción

1.1. Planteo del problema

Los patrones de movilidad urbana son aquellos que dan cuenta del desplazamiento de las personas en una determinada área urbana. Poder identificarlos resulta importante para el desarrollo de las mismas, ya que pueden ser aprovechados para la planificación de caminos y servicios de transporte, la distinción de lugares concurridos y posibles focos comerciales, entre muchos otros.

Para reconocer estos patrones, es necesario contar con información geoposicionada de individuos a lo largo del tiempo. Existen actualmente distintas maneras de obtener estos datos, sin embargo, el continuo desarrollo de la telefonía móvil y de las redes sociales hace que la información geoposicionada sea cada vez más accesible. Además, la capacidad de procesamiento y almacenaje de estos datos es cada vez mayor. Por todo esto, nuevos modelos de identificación de patrones de movilidad urbana son elaborados constantemente, haciendo uso de esta información, que no era accesible en el pasado.

El presente trabajo ofrece un posible modelo de identificación de patrones de movilidad urbana aplicable a cualquier ciudad de la cual se tengan datos geoposicionados. Como caso de estudio se ha tomado la ciudad de Valencia, España, a partir de datos geoposicionados obtenidos de Twitter. Esta información podrá ser utilizada por servicios de planificación urbana para entender donde se producen cuellos de botella en la ciudad, cuáles son los caminos más transitados, etc.

1.2. Motivación

Tener conocimiento de los patrones de movilidad urbana de una ciudad puede ayudar a las organizaciones gubernamentales en la planificación de tráfico y transporte público. Además, permite a organizaciones privadas una mejora en servicios públicos dependientes de las localizaciones y movimiento de las personas (como por ejemplo telefonía). Lo descrito anteriormente permitiría un desarrollo en el sistema de transporte en la ciudad en cuestión, ya que se entenderían mejor sus necesidades y se realizarían acciones acordes.

El transporte es un factor de vital importancia en el contexto del desarrollo sustentable dados los efectos que tiene en el medio ambiente, su impacto social y económico y su vínculo con otros sectores. El sector del transporte ha estado creciendo en los últimos años y esta tendencia no se prevé que baje, por lo que tener una estrategia para un sistema de transporte sustentable es un tema de prioridad.

1.3. Objetivos

El objetivo general de este trabajo es el de generar un modelo predictor de patrones de movilidad urbana y analizar sus variaciones para generar mejores predicciones. Para esto se generarán caminos reales de personas, los cuales se contrastarán con el camino generado por el modelo predictor.

Los objetivos específicos a cumplir son:

- Generar un modelo de actividad de una ciudad utilizando datos obtenidos de Twitter que explique los patrones de movilidad urbana.
- Aplicar este modelo al caso de la ciudad de Valencia, España.

- Generar caminos de individuos dentro de la ciudad de Valencia utilizando datos obtenidos de Twitter.
- Utilizar el modelo generado para realizar predicciones sobre los caminos de los individuos.
- Comparar los caminos predichos con los caminos reales. Obtener métricas para entender qué configuración del modelo es la que obtiene mejores resultados.
- Realizar una clasificación de caminos en base al género de la persona. Comparar los caminos predichos con los caminos reales. Obtener métricas para entender qué configuración de los modelos obtiene mejores resultados y entender las diferencias entre ellos.

1.4. Estructura

El resto de la presente memoria se estructura de la siguiente forma: en el segundo capítulo se enumerarán publicaciones relacionadas al tema, y se hará un resumen de las variaciones que estos presentan. En el capítulo tres se presentará la configuración elegida para la generación del modelo, así como los datos a utilizar. Para la comparación, en el capítulo cuatro se detallarán las variantes para la generación de caminos, y en el capítulo cinco las métricas de acierto que acompañan a cada uno. En el sexto capítulo se realizará una comparativa entre la predicción diferenciando por género, y se finalizará con las conclusiones en el capítulo siete.

***

2. Antecedentes

Existen actualmente variados estudios en torno a la identificación de patrones de movilidad urbana, que utilizan diferentes enfoques y técnicas. A continuación, se ofrece un breve resumen de aquellos que resultaron de mayor relevancia para la elaboración de este trabajo.

Bejar, Álvarez, García, Gómez, et al. [1] utilizan API públicas provistas por Twitter e Instagram, en Barcelona y Milán, como fuente de información geoposicionada. Los autores se plantearon dos objetivos. El primero fue analizar los patrones espacio-temporales para usuarios que visitan una ciudad y usuarios que viven en ella. Se emplearon técnicas de *clustering* para agrupar las coordenadas de los eventos, utilizando un radio prefijado. De esta manera, las posiciones de los clústeres se adaptaban a la densidad de los datos. Una vez realizado el proceso de discretización, se construyeron *datasets*. A estos se les aplicó la técnica de *frequent itemset* (*FP-Growth*) para encontrar patrones ocultos. De esta manera, fueron capaces de encontrar patrones significativos (conexión entre clústeres usuales) que representan caminos que deberían ser priorizados, por ejemplo, para la planificación urbana.

El segundo objetivo fue utilizar la información de estos patrones para identificar grupos o perfiles entre los usuarios. Se generó un vector de características siguiendo el modelo *bag of words*, y para los valores de los atributos se utilizaron *frecuencia de términos* y *frecuencia de documentos inversa*. Con estos datos se realizó *clustering* con *k-vecinos*, *clustering espectral* y *clustering con propagación de afinidad*.

Hawelka, Sitko, Beinat, Sobolevsky et al. [2] también realizaron un estudio utilizando datos geoposicionados obtenidos de Twitter para la detección de patrones de movilidad mundial, y estimaron la cantidad de viajes internacionales con respecto al país de residencia. A través de particiones iterativas de redes, se determinaron regiones de interés que reflejaban las

divisiones del mundo. Se utilizaron redes y grafos para la representación de la información, usando el método de optimización modular, en combinación con la técnica de optimización combo. Luego, iterativamente, se crearon subredes en cada nodo.

Sin embargo, solo estudios recientes utilizan datos de redes sociales. Siła-Nowicka, Vandrol, Oshan, Long et al. [3], través de la colaboración de voluntarios que proporcionaron información geoposicionada (contextualizada), analizaron movimientos diarios de usuarios, para poder identificar lugares "primarios" (hogar), "secundarios" (trabajo) y "terciarios" (ocio). Para lograr esto, se utilizaron técnicas de *segmentación de trayectos* en combinación con *minería de datos* con el objetivo de clasificar tipos de viajes, así como lugares de origen y fin de cada viaje. Además, se implementó una red neuronal *feed-forward* para la clasificación del tipo de transporte realizado (a pie o vehículo).

Una variación interesante fue introducida por Terroso-Sáenz, Cuenca-Jara, González-Vidal, Skarmeta et al. [4], donde mediante datos obtenidos de Twitter, se realizó una predicción online (sin entrenamiento ni pre-procesamiento) del camino que recorrerá un usuario entre dos puntos. Para esto, se mantuvo un grafo por usuario, representando los lugares importantes para una persona, cómo estos se conectan, y los tópicos comunes mencionados en esos lugares. Para la realización de dichos grafos utilizaron dos técnicas: *clustering* (*bag of words*) para la identificación de lugares importantes; y un modelo de Markov multinivel para predecir localización y actividades potenciales.

En general, los datos de redes sociales pueden ser dispersos o escasos, dependiendo de la localización. Jiang, Li, Zhou, Chen et al. [5] intentan atacar este problema utilizando información de usuarios de un servicio de telefonía móvil de China para analizar el movimiento de las personas basados en la comunidad a la que pertenecen. Se propone un método nuevo para analizar patrones de movilidad a nivel colectivo a través de diferencias entre

comunidades. Para ello, se utilizaron redes espaciales basadas en los perfiles de los usuarios, donde los nodos representaron los puntos de interés y el peso de las aristas representaron la densidad de movimiento entre esos nodos. Luego, se realizaron particiones de estas redes con algoritmos de clasificación a través de *XOR similarity matrix*.

Es interesante ver como un análisis de este estilo puede tener implicancias en escenarios reales, como es el caso de Naji, Wu, Zhang [6]. En este estudio se intentaron optimizar las ganancias de taxistas en una ciudad en China, empleando información de desplazamientos registrados por sus dispositivos GPS. Se utilizó *clustering K-Medias* para dividir a los usuarios en grupos basados en sus características. Luego, se compararon los históricos de cada grupo utilizando técnicas de *DBSCAN clustering* y *mash map matching*.

En general, existen 3 grandes tipos de enfoques a la hora de estudiar el movimiento humano, como se puede aprecia en *Tabla 1*:

- Análisis cualitativo: Trata de estudiar las trayectorias por las cuales las personas se mueven, los caminos que toman y las razones por esta elección. [1,3,4,6,7,8]
- Análisis cuantitativo: Estudia principalmente los flujos de personas entre 2 o más lugares, ignorando el camino que toman para llegar del origen al destino. [2,9,10]

Además, también hay una tercera rama que estudia las clasificaciones de sub-regiones de un lugar, para así entender las diferencias entre las personas de cada una de esas zonas basados en los movimientos que realizan. [1,3,5,10,11,12]

TIPO	ENFOQUE
Flujo de personas	Análisis cuantitativo de los desplazamientos
Trayectorias	Análisis cualitativo de los desplazamientos
Distinción de zonas	Distinción de sub-regiones en base a análisis cualitativos y cuantitativos

Tabla 1. Tipos de análisis de movilidad

De acuerdo al alcance geográfico del problema, el análisis se puede subdividir en Internacional, Nacional, Regional y Local, como se muestra en *Tabla 2*. Cada uno de estos enfoques ataca diferentes aspectos del problema, desde el más global (como flujos migratorios) al más específico (como el camino que una persona recorre de su hogar a su trabajo).

TIPO	ALCANCE
Internacional	Aspectos migratorios o turísticos [2]
Nacional	Aspectos de migración o turismo internos [8,12]
Regional	Aspectos laborales, desplazamientos urbanos-rurales [5,9,11]
Local	Aspectos laborales, de ocio y transporte diario [1,3,4,6,7,10]

Tabla 2. Alcance de estudios de movilidad

Cada problema tiene sus cualidades específicas y por tanto la "mejor técnica" a aplicar es relativa. La *Tabla 3* contiene una lista de las técnicas usadas en publicaciones recientes que se relacionan con la temática de movilidad.

TÉCNICA	VENTAJAS	DESVENTAJAS
Clustering [1,2,6]	Simple, eficiente. Fácil de implementar. Fácil de interpretar.	Determinar la cantidad de clústeres. Sensibilidad a valores atípicos. Óptimos locales.
Redes neuronales [11,12]	Buenas aproximaciones. Ataca problemas complejos.	Usa muchos datos de entrenamiento. Caja negra.
Levy Flight [7]	Describe el movimiento. Exploratoria.	Es probabilístico (requiere parámetros). Aleatoria (máximos locales).
Modelo gravitacional [8,9]	Explica la relación entre variables. Es intuitivo. Fácil de entrenar. Realiza predicciones rápido.	Asume influencia entre todas las variables. Requiere determinar parámetros. Sensibilidad a datos dispersos.

Tabla 3. Técnicas utilizadas para elaboración de modelos predictores de movilidad

Con respecto al origen de los datos, en el pasado la recolección de datos geoposicionados era manual, o con la colaboración de voluntarios [3]. El siguiente paso fue la utilización de datos

registrados por dispositivos GPS en vehículos, con las limitaciones de cantidad de datos que esto presentaba [6].

Hoy en día, con el desarrollo tecnológico incesante, esta información no solo es fácilmente calculable y almacenable, sino que también es muy accesible. Prácticamente no existe zona poblada sin alcance de redes de telefonía móvil, así como personas sin un dispositivo (especialmente en países desarrollados).

Varios trabajos han sido realizados utilizando información de las redes telefónicas [5,7,9], con las limitaciones legales y de privacidad que presentan. Sin embargo, hoy en día la mayor cantidad de información geoposicionada se encuentra en las redes sociales. La combinación de redes móviles y dispositivos móviles accesibles, el auge en las redes sociales y las particularidades de los humanos como seres sociales, resultan en una cantidad de información pública nunca antes vista.

Diversas publicaciones utilizaron información de redes sociales (usualmente con API pública) para el análisis de información geoposicionada. Entre ellas se encuentran: Twitter [1,2,4,10,11], Foursquare [10] y Flickr [8,12]. Un resumen de las diferentes fuentes de los datos, así como sus ventajas y desventajas se pueden encontrar en *Tabla 4.*

MÉTODO	VENTAJAS	DESVENTAJAS
Voluntarios/ Observación	Alta precisión. Fácil interpretación.	Poca información. Difícil de obtener.
Dispositivos GPS	Alta precisión. Sin mucha interacción humana.	Solo identifica transporte de vehículos. No está universalmente extendido.
Redes móviles	Relativamente buena precisión. Registro automático. Gran cantidad de datos.	Limitaciones por privacidad La precisión depende de la calidad de la cobertura.
Redes sociales	Relativamente buena precisión. Registro automático. Enorme cantidad de datos con contexto.	La precisión depende de la calidad de la conexión del dispositivo usado. Requiere la participación activa del usuario.

Tabla 4. Fuente de datos para la generación de modelos de predicción de movilidad

La *Tabla 5* muestra un resumen de las investigaciones relacionadas al tema, junto con su tipo, alcance, técnica y método de recolección de datos.

INVEST	TIPO	ALCANCE	TÉCNICA	DATOS
[1]	Cualitativo, Distinción de zonas	Local	Clustering	Redes sociales
[2]	Cuantitativo	Internacional	Clustering	Redes sociales
[3]	Cualitativo, Distinción de zonas	Local	Segmentación de trayectorias	Voluntarios
[4]	Cualitativo	Local	Bag of Words	Redes sociales
[5]	Distinción de zonas	Regional	Técnica propia	Redes móviles
[6]	Cualitativo	Local	Clustering	Dispositivos GPS
[7]	Cualitativo	Local	Levy Flight	Redes móviles
[8]	Cualitativo	Nacional	Modelo gravitacional	Redes sociales
[9]	Cuantitativo	Regional	Modelo gravitacional	Redes móviles
[10]	Cuantitativo, Distinción de zonas	Local	Kernel Density	Redes sociales
[11]	Distinción de zonas	Regional	Redes neuronales	Redes sociales
[12]	Distinción de zonas	Nacional	Redes neuronales	Redes sociales

Tabla 5. Resumen de investigaciones relacionadas

Esta investigación permitió entender la realidad actual y los alcances de los documentos relacionados al tema, lo que ayudó tener un punto de partida a la hora de iniciar este trabajo. El siguiente capítulo explica el modelo elegido, que fue basado en las técnicas y los resultados de los trabajos mencionados en este capítulo.

3. Generación del modelo

Partiendo de las conclusiones de las investigaciones actuales, así como las bondades y limitaciones de cada una de las fuentes y técnicas utilizadas, se decide proseguir con la siguiente configuración:

- Fuente de los datos: Redes sociales (más específicamente: Twitter).
- Técnica utilizada: Modelo gravitacional.
- Escala: Local (Valencia).
- Enfoque: Análisis de trayectorias.

La elección de utilizar redes sociales está relacionada con el auge que estas han tenido en los últimos tiempos. En la *Figura 1* se puede observar el incremento en la cantidad de usuarios activos en Twitter, especialmente el incremento de usuarios móviles. Estos últimos tienen la capacidad de generar datos geoposicionados, un caso ideal para el problema a resolver.

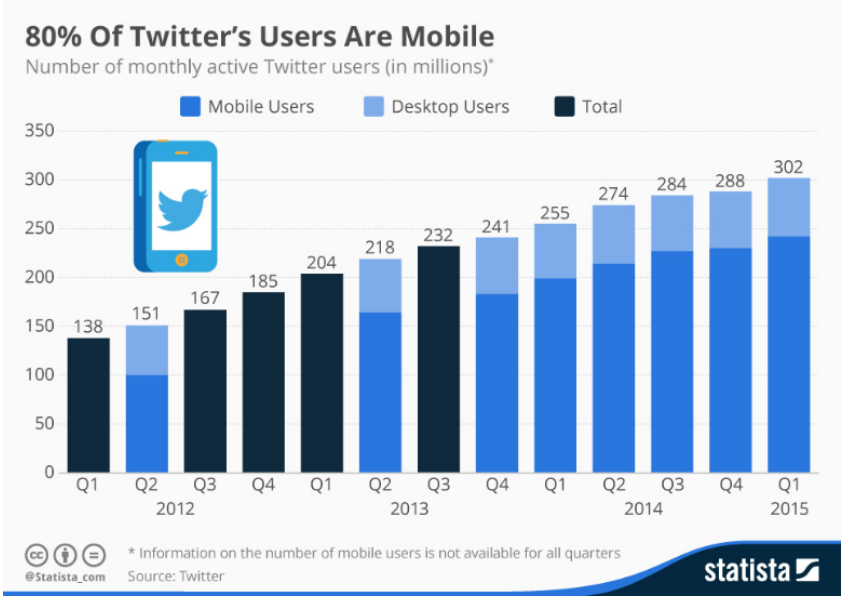

Figura 1. Incremento de usuarios en móviles Twitter [13]

Para este trabajo, se tomaron datos de la API de Twitter, recolectando *tweets* geoposicionados de usuarios en la ciudad de Valencia entre las fechas 04/02/2015 al 26/02/2018. Esto deja un total de 1.483.338 *tweets* de 72.717 usuarios. En la *Figura 2* se pueden observar la distribución de *tweets* con respecto al tiempo.

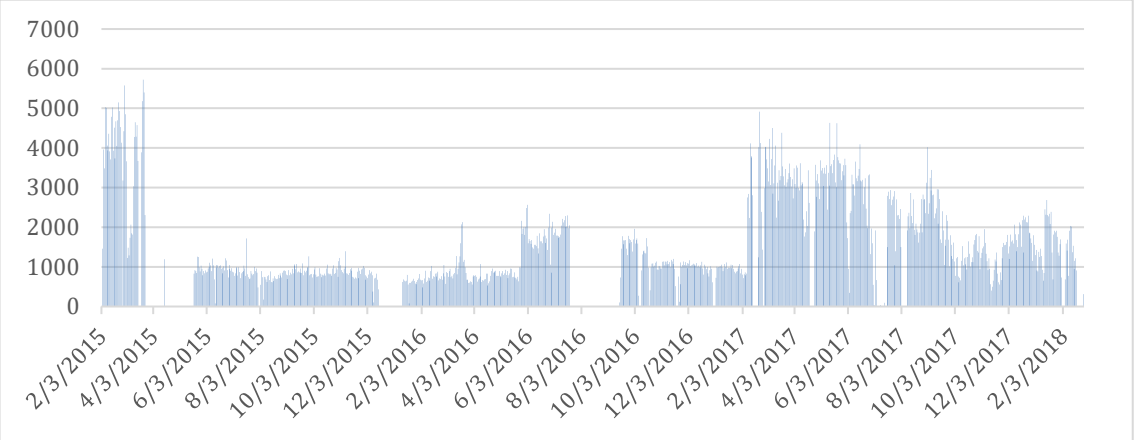

Figura 2. Distribución de tweets utilizados para la generación del modelo

Los espacios o huecos representan caídas del servicio de recolectado, o problemas con la API de Twitter. Esto sin embargo no afecta al proceso realizado, ya que se abstrae el momento de obtención del *tweet*: lo que importa es la geoposición.

La API de Twitter expone muchos campos en cada *tweet* obtenido, como perfil del usuario, mensaje, archivos adjuntos, entre otros. Para este trabajo, solo se tomaron datos relevantes: geoposición, fecha, identificador del usuario y nombre.

A continuación, se detallarán los 3 tipos de modelos que se generarán y utilizarán en este trabajo:

- Modelo gravitacional: Inspirado en el campo potencial gravitatorio, supone que cada *tweet* tiene masa y realiza deformaciones en el mapa acorde a ellos.

- Modelo de atracción: Expande el modelo gravitacional, dándole un peso especial al punto de destino de un camino que se intenta predecir.

- Geodésica: Busca el camino de menor coste entre 2 puntos en un espacio curvo. Para este caso particular, busca el camino de menor coste en el modelo gravitacional, teniendo en cuenta sus deformaciones tridimensionales.

Modelo gravitacional

Siguiendo las aplicaciones y técnicas de la literatura, para predecir el movimiento de las personas por una ciudad se utilizará un modelo inspirado en el campo de potencial gravitatorio y de cómo este se vería influido si cada *tweet* tuviera masa. Un detalle de la aplicación de las leyes gravitacionales puede encontrarse en la publicación de Barthélemy. [14.

El modelo generado emula un modelo gravitacional, donde las zonas *"pesadas"* es donde se encuentran más *tweets*, mientras que las zonas *"livianas"* o *"planas"* son donde se encuentran menos. Para generar este modelo, se toma el mapa de Valencia y se divide en una grilla de 50x60 subregiones, como se muestra en *Figura 3*.

Luego, cada *tweet* procesado se coloca en cada una de las divisiones según sus coordenadas. Esto resulta en una especie de histograma tridimensional, donde cada región acumula las cantidades de *tweets* que le corresponden.

Figura 3. División de la ciudad de Valencia para la generación del modelo

Por cada *tweet* procesado, en dicho mapa se aplica un peso, siguiendo una distribución gaussiana.

$$G_i = -e^{-\frac{1}{2}\left(\frac{x-lon(i)}{len(x)}\right)^2} \times e^{-\frac{1}{2}\left(\frac{y-lat(i)}{len(y)}\right)^2}$$

Donde G_i es la matriz de pesos que genera el punto *i* en el mapa, *x* es el vector con las coordenadas horizontales del mapa e *y* es el vector con las coordenadas verticales del mapa. Además, *len(x)* y *len(y)* representan el largo y el alto de cada sub-región en particular, y *lon(p)* y *lat(p)* representan la longitud y latitud del punto p. El efecto que se genera en el mapa se puede ver en la *Figura 4*.

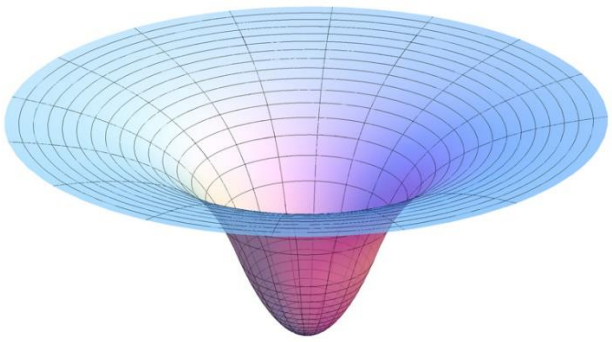

Figura 4. Efecto del potencial gravitatorio

Luego, al aplicar un segundo *tweet* en el mapa, se puede observar como ambos potenciales gravitatorios se afectan, como muestra la *Figura 5*.

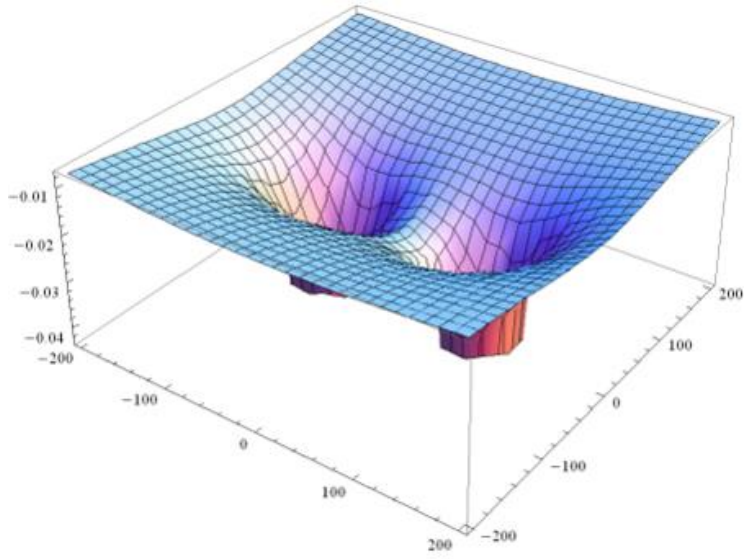

Figura 5. Efecto de 2 potenciales gravitatorios cercanos

Una vez calculados todos los efectos gravitacionales, estos se acumulan en una matriz final *G* y posteriormente se aplica un logaritmo para normalizar.

$$G = \log_{10} \sum_{i} G_i$$

Si este efecto se repite para todas las regiones del mapa, se obtiene un modelo gravitacional más rico, con mayor efecto entre las regiones, como se puede observar en la *Figura 6.* Esta figura representa el modelo gravitacional de la ciudad de Valencia generada por los *tweets* recolectados para este trabajo. Luego, en la *Figura 7* se puede ver el efecto de este modelo en la ciudad con curvas de nivel en un mapa de contorno. Esto muestra las regiones de mayor concentración de *tweets*, como es el centro de la ciudad, y las de menor concentración, como es el mar. Además, las líneas de contorno (y la separación entre ellas) explican la pendiente y la dirección de dicha pendiente en zonas en particular, indicando la capacidad de atracción en cada punto.

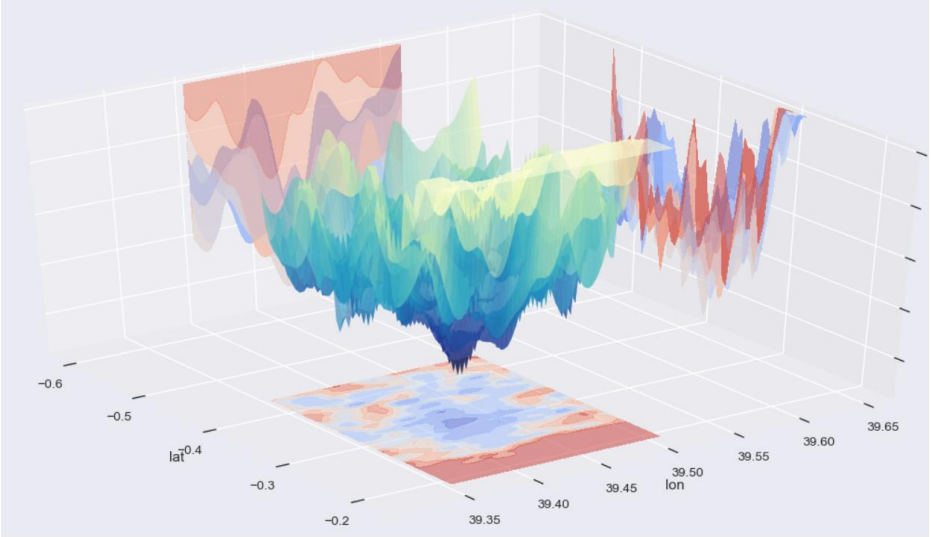

Figura 6. Vista tridimensional del modelo gravitacional

Figura 7. Mapa de contorno del modelo gravitacional

Dado que el objetivo final es predecir el camino que recorrerá una persona para ir de un punto A a un punto B, lo más sencillo que se podría hacer con los datos existentes es simplemente utilizar el camino "de caída" en base a las pendientes (gradientes) de cada punto. La *Figura 8* muestra las direcciones que se tomarían en cada punto a la hora de generar un camino. En este caso, un camino se da por "finalizado" cuando: se llega al punto de destino, se fuese a repetir un punto o se llegue a un número máximo de puntos (100 pasos).

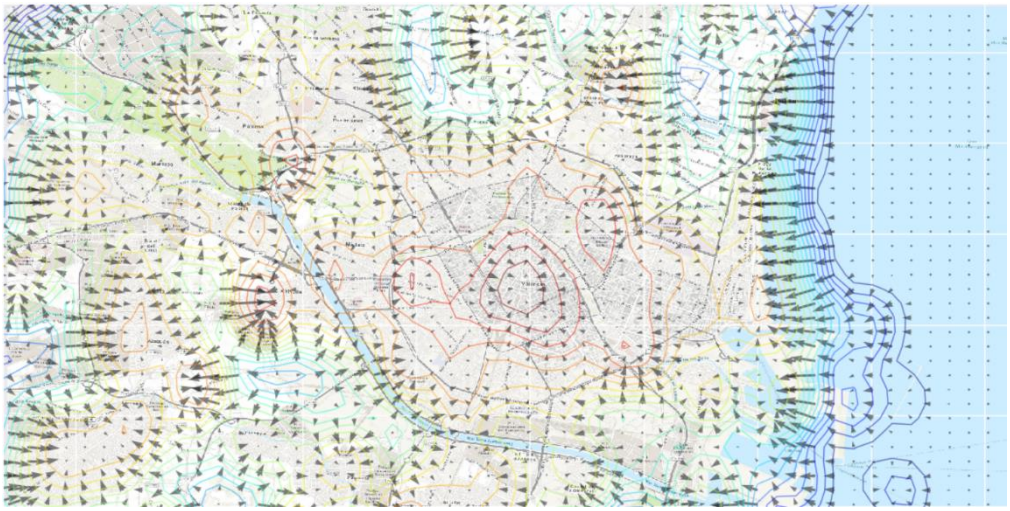

Figura 8. Direcciones de gradiente en mapa gravitacional

Modelo de atracción

Seguir los gradientes del modelo gravitacional para la generación de caminos conlleva a que en muy pocos casos se llegue al destino, debido a que hay muchos mínimos locales y no hay una tendencia explícita del punto de inicio al de fin. Es por esto que, como primera idea, se aplica un modelo de atracción, donde el punto de destino tiene un gran peso sobre el modelo gravitacional, así haciendo tender el camino al punto final.

A la matriz *G* calculada para el modelo gravitacional se le suma un peso en el punto final *f* dependiente de un factor que llamaremos *A*. Este factor regula la deformación que genera este punto sobre el mapa gravitacional original. Poco peso hará que haya poca influencia, pero mucho peso transformará al camino en una línea recta. Por lo tanto, es necesario probar con diferentes pesos para encontrar el apropiado.

$$G_A = G + A\left(e^{-\frac{1}{2}\left(\frac{x-lon(f)}{len(x)}\right)^2} \times e^{-\frac{1}{2}\left(\frac{y-lat(f)}{len(y)}\right)^2}\right)$$

Esto se puede ver en la *Figura 9*, donde las flechas representan el gradiente de cada región. Además, en la *Figura 10* se puede ver otro ejemplo, donde se aplica una atracción fuerte en el centro del mapa. Los niveles de profundidad del mapa muestran una clara tendencia de bajada en el centro. Esto se puede apreciar de mejor manera en la *Figura 11*, donde se ve claramente que el centro del mapa es el que atrapa el mayor poder gravitacional. En este caso, un camino se da por "finalizado" cuando: se llega al punto de destino, se fuese a repetir un punto o se llegue a un número máximo de puntos (100 pasos).

Figura 9. Efecto de atracción en el punto final de un camino (punto rojo)

Figura 10. Mapa gravitacional con atracción fuerte en el centro

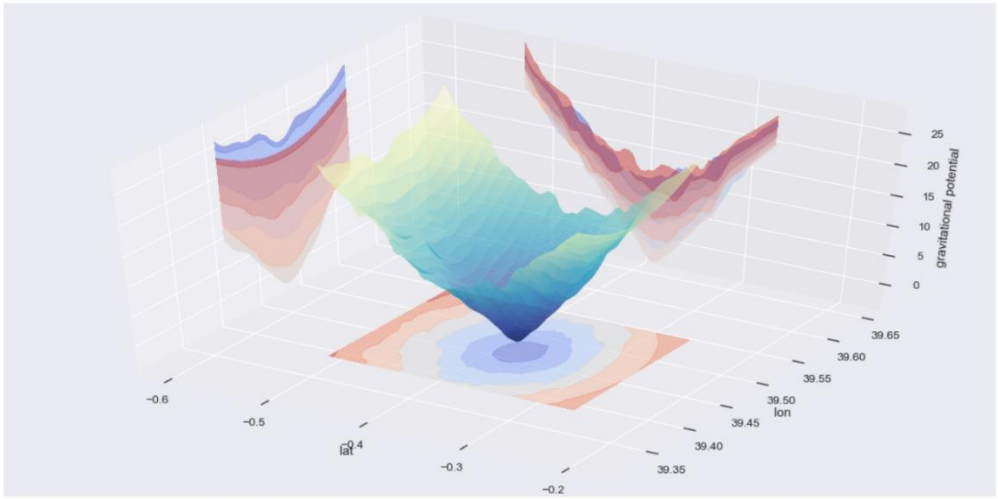

Figura 11. Efecto de atracción en el centro del mapa gravitacional

Geodésica

Cómo última variación se probó la generación del camino geodésico del modelo; es decir, el camino de menor coste entre el punto inicial y final en un espacio tridimensional. Esto está inspirado en el estudio de superficies, en el cual, al tratar de encontrar el camino más corto entre 2 puntos, se deben tener en cuenta las irregularidades de la superficie como factor de coste. Por lo general, se busca que el camino sea lo más "recto" posible, evitando pendientes grandes.

Para lograr esto se transformó el modelo en un grafo, donde cada nodo es un cuadrado en el mapa, las aristas son conexiones a sus puntos adyacentes y el peso de la arista es la diferencia de altura presentada en el modelo. Luego, se realizó una búsqueda del camino de menor coste utilizando el algoritmo de Dijkstra. La *Figura 12* muestra un ejemplo de un camino generado por esta técnica, donde se puede apreciar que los puntos se mantienen usualmente en las curvas de nivel contiguas, es decir, en "alturas" parecidas.

Figura 12. Ejemplo de camino a través de la geodésica

4. Generación de caminos reales

Una vez generado el modelo predictor, es necesario un set de datos para poder comprobarlo. Para esto, se utilizan los *tweets* almacenados como descriptores del camino que una persona realiza a lo largo de un día.

Para la generación de caminos se siguió la siguiente regla general: "Un camino consta de puntos donde una misma persona realizó un *tweet* en un mismo día, siempre que estén a una distancia significativa y las transiciones de 3 puntos no representen ángulos cerrados". En lo específico, lo que se realiza es:

- Por cada día, se toman todas las personas que hayan realizado un *tweet* al menos 3 veces.
- Por cada persona, se eliminan *tweets* consecutivos cuya geoposición sea la misma, dejando solo 1 (para evitar repetidos).
- Se genera un camino temporal, que consta de los puntos donde la persona realizó un *tweet*.
- Ese camino se descompone en 2 o más caminos, realizando la división por cada combinación de 3 puntos que genera un ángulo menor a 90 grados.
- Por último, se eliminan los caminos de menos de 2 puntos.
- Se recompilan todos los caminos encontrados de todos los usuarios para todos los días.

La *Figura 13* muestra la división por ángulos mencionada. El ángulo formado por los puntos 2, 3 y 4 es menor a 90 grados, por lo que el camino compuesto por los puntos [0,1,2,3,4,5,6] se descompone en 2 caminos compuestos por: [0,1,2,3] y [3,4,5,6]

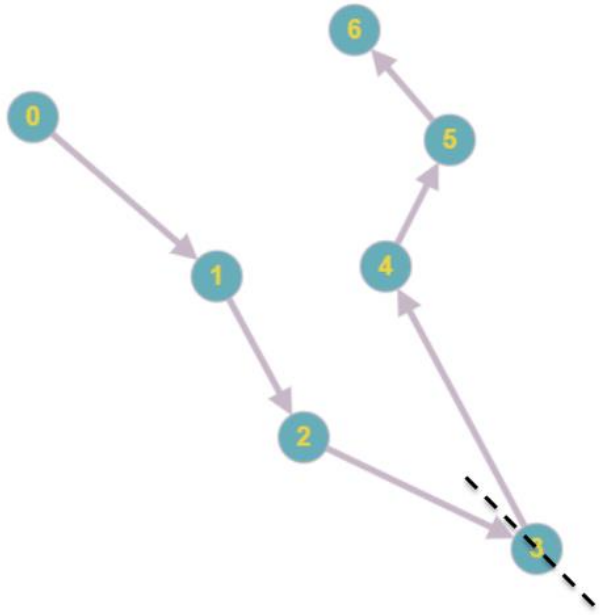

Figura 13. División de camino por ángulo

Esto genera un total de 6.972 caminos. Al realizar un análisis de los caminos generados, se puede observar que la gran mayoría de estos son cortos. Casi el 97% de los caminos son de solo 3 y 4 puntos. La *Tabla 6* compara las distancias totales y promedios de los caminos, basado en la cantidad de pasos que tiene cada uno.

Cantidad de pasos	1	2	3	4	5	6	7	8
Caminos	0	0	5876	879	162	43	9	6
Porcentaje	0	0	84,27%	12,61%	2,32%	0,49%	0,13%	0,09%
Distancia total (km)	0	0	27.766	4.977	1.001,8	188,2	96	23
Promedio (km)	0	0	4,73	5,66	6,18	4,38	10,66	3,83

Tabla 6. Análisis de los caminos generados

Al analizar las distancias de cada uno el histograma *Figura 14* muestra que la mayoría de los caminos son cortos. La distancia total de los caminos es de 353.2 km.

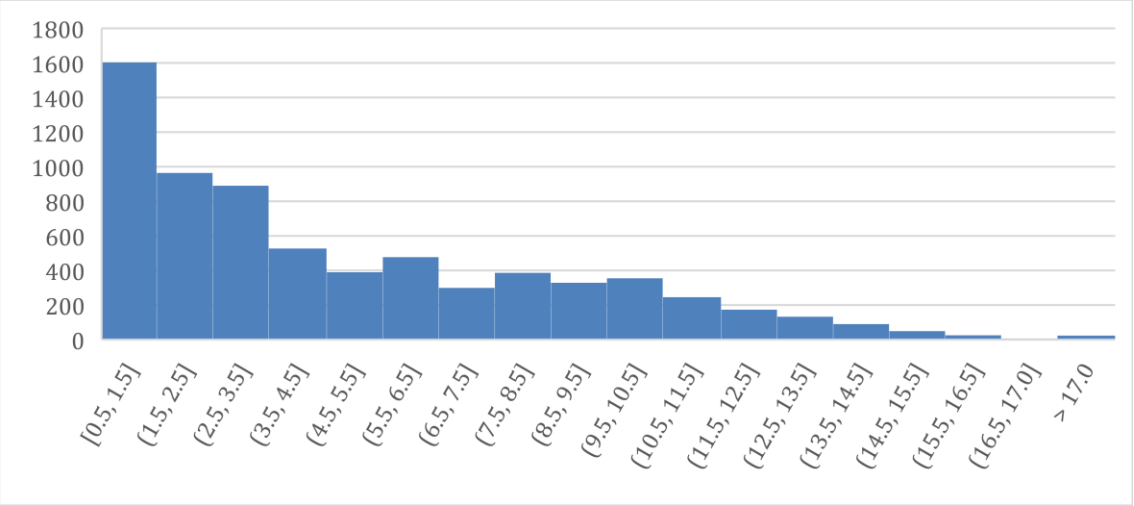

Figura 14. Largo de los caminos en km

Métrica para comparar predicciones

El objetivo final del modelo es de predecir caminos que realizará una persona para ir de un punto a otro. Para probar la bondad de este modelo, es necesario comparar caminos reales con caminos predichos. Los caminos reales son los mencionados en el apartado anterior, obtenidos de los *tweets* de una persona a lo largo del día. Los caminos predichos son los caminos generados por el modelo.

Para poder probar la precisión del camino predicho, se debe utilizar alguna métrica para poder compararlo con el camino real. Para este trabajo, se elige trabajar con la Distancia de Hausdorff [15], que ha sido usada en la anterioridad para la similitud de curvas y de superficies [16,17]. Esta técnica propone que:

$$d_H(X,Y) = \max\{ \sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y) \}$$

Siendo X e Y el camino real y el camino predicho, respectivamente, y $d(x,y)$ la distancia euclídea entre el punto $x \in X$ y el punto $y \in Y$. La técnica plantea que los caminos están cerca si cada punto de cada camino está cerca de un punto del otro camino. La distancia de Hausdorff es la distancia más larga entre 2 puntos de caminos diferentes, dentro del más cercano posible que cada punto puede tener. Para este trabajo, siempre que se hable de "distancia" o "error" se hace referencia a la distancia de Hausdorff en metros entre el camino predicho y el camino real.

5. Predicción de caminos

5.1. Experimento 1: Cálculo inicial de métricas

Línea recta

A la hora de realizar predicciones, se decidió utilizar como modelo predictor base la línea recta.

Es decir, tomar el punto inicial y el punto final del camino, y trazar una línea recta entre ellos.

En la *Figura 15* se puede ver el camino real en azul, y el camino línea recta en verde.

Figura 15. Comparación del camino real con la línea recta

Al realizar el cálculo de la distancia de Hausdorff, se obtiene que:

- Distancia total (d_H): 4.763.400 m.
- Distancia promedio ($\overline{d_H}$): 683,2 m.

Este resultado, al ser inicial, se puede tomar como base para comparación con otras técnicas. 683.2 metros de distancia es alentador, teniendo en cuenta las distancias de los caminos utilizados para la comparación.

Línea recta a 1 paso

Luego, aprovechando la grilla de 50x60 generada para el modelo, se prueba la posibilidad de hacer pasos de 1 casilla de manera horizontal, vertical o diagonal, teniendo en cuenta solo el punto de inicio y el punto de fin. La *Figura 16* muestra el camino generado de esta manera, oscilante buscando acercarse al punto final.

Figura 16. Comparación del camino real con la línea recta a 1 paso

Al realizar el cálculo de la distancia de Hausdorff, se obtiene que:

- d_H: 8.726.842 m.
- $\overline{d_H}$: 1251,7 m.

En comparación con la línea recta, los resultados son peores como era esperable. Sin embargo, 1.251,7 metros parece un buen resultado.

Distancia de Manhattan

Como variación del punto anterior, se probó Manhattan, que propone que los movimientos de 1 casilla sean solo horizontales o verticales, como se ve en la *Figura 17*.

Figura 17. Comparación de camino real con Manhattan

Al realizar el cálculo de la distancia de Hausdorff, se obtiene que:

- d_H: 8.613.866 m.
- $\overline{d_H}$: 1235,5 m.

Como es de esperar, los resultados son muy parecidos a la línea recta a 1 paso, aunque levemente mejores.

Modelo gravitacional

Al momento de probar el modelo en sí, lo primero que viene a la mente es intentar utilizar el modelo como fue calculado. Es decir, empezar en el punto inicial y luego seguir las pendientes de cada punto de manera que se busque el punto de mayor "densidad". Esto por supuesto da malos resultados, ya que en la mayoría de los casos ni siquiera se llega al destino, sino que se atasca en un mínimo local, como se puede ver en la *Figura 18*.

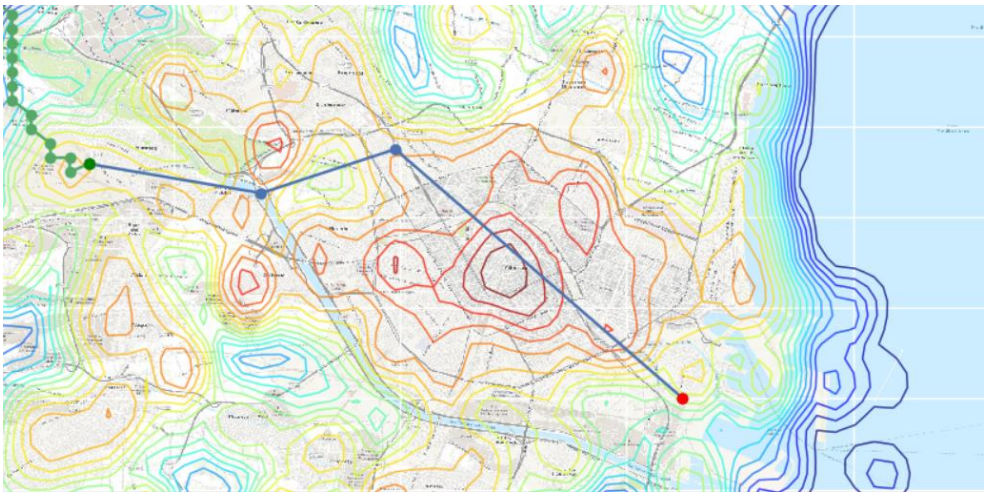

Figura 18. Comparación de camino real con modelo gravitacional

Los resultados dan métricas muy lejanas a las anteriores. Al realizar el cálculo de la distancia de Hausdorff, se obtiene que:

- d_H: 21.879.245 m.
- $\overline{d_H}$: 3138,2 m.

Los resultados son muy alejados a los obtenidos en las técnicas anteriores, y esto es esperable: no hay dirección desde el punto de inicio al de fin, y son muy pocos los caminos que llegan a destino.

Modelo de atracción

Para remediar el problema de los mínimos locales, se prueba con un modelo de atracción, donde al punto final del camino real se le aplica un peso que hace tender los gradientes de todos los puntos del mapa ligeramente hacia él.

La *Figura 19* muestra el efecto de aplicar atracción sobre el punto final (rojo) del camino original (azul). Las líneas de colores muestran la profundidad del mapa en esa zona, mientras que las flechas negras representan el gradiente en cada punto. Por supuesto, dependiendo del factor elegido, el camino no siempre llegará a destino ya que un peso pequeño no será capaz de eliminar los mínimos locales.

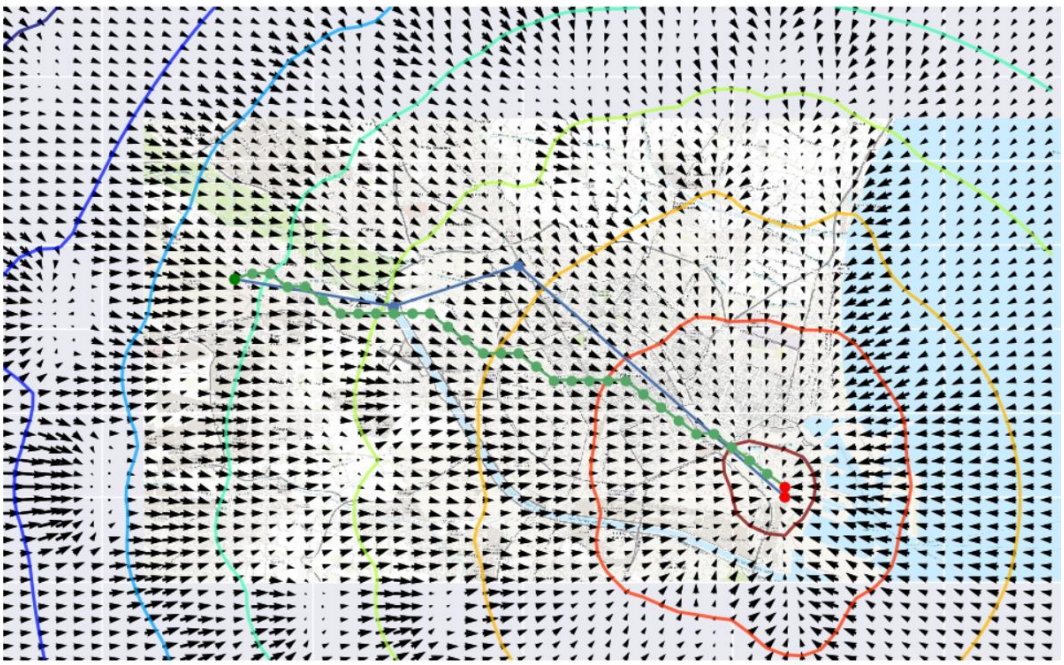

Figura 19. Comparación de camino real con camino de atracción

Al realizar pruebas con esta técnica, se probaron los siguientes posibles factores de atracción: [-5.000, -2.000, -1.000, -50, -10, 1, 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1.000, 2.000, 5.000]. Los primeros números, negativos, deberían alejar al camino del punto final, por

lo que un valor muy malo de ellos sería correcto. La *Tabla 7* muestra los resultados obtenidos al probar estos valores.

Factor A	d_H	$\overline{d_H}$
-5.000	41.832.405	6.000,1
-2.000	41.908.219	6.010,9
-1.000	42.255.785	6.060,8
-50	18.648.167	2.674,7
-10	17.252.762	2.474,6
1	16.986.703	2.436,4
10	16.640.672	2.386,8
50	15.653.835	2.245,2
100	13.794.129	1.978,5
200	9.907.931	1.421,1
300	8.998.555	1.290,7
400	8.965.059	**1.285,9**
500	9.015.788	1.293,1
600	9.060.297	1.299,5
700	9.088.653	1.303,6
800	9.107.345	1.306,3
900	9.121.401	1.308,3
1.000	9.126.864	1.309,1
2.000	9.131.151	1.309,7
5.000	9.143.947	1.311,5

Tabla 7. Resultados al variar el factor de peso en el modelo de atracción

Se puede observar que el factor que mejor funciona, en base a la distancia promedio que arroja, es el de 400. Así mismo, aumentar este factor más allá de 400 no parece variar mucho los resultados obtenidos, dando resultados levemente peores.

Geodésica

Otra alternativa de utilización del modelo gravitacional es el de encontrar el camino a través de la geodésica, buscando el trayecto de menor coste entre el punto de inicio y de fin teniendo en cuenta las pendientes que presenta cada punto. La *Figura 20* muestra un ejemplo de este camino, donde se puede observar cómo, desde el punto inicial al punto final, el camino generado (verde) intenta seguir los caminos más planos hasta llegar al destino.

Figura 20. Ejemplo de generación de un camino geodésico

Al calcular la distancia entre los caminos reales y los caminos predichos, los resultados obtenidos fueron:

- d_H: 9.628.385 m.
- $\overline{d_H}$: 1.380,8 m.

Los resultados obtenidos son buenos, aunque levemente peores que el del modelo de atracción. El hecho de que todos los caminos lleguen a destino representa una ventaja sobre el modelo gravitacional.

Comparación

La *Tabla 8* muestra todos los resultados obtenidos con cada una de las estrategias.

TÉCNICA	$\overline{d_H}$
Línea recta	683,2
Línea recta a 1 paso	1.251,7
Manhattan	1.235,5
Modelo gravitacional	3.138,2
Modelo de atracción	1.285,9
Geodésica	1.380,8

Tabla 8. Comparativa de resultados obtenidos

Se puede observar que la línea recta obtiene excelentes resultados en comparación con cualquiera de las otras técnicas. Además, en el modelo de atracción, los mejores resultados se dan cuando el factor de atracción es muy alto, convirtiendo a este predictor casi en una línea recta a 1 paso. Esto podría corresponder a la linealidad de los caminos originales, o que el modelo no es tan buen predictor. Esta linealidad en los caminos originales posiblemente no representa el verdadero camino que realizaron los sujetos, por lo que se procede al experimento 2.

5.2. Experimento 2: Análisis por largo del camino

Los valores obtenidos en el experimento 1 no fueron buenos, siendo que el mejor resultado propuesto por el modelo arroja valores del doble de error que la línea recta. La suposición hecha es que como los caminos son muy cortos, la línea recta siempre va a predominar. Por esto, se intentó realizar una comparación de métricas si se tomaran solo caminos de un cierto largo. La *Tabla 9* muestra los errores promedio obtenidos al limitar los caminos generados a un cierto número de puntos.

LARGO MÍNIMO	3	4	5	6	7
Número de caminos	6.972	1.096	218	55	13
Línea recta	683,2	970,8	1.099,5	1.008,7	1.619
Línea recta de a 1 paso	1.251,7	1.395,4	1.329,5	1.148,7	1.836,5
Manhattan	1.235,5	1.393,4	1.378,6	1.135,4	1.852,9
Modelo gravitacional	3.138,2	3.302,7	3.344,8	2.990,5	3.360,5
Geodésica	1.380,8	1.557,9	1.603,4	1.390,7	1.835,4
Atracción (mejor parámetro)	1.285,9 (400)	1.467,3 (300)	1.456,6 (300)	1.273,2 (300)	2.011 (300)

Tabla 9. Comparación de métricas al restringir el número mínimo de puntos de un camino

En la *Figura 21* se puede observar cómo, al aumentar las restricciones (el número mínimo de puntos para generar un camino) la cantidad de caminos generados decrece drásticamente. De la misma manera, los errores crecen un poco, aunque no es determinante. Esto puede devenir de que, posiblemente, los caminos de 3 o 4 puntos eran los que más tendían a ser una línea recta. Al quitar esos caminos, los que quedaron fueron más complejos, generando un error mayor promedio.

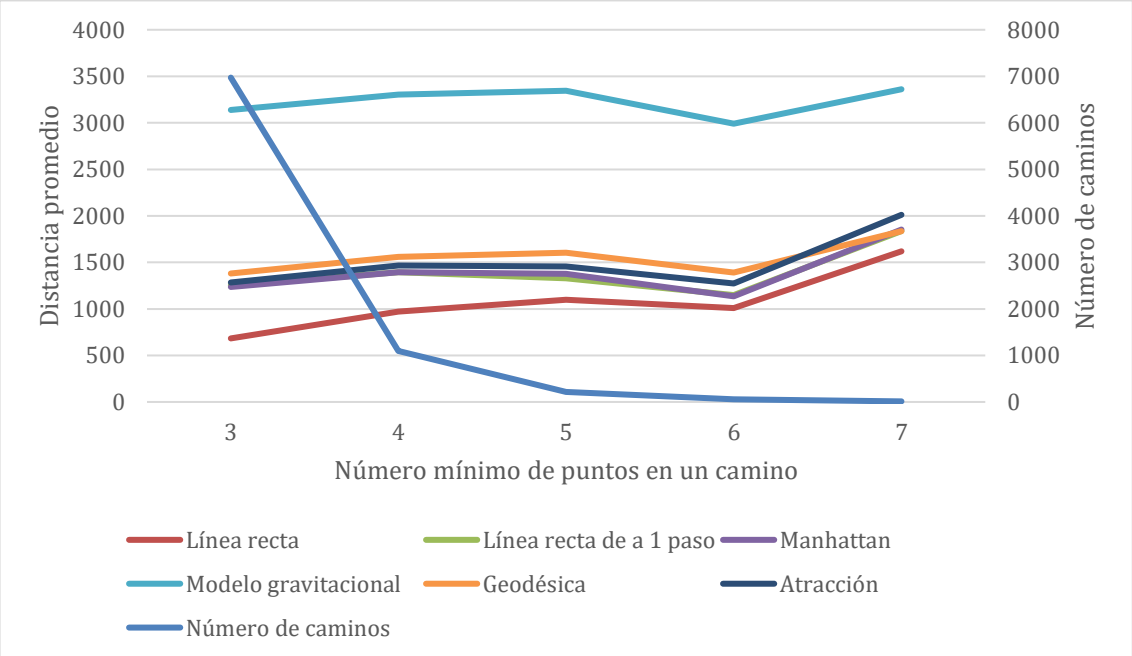

Figura 21. Comparación de métricas al restringir el número mínimo de puntos de un camino

5.3. Experimento 3: Reconstrucción de caminos

El siguiente experimento constó en quitar una de las restricciones al momento de generar los caminos: los caminos ya no se subdividen permanentemente si el ángulo de tres puntos consecutivos es menor a 90 grados, sino que se dividen a la hora de la predicción, pero se reconstruyen a la hora de comparar con los caminos reales. De esta manera, todos los *tweets* que haya hecho una persona durante un día formarán parte de un camino. Esto requería mantener caminos de 2 puntos, que sirven de conexión entre potenciales caminos más largos.

La *Figura 22* muestra como el camino original, [0,1,2,3,4,5,6,7] se divide en 3 caminos a la hora de la predicción: [0,1,2,3], [3,4,5,6] y [6,7]. Una vez realizada la predicción de estos 3 caminos por separado, estas predicciones se reconstruyen en base a los puntos de separación [3] y [6] así como el camino original, volviendo a [0,1,2,3,4,5,6,7]. Luego, se comparan el camino original reconstruido y el camino predicho reconstruido utilizando las técnicas mencionadas.

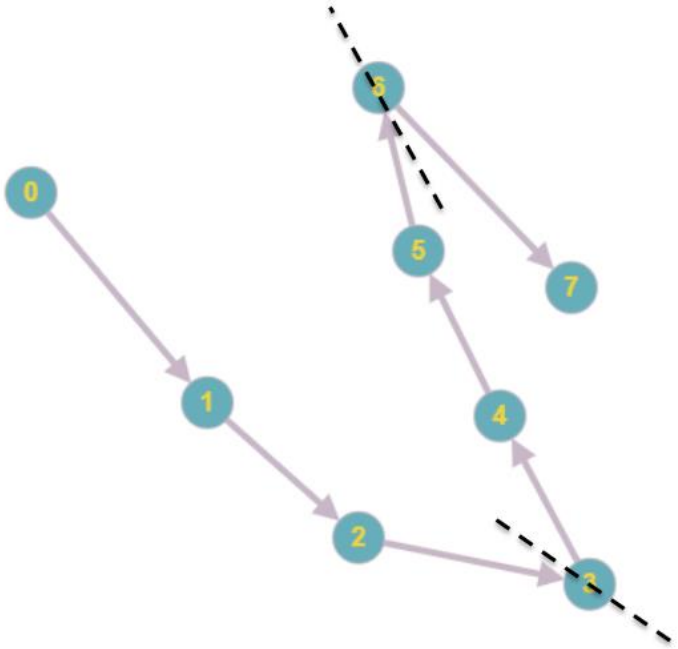

Figura 22. Ejemplo de reconstrucción de caminos

La *Tabla 10* y *Figura 23* muestra las cantidades y los largos de los caminos reales generados. Se puede ver que la mayoría de los caminos son de 2 y 3 puntos, y que incluso los caminos de menor cantidad de puntos presentan el mayor largo total.

NÚMERO DE PUNTOS	1	2	3	4	5	6	7	8	9	10	11
Distancia total	0	31.168	5.274	794	171	38	23	9	0	1	1
Porcentaje	0	83,2	14,1	2,1	0,5	0,1	0,06	0,02	0	0,003	0,003

Tabla 10. Generación de caminos para reconstrucción

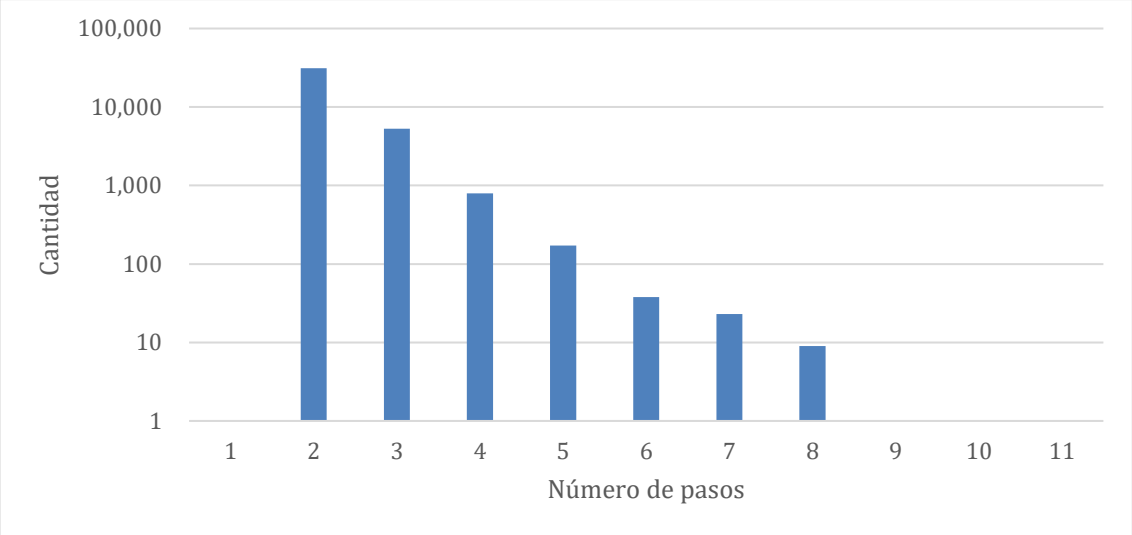

Figura 23. Generación de caminos para reconstrucción

Al realizar los experimentos sobre este conjunto de caminos, la línea recta obtuvo valores mucho mejores que el resto de las métricas, como puede verse en la *Tabla 11*.

TÉCNICA	$\overline{d_H}$
Línea recta	164,2
Línea recta a 1 paso	1180
Manhattan	1.167,4
Modelo gravitacional	2.085,7
Modelo de atracción	1.037,2
Geodésica	1.298,8

Tabla 11. Métricas al usar reconstrucción de caminos

Los valores obtenidos siguen siendo malos, y esto se debe principalmente a que, para poder garantizar la reconstrucción de caminos, se necesitan mantener muchos caminos de 2 puntos, que no son realistas a la hora de la predicción, ya que no habrá mejor predictor que la línea recta. Esto se acentúa por el hecho de que el 83.2% de los caminos son de 2 puntos.

5.4. Experimento 4: Caminos completos

Luego se decidió liberar completamente las restricciones de caminos, excluyendo solo a los caminos de 2 puntos. Se usaron todos los puntos donde una persona realizó un *tweet* a lo largo de un día para la creación del camino, su predicción y su comparación. De esta manera se libera la generación de caminos, potencialmente generando caminos más realistas. En la *Figura 24* se pueden observar los largos de los caminos generados.

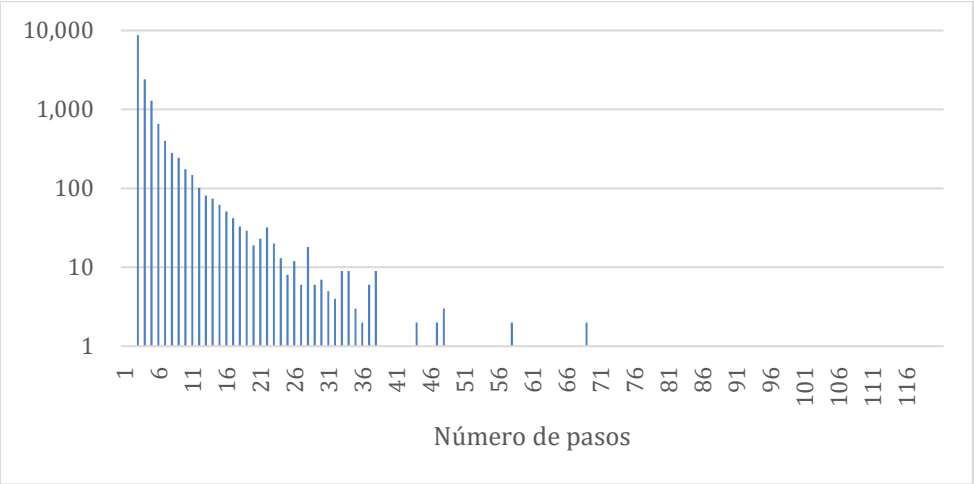

Figura 24. Largo de caminos completos

Al calcular las métricas para estos caminos, los resultados obtenidos vuelven a mostrar una ventaja de la línea recta, pero ya esa ventaja deja de ser abrumadora. Caminos más reales y complejos se adaptan mejor al modelo, como se puede observar en la *Tabla 12*.

TÉCNICA	$\overline{d_H}$
Línea recta	1.594,4
Línea recta a 1 paso	1.894,8
Manhattan	1.889,4
Modelo gravitacional	3.374,5
Modelo de atracción	1.901,5
Geodésica	2.652,9

Tabla 12. Comparación de métricas para caminos completos.

Nuevamente, los valores obtenidos no son muy buenos, lo que puede atribuirse a la cantidad de caminos cortos que no son realistas y fáciles de predecir con líneas rectas. Por esto, se decidió comparar las métricas en base a los diferentes valores de largo mínimo de un camino, como se hizo en el Experimento 2. Los resultados pueden observarse en la *Tabla 13* y en la *Figura 25*.

LARGO MÍNIMO	4	5	6	7	8	9	10	15	20
Número de caminos	6.318	3.911	2.616	1.960	1.559	1.278	1.033	453	236
Línea recta	1.463,4	1.708,7	1.818,6	2.000,4	2.105,2	2.219,3	2.346,1	2.810,5	3.093,1
Línea recta a un paso	1.896,9	2.086,7	2.226,9	2.400,8	2.524,4	2.627,4	2.769,1	3.255,2	3.553,1
Manhattan	1.890,5	2.088,7	2.227,4	2.409,1	2.532,6	2.637,8	2.780,9	3.265,8	3.581
Gravitacional	2.844,2	2.910,4	3.077,1	3.198,8	3.346,3	3.424,1	3.563,6	4.072	4.372,7
Geodésica	2.899,9	3.230,6	3.140,6	3.333,5	3.365	3.500,5	3.582,3	3.878	4.091,6
Atracción (factor)	1.907,1 (400)	2.098,2 (400)	2.236,5 (400)	2.413,1 (400)	2.536,4 (400)	2.639,6 (400)	2.783,1 (400)	3.273,5 (300)	3.566,8 (5000)

Tabla 13. Comparativa de métricas al restringir el número mínimo de puntos de un camino

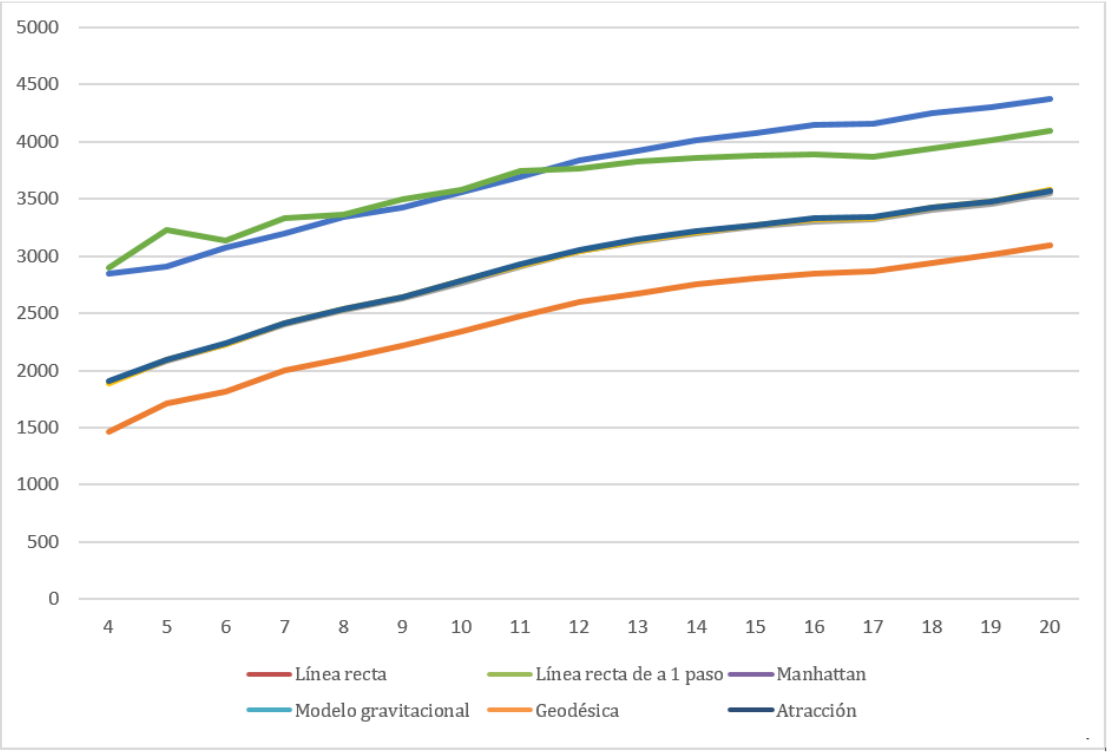

Figura 25. Comparación de métricas al restringir el número mínimo de puntos de un camino

Los valores obtenidos son peores a medida que aumenta la cantidad mínima de puntos del camino, y esto puede atribuirse a que los valores pequeños son fáciles de predecir, y al descartarlos, el promedio de error sube.

5.5. Experimento 5: Utilización de rutas generadas

Los experimentos anteriores hicieron llegar a la conclusión de que los *tweets* que realiza una persona no son suficiente información para la generación del camino que recorre al desplazarse. Sin embargo, esta es la única información disponible.

Para poder atacar este problema, se decide reemplazar la forma en la que se unen 2 puntos (*tweets*) consecutivos. En vez de utilizar una línea recta, se utiliza un camino real, generado por un servicio de planificación de rutas. Este generador ignora los factores como

embotellamiento, preferencias, y otros que el modelo intenta predecir, por lo que da un buen punto de partida.

Se utilizó un servidor de la UPV montado con una instancia de Project OSRM [18], a la cual se le hicieron solicitudes para la generación de los nuevos caminos. La *Figura 26* muestra un camino generado por *tweets*, como se realizaban anteriormente, mientras que la *Figura 27* muestra un camino generado por OSRM utilizando los *tweets*.

Figura 26. Camino generado por tweets

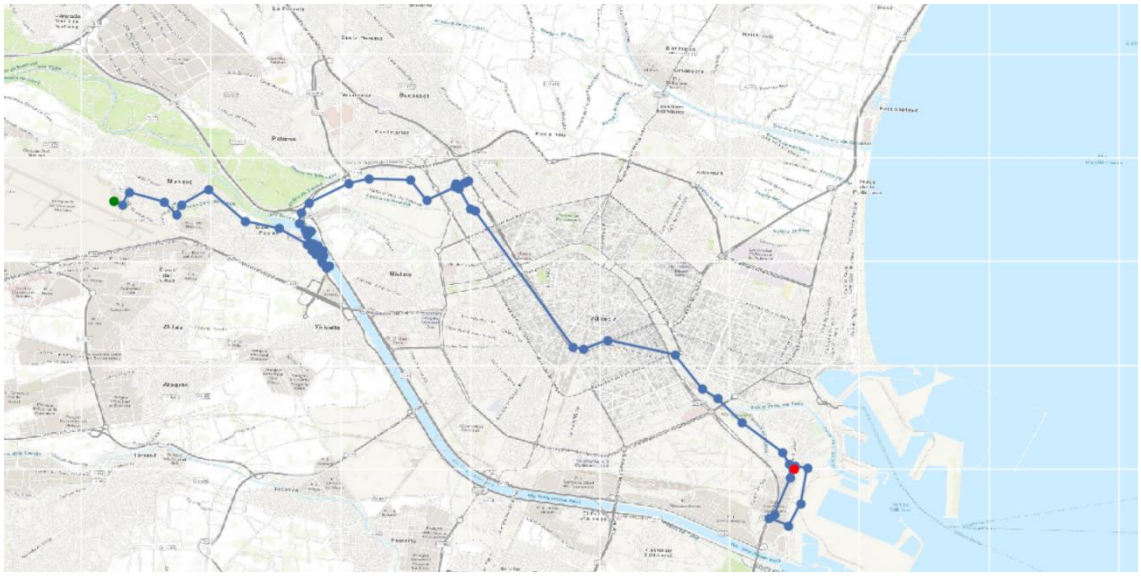

Figura 27. Camino generado por OSRM utilizando puntos de tweets

De esta manera, los caminos originales son reemplazados por caminos más detallados. Al realizar un análisis del largo del camino en la *Figura 28*, se observa una distribución que tiene más sentido al pensar en el *"trayecto de una persona entre 2 puntos".*

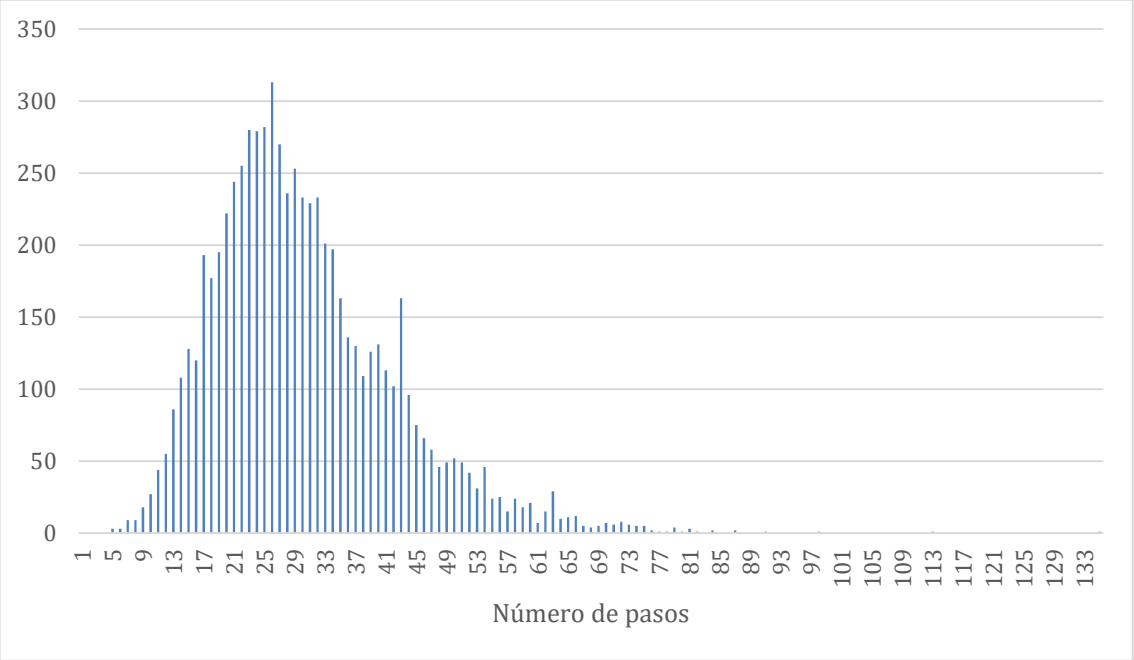

Figura 28. Largo de caminos utilizando rutas reales

Al realizar los cálculos de las métricas con las diferentes técnicas, la *Tabla 14* muestra que el modelo de atracción obtiene resultados mejores que las demás técnicas, pero con un factor muy alto: se comporta casi como una línea recta a 1 paso.

TÉCNICA	DISTANCIA PROMEDIO
Línea recta	1.185,1
Línea recta a 1 paso	1.196,7
Manhattan	1.203,9
Modelo gravitacional	3.384,1
Geodésica	1.417,9
Modelo de atracción	1.176,7 (1.000)

Tabla 14. Comparativa de métricas al utilizar rutas reales

Es interesante analizar la cantidad de caminos que no llegan a destino usando el modelo gravitacional: 6.675 de 6.972. Esto sucede ya que los caminos caen en mínimos locales de los cuales no pueden salir, porque obedecen a los gradientes. Sin embargo, se puede ver como este número va decreciendo a medida que aumenta el factor de atracción: cada vez más caminos llegan a destino si el punto de fin tiene un poder gravitacional muy alto, como se ve en la *Tabla 15*.

PARÁMETRO	-5.000	-10	1	10	50	100	500	1.000	2.000	5.000
ERRORES	6.675	6.384	6.273	6.049	5.118	2.855	634	597	544	513

Tabla 15. Comparativa de caminos fallidos cuando aumenta el factor de atracción

Para intentar paliar esto, se intenta no permitir puntos repetidos a la hora de predecir caminos. Cuando esto fuese a suceder, se toma un paso en otra dirección, así intentando siempre explorar otras opciones (a diferencia de interrumpir la generación del camino como sucedía en experimentos anteriores).

Al calcular métricas con diferentes parámetros de atracción, se puede observar un decremento de la cantidad de caminos fallidos, a la vez que decrece el error. El mejor resultado llega en el punto de inflexión donde casi todos los caminos son predichos correctamente, como se muestra en la *Tabla 16*. Un factor de atracción de 40 genera correctamente todos los caminos excepto 2. Es interesante ver que, si se incrementara el valor del factor, el error crecería, aunque todos los caminos se sigan generando correctamente. Esto está relacionado con que, si el factor es extremadamente grande, los caminos generados sería líneas rectas que no se amoldarían al camino real.

PARÁMETRO	DISTANCIA PROMEDIO	CAMINOS FALLIDOS
-5.000	6.929,7	6.570
-2.000	6.933,1	6.570
-1.000	6.926,5	6.570
-50	6.748,6	6.570
-10	5.704,5	6.169
1	2.907,1	3.777
10	1.561,2	990
20	1.286,5	137
30	1.212,5	16
40	**1.197,5**	2
50	1.201,8	0
100	1.210,9	0
500	1.203,9	0
1.000	1.202,8	0
2.000	1.202,5	0
5.000	1.202,8	0

Tabla 16. Distancia promedio y caminos fallidos al incrementar el factor de atracción

Esto lleva a la conclusión de que el mejor resultado posible se obtiene al utilizar un modelo de atracción con un factor de 40. Las figuras 29 a la 37 muestran ejemplos de caminos generados con el modelo de atracción con este factor en comparación con el camino real.

Figura 29. Ejemplo de un camino original (azul) y un camino generado (verde)

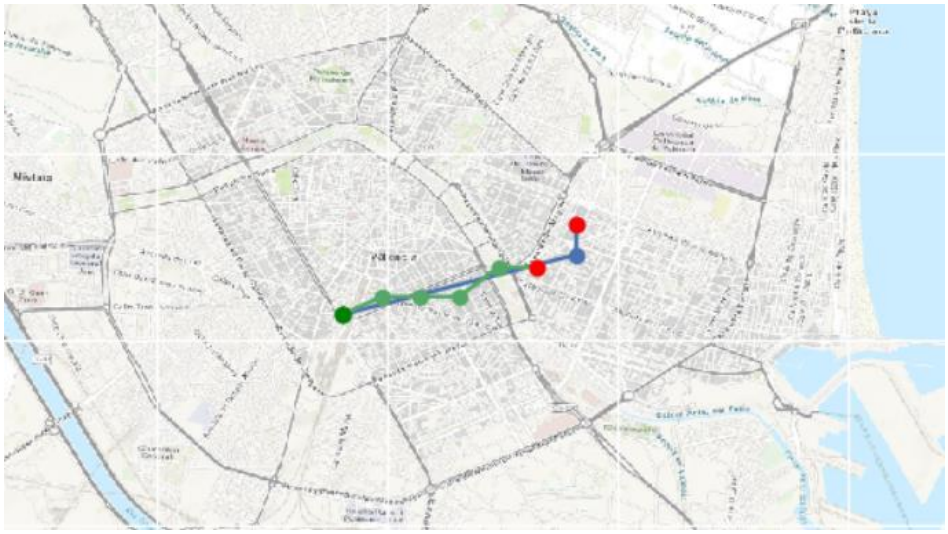

Figura 30. Ejemplo de un camino original (azul) y un camino generado (verde)

Figura 31. Ejemplo de un camino original (azul) y un camino generado (verde)

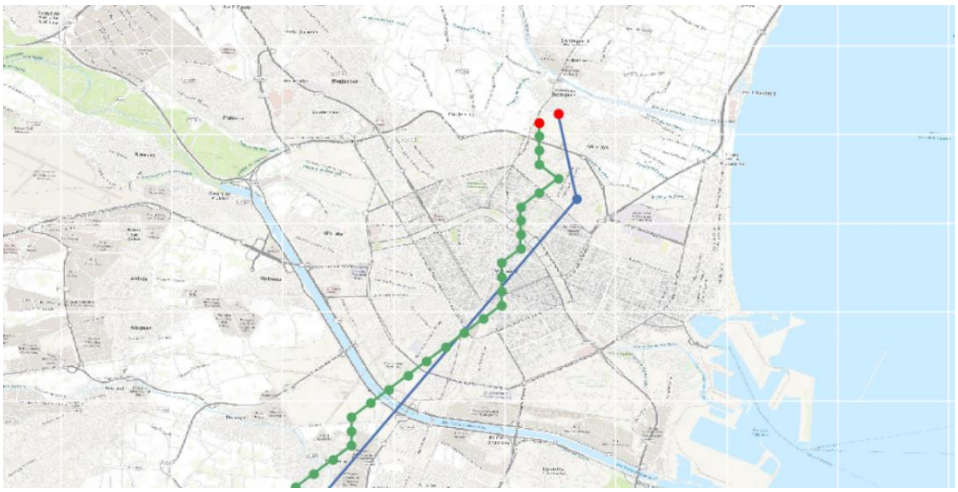

Figura 32. Ejemplo de un camino original (azul) y un camino generado (verde)

Figura 33. Ejemplo de un camino original (azul) y un camino generado (verde)

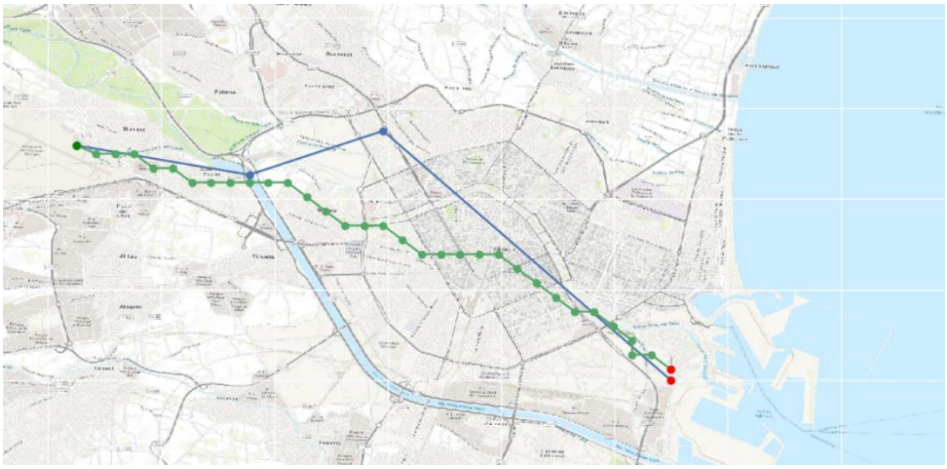

Figura 34. Ejemplo de un camino original (azul) y un camino generado (verde)

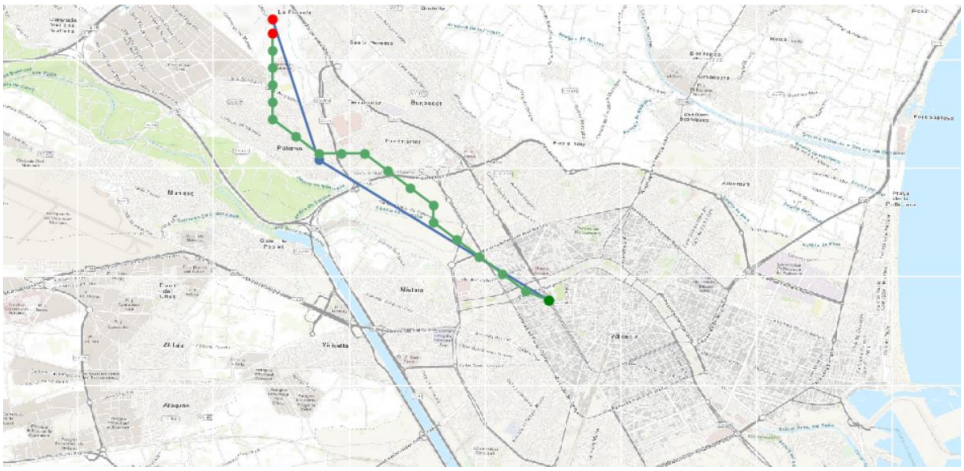

Figura 35. Ejemplo de un camino original (azul) y un camino generado (verde)

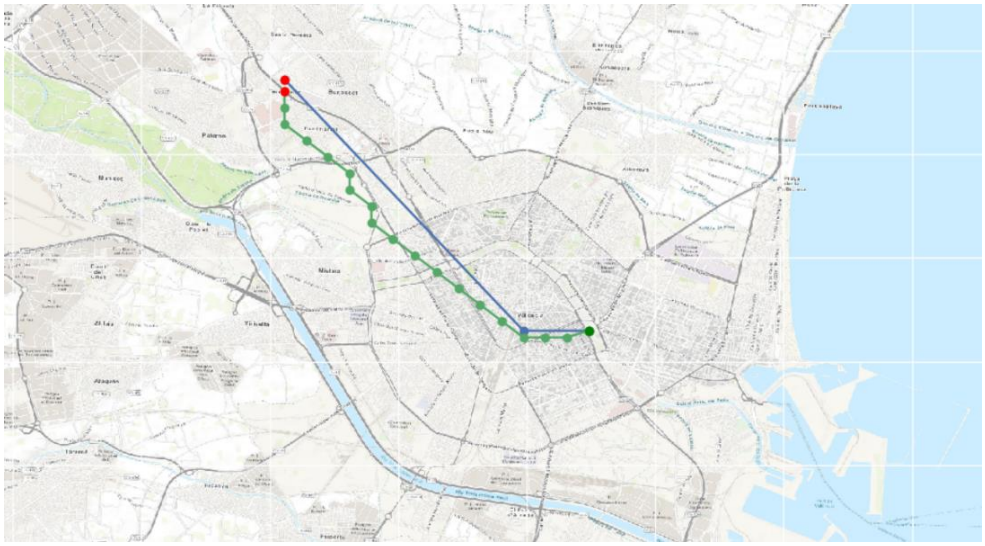

Figura 36. Ejemplo de un camino original (azul) y un camino generado (verde)

Figura 37. Ejemplo de un camino original (azul) y un camino generado (verde)

5.6. Evaluación de los resultados

En este capítulo se realizaron pruebas para la generación de caminos utilizando *tweets*, y se midieron esas pruebas en base a las 6 técnicas explicadas (Línea recta, Línea recta a 1 paso, Manhattan, Modelo gravitacional, Manhattan y Geodésica). Los resultados obtenidos dan a entender que es de vital importancia contar con una buena calidad de datos de caminos reales

para poder contrastar con los generados por el modelo. Es por esto que los mejores resultados se obtuvieron al contrastar los caminos generados por el modelo con los obtenidos a través del servicio de navegación, que es el que más detalle presentaba.

En general, cuando el camino real es suficientemente complejo, el modelo de atracción suele adaptarse bien, y la geodésica se encuentra muy cerca. Sin embargo, el modelo gravitacional no suele presentar buenos resultados, lo cual tiene sentido ya que no hay una tendencia explícita del punto de inicio al punto de fin.

6. Comparación entre géneros

Diversos estudios han analizado la diferencia en el comportamiento de las personas en base a su género, incluso en lo relacionado al transporte y la movilidad [19-23]. Estos estudios en general concluyen que hay una diferencia entre los caminos que siguen un hombre y una mujer a la hora de transportarse, y es lo que se intenta probar con el modelo generado en este trabajo. Se intentará rechazar la hipótesis nula H_0: "*no hay diferencias en los patrones de movilidad de hombres y mujeres*".

Aprovechando las capacidades de los datos utilizados para generar el modelo, es posible clasificarlos para realizar un análisis descriptivo de los grupos generados al dividir por género. Para la clasificación de los caminos por género, se tomó la información de perfil de los usuarios de Twitter que fueron usados para generar los caminos. Como la API no provee información de género, se debió utilizar otra estrategia para su distinción. Se utilizó el *Nombre* de la cuenta, que fue comparado con la lista de nombres y géneros tomada del proyecto GenderReader [24]. Este archivo provee una lista internacional de nombres, con su género más probable asociado.

Por supuesto, no todas las personas usan su nombre real a la hora de hacer una cuenta de Twitter, por lo que no se pudo obtener información de género de todos los usuarios. Del total de caminos generados anteriormente, 6.972, solo se pudieron clasificar 2.826, como muestra la *Figura 38*. En total, se pudieron clasificar 1.910 caminos de hombres y 916 caminos de mujeres.

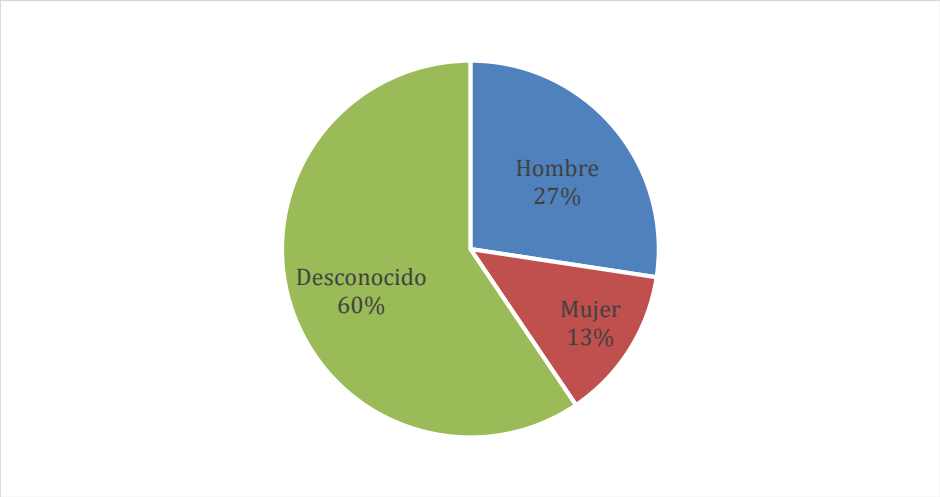

Figura 38. Distribución de género en los caminos

6.1. Experimento 1: Longitud de los caminos

La *Figura 39* y la *Figura 40* muestran la distribución de largos de caminos para hombres y mujeres, respectivamente. Se puede observar que presentan una distribución parecida, con una ligera inclinación a caminos largos por parte del género masculino. Los hombres presentan una media de 4,73 km, mientras que el de las mujeres es de 4,15 km. La desviación estándar en el caso de los hombres es de 4,1 km y 3,8 km en el caso de las mujeres.

Al realizar el test T de Student para analizar las distribuciones de las distancias de los caminos de estos 2 grupos, se obtiene un p valor de 0.000334. Esto significa que se rechaza la hipótesis nula, reafirmando que existen diferencias entre los largos de caminos de hombres y mujeres.

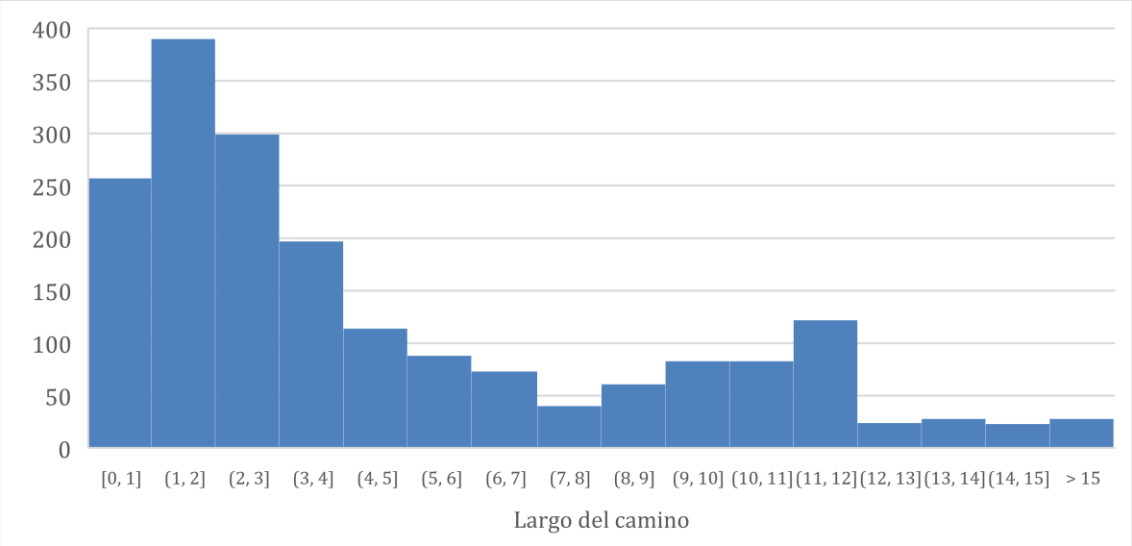

Figura 39. Distribución de largos de caminos para el género masculino

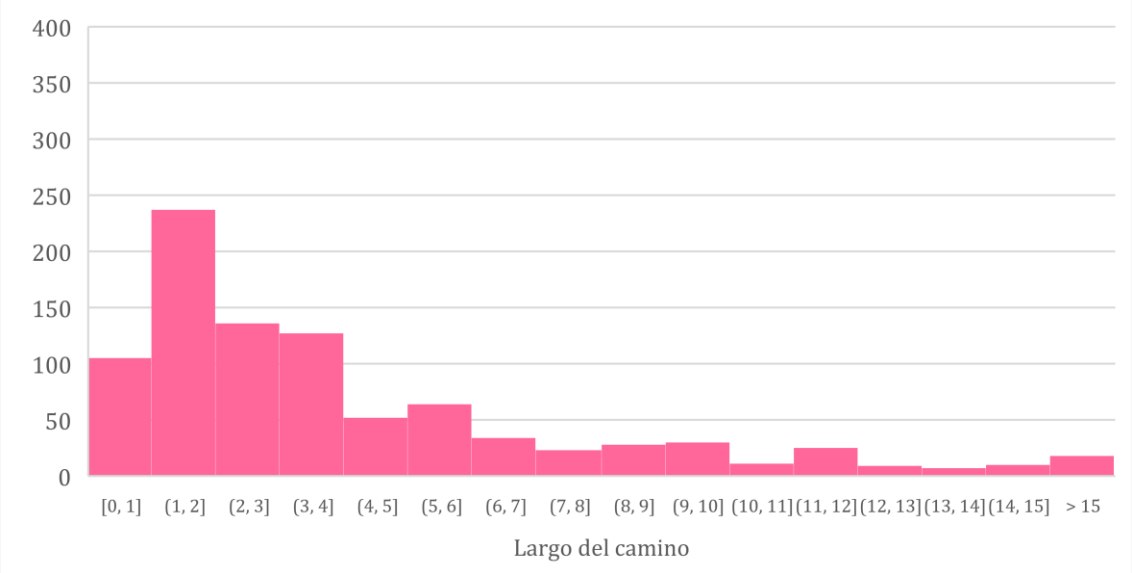

Figura 40. Distribución de largos de caminos para el género femenino

6.2. Experimento 2: Predicción de caminos por género

Al generar métricas específicas de acuerdo a las técnicas utilizadas anteriormente, se ve una pequeña diferencia, como se aprecia en la *Tabla 17*.

TÉCNICA	d_HHOMBRE	d_HMUJER
Línea recta	526,3	508,7
Línea recta a 1 paso	1.239,1	1.131,6
Manhattan	1.217,2	1.069,3
Modelo gravitacional	3.148,4	2.857,1
Modelo de atracción	1.253,9	1.121,7
Geodésica	1.280,6	1.166,9

Tabla 17. Resultados de comparar métricas por género

Tomando la mejor configuración del modelo de atracción, con factor igual a 100, se generaron las distribuciones de las distancias de Hausdorff, como se puede apreciar en la *Figura 41* y *Figura 42* que muestran estas estadísticas para hombres y para mujeres, respectivamente.

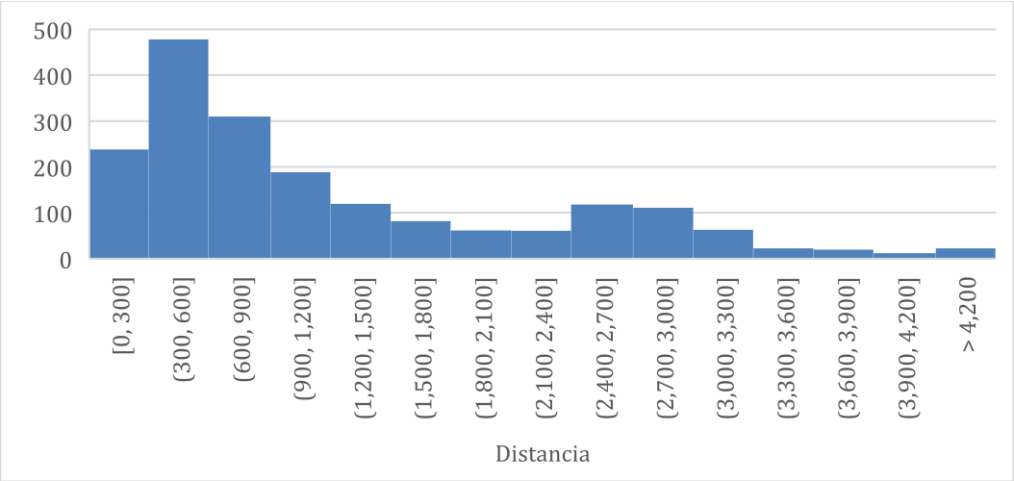

Figura 41. Distribución de distancias de Hausdorff para hombres

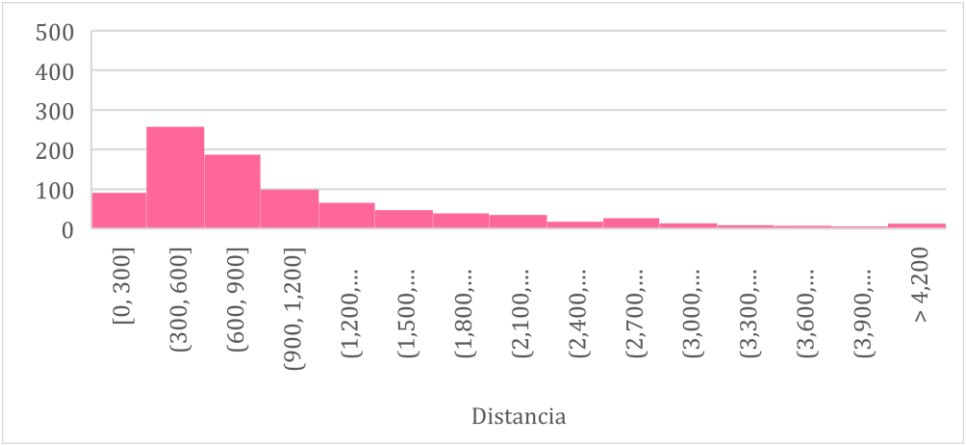

Figura 42. Distribución de distancias de Hausdorff para mujeres

Se puede observar que la distribución es muy parecida, con un pequeño sesgo hacia los errores más grandes para el grupo de los hombres. El promedio de las distancias de Hausdorff para el caso de los hombres es de 1,25 km y de 1,12 km para el caso de las mujeres. En el caso de la desviación estándar, tiene un valor de 1,06 km para los hombres y de 0,99 km para las mujeres.

Se tomó esta configuración (modelo de atracción con un factor de 100) y se realizó el test T de Student para entender su distribución. El p valor obtenido fue de 0.0014 por lo que se puede rechazar la hipótesis nula, reafirmando que existen diferencias entre los errores a la hora de predecir caminos de hombres y mujeres.

6.3. Experimento 3: Sinuosidad de los caminos

Para entender mejor estos resultados, se calculó la medida de Sinuosidad promedio de los caminos. Esta métrica es el índice que representa cuánto el trazado de un camino se desvía de la línea recta. El cálculo de la sinuosidad *S* para el camino *p* está dada por:

$$S_p = \frac{len(p)}{\sqrt{\left(p_{ix} + p_{fx}\right)^2 + \left(p_{iy} + p_{fy}\right)^2}}$$

Siendo: *len(p)* el largo del camino p; p_{ix} la coordenada x del punto inicial; p_{fx} la coordenada x del punto final; p_{fx} la coordenada y del punto inicial; p_{fy} la coordenada y del punto final. Los resultados de sinuosidad promedio para los hombres es de 1,168 (desviación estándar de 0,66), y para las mujeres de 1,201 (desviación estándar de 1,6). Es decir, los caminos de las mujeres presentan mayor sinuosidad que los de los hombres. Al realizar el test T de Student para estas distribuciones se obtiene un p valor de 0.042

Como una métrica relacionada, se calculó el promedio de los ángulos de cada camino de cada grupo. Es decir, por cada camino, se tomaban tres puntos consecutivos y se acumulaba el ángulo generado por los dos segmentos. En la *Figura 43* se puede observar un ejemplo de camino, cuyo ángulo total acumulado sería de $\alpha + \beta$ y su promedio de $(\alpha + \beta)/2$. Al realizar este cálculo para todos los caminos y promediar se obtuvo que el ángulo promedio de los hombres es de 26,1 grados con desviación estándar de 86,39, mientras que el de las mujeres es de 32,2 grados con desviación estándar de 76,38. Al realizar el test T de Student da un p valor de 0.006, por lo que, en conjunto con el valor obtenido para la sinuosidad, se rechaza la hipótesis nula de que no existen diferencias entre la sinuosidad de los caminos de hombres y mujeres.

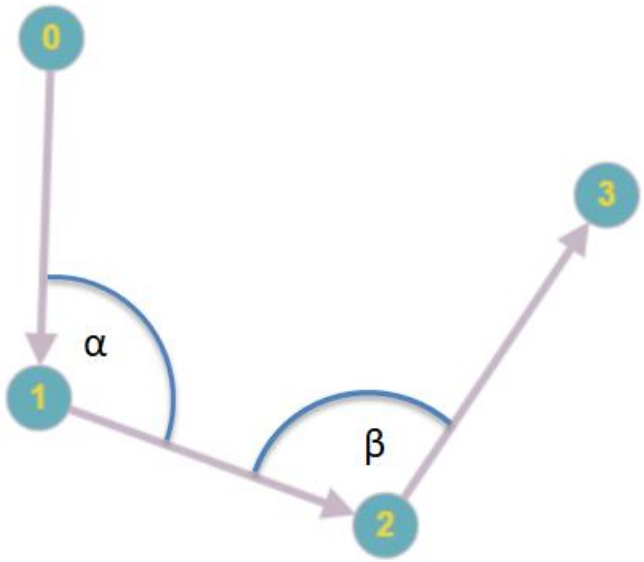

Figura 43. Ángulos de un camino

6.4. Evaluación de los resultados

Estas métricas apoyan con evidencias las hipótesis sobre las diferencias de género en los tipos de desplazamiento urbano. Muestran que los hombres usualmente toman caminos más largos, de ángulos más cerrados, pero menos sinuosos. En cambio, las mujeres toman caminos ligeramente más cortos, aunque de ángulos más abiertos y más sinuosos.

Al comparar esto con los resultados de predicción obtenidos, resulta interesante ver que la predicción de caminos para el grupo de mujeres tiene menor error. Esto, en conjunto con el resultado que dice que las mujeres realizan caminos cortos y de ángulos abiertos y sinuosos, hace entender que el modelo se aplica mejor para este tipo de trayectos. Esto de alguna manera tiene sentido, ya que, dada la complejidad del modelo gravitacional, y la cantidad de mínimos/máximos generados, intentar generar un camino de una gran distancia puede ser dificultoso. En cambio, caminos cortos y sinuosos se adaptan más al modelo, ya que obedecen a los pocos máximos/mínimos que pueden encontrarse entre el punto de inicio y el punto de fin.

7. Conclusiones

En este trabajo se han estudiado los patrones de movilidad de una persona en una ciudad a través de su actividad en redes sociales. Para ello, se han presentado y explicado publicaciones relacionadas al tema de predicción de patrones de movilidad urbana, y se ha hecho un resumen comparativo entre ellas. Luego se han generado varios modelos de actividad en la ciudad de Valencia, España, a través de datos obtenidos de Twitter. Se generaron caminos de individuos utilizando los mismos datos, y se compararon estos con predicciones hechas con el modelo. Por último, se realizó una comparación de predicción por género, y se estudiaron los resultados. Es por esto que se han alcanzado los objetivos propuestos en la introducción.

En general, el modelo gravitacional con atracción o geodésico parecen predecir buenos caminos ajustando los parámetros correspondientes. Sin embargo, a la hora de contrastar los caminos predichos contra caminos reales, se puede ver que los resultados podrían mejorar aplicando nuevas técnicas.

Al realizar un análisis de los caminos generados para la comparación, resulta que la mayoría de los caminos reales utilizados no eran buenos. Y esto tiene sentido, ya que, como se usó datos de redes sociales, estos no son objetivos. La gente realiza un *tweet* en los puntos en que se encuentra quieto, pero no necesariamente durante el camino. Se puede saber que una persona fue de un punto A a un punto B, pero no se puede saber por dónde pasó.

Es por esto, tal vez, que los mejores resultados obtenidos fueron cuando este camino generado de Twitter se ampliaba con caminos reales generados por un servicio de navegación urbano. Es decir, mientras más información se tenga de los caminos reales, mejores resultados se obtienen cuando se compara con los caminos predichos. Esto es alentador: mientras más real sea el camino original, mejor es el predictor.

Potencialmente, si se tuviera información perfecta del camino por el que se desplazó una persona, mejor serían los resultados al comparar con el camino predicho por el modelo desarrollado en este trabajo.

Por último, al comparar la predicción de caminos basada en género, los resultados concluyeron que los caminos generados por el género femenino presentan un error de predicción levemente menor que el del género masculino, y que las mujeres presentan, en promedio, caminos más cortos y sinuosos. Estos resultados validan los estudios de géneros que se han realizado en este campo y dan a entender que el modelo se ajusta mejor a este tipo de caminos, y que presenta dificultades para predecir caminos más largos y de ángulos más cerrados.

8. Referencias

1. Béjar Alonso Javier, Álvarez Napagao Sergio, Garcia Gasulla Dario, Gómez Sebastià Ignasi, et al. (2015) Discovery of Spatio-Temporal Patterns from Location Based Social Networks. doi: 10.1080/0952813X.2015.1024492
2. Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, et al. (2015) Geo-located Twitter as the proxy for global mobility patterns. doi: 15230406.2014.890072
3. Katarzyna Siła-Nowicka, Jan Vandrol, Taylor Oshan, Jed A. Long, et al. (2016) Analysis of Human Mobility Patterns from GPS Trajectories and Contextual. doi: 10.1080/13658816.2015.1100731
4. Fernando Terroso-Sáenz, Jesús Cuenca-Jara, Aurora González-Vidal, Antonio F. Skarmeta (2016) Human Mobility Prediction Based on Social Media with Complex Event Processing. doi: 10.1177/155014775836392
5. Hao Jiang, Qian Li, Xian Zhou, Yanqiu Chen, et al. (2017) A collective human mobility analysis method based on data usage detail records. doi: 10.1080/13658816.2017.1370715
6. Hasan A. H. Naji, Chaozhong Wu, Hui Zhang (2017) Understanding the Impact of Human Mobility Patterns on Taxi Drivers' Profitability Using Clustering Techniques: A Case Study in Wuhan, China. doi: 10.3390/info8020067
7. Marta C. González, César A. Hidalgo, Albert-László Barabási (2008) Understanding individual human mobility patterns. doi: 10.1038/nature06958
8. Mariano G Beiró, André Panisson, Michele Tizzoni, Ciro Cattuto (2016) Predicting human mobility through the assimilation of social media traces into mobility models. doi: 10.1140/epjds/s13688-016-0092-2
9. Carlos Andre Reis Pinheiro (2014) Revealing Human Mobility Behavior and Predicting Amount of Trips Based on Mobile Records. Doi: 10.1080/23249935.2017.1412370
10. Samiul Hasan, Xianyuan Zhan, Satish V. Ukkusuri (2013) Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. doi: 10.1145/2505821.2505823
11. Nibir Bora, Vladimir Zaytsev, Yu-Han Chang, Rajiv Maheswaran. (2013) Spatiotemporal Patterns in Social Networks

12. Daniele Barchiesi, Tobias Preis, Steven Bishop, Helen Susannah Moat (2015) Modelling human mobility patterns using photographic data shared online. doi: 10.1098/rsos.150046
13. 80% of Twitter's Users are Mobile. In Statista - The Statistics Portal. Visitado el 01/06/2018. https://www.statista.com/chart/1520/number-of-monthly-active-twitter-users/
14. Barthélemy, Marc (2006). Spatial Networks. 26-31.
15. D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge (1993). Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(9):850– 863
16. Daniel P. Huttenlocher, Klara Kedem, Jon M. Kleinberg (1992). On dynamic Voronoi diagrams and the minimum Hausdorff distance for point sets under Euclidean motion in the plane. The eighth annual symposium on Computational geometry.; 110-119
17. Lockett H, Guenov M (2008). Similarity measures for mid-surface quality evaluation. Comput Aided Design. 40(3): 368- 380.
18. Luxen, D., and C. Vetter (2011). Real-time routing with OpenStreetMap data. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 513–516. GIS '11. New York, USA: ACM.
19. Best, Henning (2005). Division of labour and gender differences in metropolitan car use: An empirical study in Cologne, Germany. doi: https://doi.org/10.1016/j.jtrangeo.2004.04.007
20. Hamilton Kerry, L Jenkins (2000). A Gender Audit for Public Transport: A New Policy Tool in the Tackling of Social Exclusion. doi: https://doi.org/10.1080/00420980020080411
21. Turner, Jeff (2000) Gender and Time Poverty: The Neglected Social Policy Implications of Gendered Time, Transport and Travel. doi: https://doi.org/10.1177/0961463X00009001007.
22. Peter Gordon, Ajay Kumar, Harry W. Richardson (1988). Gender Differences in Metropolitan Travel Behaviour. doi: https://doi.org/10.1080/00343408912331345672.
23. Peter Turnbull, Julia Lear, Huw Thomas (2013). Women in the transport sector. Transport Policy Brief ILO.
24. Jörg, Michael. Archivo "nam_dict.txt". Visitado el 01/06/2018. https://github.com/cstuder/genderReader