UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARACTERIZACIÓN Y RECONOCIMIENTO DE USUARIOS A TRAVÉS DE LA OBSERVACIÓN DE SU MOVILIDAD EN TRANSPORTE PÚBLICO

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS MENCIÓN COMPUTACIÓN
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL EN COMPUTACIÓN

CATALINA ANDREA ESPINOZA INAIPIL

PROFESORES GUÍA:

BENJAMÍN BUSTOS CÁRDENAS
MARCELA MUNIZAGA MUÑOZ

MIEMBROS DE LA COMISIÓN:

PABLO BARCELÓ BAEZA
SERGIO OCHOA DELORENZI
JORGE RIVERA CAYUPI

Este trabajo ha sido parcialmente financiado por el Instituto de Sistemas Complejos de Ingeniería (CONICYT–PIA_FB0816), el Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT 116_1589, FONDEF D10E_1002), y el programa de estadías cortas de investigación del departamento de postgrado y postítulo de la Vicerrectoria de Asuntos Académicos de la Universidad de Chile

---

SANTIAGO DE CHILE
2017

Resumen

La integración de tarjetas inteligentes en sistemas de transporte público ha permitido que los operadores, autoridades e investigadores, tengan acceso a una mejor perspectiva del servicio. La colección diaria de transacciones de tarjetas inteligentes trae consigo problemas propios del manejo de grandes volúmenes de datos: problemas metodológicos y nuevos desafíos. Por ejemplo, para hacer un análisis longitudinal del comportamiento de los usuarios, se necesitan los registros de transacciones en largos periodos de tiempo. No obstante, no todos los sistemas de transporte asocian información del usuario a las tarjetas inteligentes. Luego, las tarjetas se vuelven intercambiables, y no es posible asegurar la relación uno a uno entre usuarios y tarjetas. Este problema se vuelve una limitación mayor cuando el sistema presenta una alta renovación de tarjetas.

El objetivo de esta tesis es medir la estabilidad y unicidad del comportamiento de usuarios de transporte público, con el fin de evaluar la posibilidad de reconocer a los usuarios a través de su movilidad. Para lograr lo anterior, se implementan tres algoritmos de caracterización y comparación de la movilidad humana: dos adaptaciones de algoritmos de la literatura, y uno diseñado y calibrado en esta tesis. Los tres algoritmos se evalúan bajo dos perspectivas:

1. Cuán variable es la movilidad de los usuarios a través del tiempo.
2. Cuán característica es la movilidad de los usuarios respecto a sus pares.

Para medir cuán variables son los usuarios se utilizó una base de datos de registros de dos años del sistema de transporte público de Gatineau, Canadá. Para medir la capacidad de reconocer usuarios se utilizó una base de datos de registros de dos semanas, separadas por un intervalo de cinco meses, del sistema de transporte público de Santiago, Chile.

Los tres algoritmos reportan diferencias en el grado de variabilidad de los usuarios y en la capacidad de distinguirlos según su movilidad. Sin embargo, se observa de manera transversal que a mayor cercanía entre los periodos observados menor es la variabilidad medida. Del mismo modo se observa que a mayor tamaño de los periodos comparados, mayor es la estabilidad de la movilidad. Por otra parte, si bien los usuarios presentan una alta variabilidad en la movilidad, la mayoría posee una componente estable en el tiempo. En relación a la capacidad de reconocer a los usuarios, se observa que hay un grupo de usuarios distinguible por los tres algoritmos, del mismo modo, hay usuarios no distinguibles por ninguno. Se concluye que es posible reconocer a usuarios mediante su movilidad en transporte público, pero con una alta tasa de error.

A los momentos que no se viven por vivir estos. Especialmente, a la falta de peleas con Valentina o Belén.

Agradecimientos

Nadie debiese mirar hacia el pasado si no es suficiente paciente para antes recordar al menos a la mitad de sus abuelos. La verdad es que en estos años de universidad he visitado menos de lo que debería a mis abuelos, por eso, un poco como disculpas, mis primeros agradecimientos son a ellos. Gracias Coni por su constante preocupación y cariño, por recibirme en su casa y por aguantar mis ausencias. Gracias Tata por los infinitos tableros, sin duda contribuyeron a mi formación personal y profesional.

Antes de continuar con los agradecimientos personales, me gustaría agradecer a todos quienes han leído, corregido o aportado de alguna forma a esta investigación.

Gracias a Andrés Riquelme. Mirando hacia atrás, me parece que marcó el inicio de los sucesos que determinarían esta tesis.

Gracias a mis amigas y amigos de Santiago. Agradezco a mis amigas de handball, juntarme con ustedes fue fundamental para mi salud mental. Gracias a Jose, Pablo y Tamarindo, juntarme con ustedes fue pésimo para mi salud mental, pero igual gracias por tanto. Especialmente te agradezco Tami por soportarme estos meses.

Gracias a Carolina Salhazar, Made, Conan y Dario, por su apoyo a la distancia y por las visitas esporádicas. Que felicidad que después de todo este tiempo podamos seguir compartiendo.

Gracias a los tesistas y ex-tesistas de transporte. Conocerlos fue como volver a entrar a la universidad, gracias por recibir a la gente nueva tan abiertamente. Leo, sos un sol. Gracias a Felipe, por su amistad por supuesto, y particularmente por contarme de su tesis y de transporte.

Gracias a mis profesores guía Benjamín y Marcela, por todo su apoyo, por las conversaciones y correcciones. Especialmente agradezco a la profesora Marcela por su seguimiento, por su apoyo en las conferencias, por ayudarme a trabajar con el profesor Trépanier. Pero por sobre todo, gracias por su trato en la cotidianidad, siempre atenta y propositiva.

Gracias a toda mi familia por su apoyo. Gracias a mis tías por su cariño y porque desde pequeña han sido mis referentes de mujeres fuertes. Gracias a mi hermanita, por ser como eres, especialmente gracias por tu alegría, echo mucho de menos crecer junto a ti. Gracias a mi papá, por inculcarme las ganas de estudiar en esta universidad, también gracias por enseñarme a confiar en mí y a disfrutar la vida no importando las circunstancias. Gracias a mi mamá, por cada una de las conversaciones a corazón abierto, cada una de las llamadas, todo el apoyo incondicional, todas las visitas; sinceramente no sé si estaría terminando la carrera de no ser por ti, muchas gracias.

Me van quedando unas gracias,
guardadas para el final,
por Santiago y Montreal,
vienen haciendo acrobacias,
y aquí no digo falacias,
siempre estuviste conmigo,
mi compañero, mi amigo,
me acompañó a traducir,
a viajar y a corregir,
pongo a un shandy de testigo.

Entre smart cards y esqueletos,
en las buenas y en las malas,
cual dinosaurio con alas,
te saltaste los libretos,
como el mejor amuleto,
fuiste apoyo emocional,
constante e incondicional,
obrigada, merci, gracias,
digo Ale, sin diplomacias,
arigato exponencial.

Tabla de Contenido

Capítulo 1

Introducción

La movilidad humana ha sido estudiada con múltiples propósitos desde distintas áreas de investigación. Por ejemplo, se ha investigado desde el área de las telecomunicaciones para mejorar servicios de telefonía e internet; desde el transporte para mejorar decisiones operacionales y de planificación; desde el urbanismo para planificación y desarrollo; desde la biología para comprender la evolución de epidemias; desde el marketing para optimizar la oferta de servicios y productos. Todas estas áreas se han visto potenciadas por la disponibilidad de nuevos métodos de recolección de datos, entre los cuales destacan: las tarjetas inteligentes de sistemas de transporte, los dispositivos móviles con geolocalización y los registros de actividad georeferenciados en redes sociales. La disponibilidad de estos datos permite profundizar la investigación de la movilidad humana y al mismo tiempo plantea nuevos desafíos.

En esta tesis se trabaja con registros de transacciones de tarjetas inteligentes de transporte público. De manera simplificada, cada registro de tarjetas inteligentes corresponde a la hora del pago de un viaje y la posición del vehículo en que se realizó la transacción. Estos datos se obtienen mediante la asociación de registros provenientes del sistema de recolección tarifaria automática (AFC, Automatic Fare Collection) y del sistema de localización vehicular automatizada (AVL, Automatic vehicle location). Se han propuesto diversas metodologías para asociar los datos de ambos sistemas, las cuales permiten reconstruir los viajes (o la trayectoria) de cada usuario y generar matrices origen-destino de todos los viajes realizados en el sistema de transporte.

En esta tesis se evalúa si es posible distinguir a los usuarios de transporte público a partir de registros de tarjetas inteligentes. En sistemas de transporte donde las tarjetas inteligentes no almacenan información del usuario, distinguir a los usuarios requiere construir perfiles individuales para representar los patrones de movilidad de cada usuario. En esta tesis se evalúan tres algoritmos de extracción y comparación de perfiles de movilidad, los cuales reciben como entrada dos tablas de transacciones de transporte público, y entregan un indicador de la similitud de la movilidad registrada en las tablas. La evaluación de estos algoritmos permitirá avanzar en la comprensión de los factores que determinan la unicidad de los usuarios, y ahondar en el estudio de la variabilidad de la movilidad humana.

La metodología de trabajo de esta tesis puede ser replicada con cualquier base de datos

de tarjetas inteligentes que haya sido procesada para obtener matrices origen-destino. En particular en esta tesis se trabajó con bases de datos anonimizadas del sistema de transporte público de Santiago, Chile, previamente procesadas con la metodología propuesta por Munizaga y Palma (2012), y del sistema de transporte de Gatineau, Canadá, previamente procesadas con la metodología propuesta por Trépanier et al. (2007).

1.1. Motivación

La tarjeta inteligente bip! es el medio de pago más frecuente de Transantiago, sistema de transporte público de Santiago de Chile. A diferencia de otros sistemas de transporte, Transantiago es un sistema integrado de buses y metro, en el cual todos los buses cuentan con GPS, y el 97 % de las transacciones son realizadas con la tarjeta inteligente (Beltrán et al., 2011). Otra característica del sistema, es que si bien cada tarjeta bip! posee un identificador, Transantiago no almacena información que vincule al dueño con su tarjeta, por lo que a priori no es posible asociar una o más tarjetas a una misma persona.

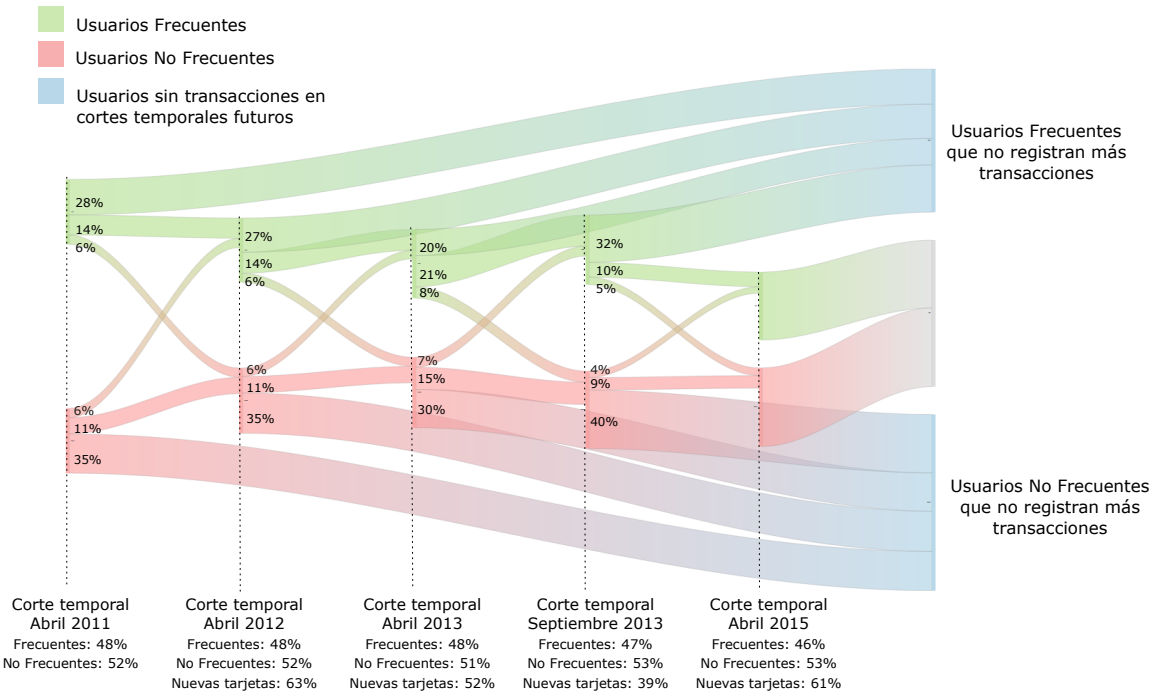


Figura 1.1: Flujo de tarjetas entre distintos cortes temporales separadas en las categorías: usuario frecuente y no frecuente.

La Figura 1.1. muestra el flujo de tarjetas entre distintos cortes temporales por categorías: usuario frecuente, usuario no frecuente y sin transacciones futuras [1]. Es posible observar como en cada corte temporal hay alrededor de un 48 % de usuarios frecuentes, sin embargo las tarjetas que componen este grupo varían significativamente año a año. Anualmente se

[1]Por simplicidad se omite la visualización de tarjetas que reaparecen en cortes temporales que no son el contiguo, ya que representan un porcentaje muy menor de cada categoría. Para ver el porcentaje de tarjetas que fluctúa entre cada categoría dirigirse al Anexo A.

observa un promedio de 50 % de tarjetas nuevas. De este análisis se concluye que el uso de más de una tarjeta, ya sea por renovación o por uso combinado de ellas, es un fenómeno presente y significativo.

Al momento de realizar investigaciones relacionadas a los registros de viajes en transporte público, no poder relacionar un usuario con varias tarjetas de su pertenencia tiene las siguientes consecuencias negativas:

- Se posee información incompleta de la trayectoria de un usuario. Tanto al reemplazar o utilizar más de una tarjeta, la trayectoria del usuario queda segmentada y no es posible observar el comportamiento del usuario en largos periodos de tiempo.
- Se posee información incompleta de la frecuencia de viaje de un usuario. Es posible que usuarios calificados como no frecuentes realicen más viajes registrados en otras tarjetas.
- Limita el análisis dinámico del comportamiento de viajeros. En particular se hace complejo distinguir la fuga de usuarios del extravío de tarjetas. Del mismo modo, se hace difícil distinguir los nuevos usuarios de viejos usuarios con distintas tarjetas.

Ante la alta renovación de tarjetas de Transantiago y las limitaciones que esto provoca, surge la pregunta de investigación: ¿Es posible reconocer, en cortes temporales independientes, tarjetas pertenecientes a un mismo usuario a través de la observación de su movilidad en transporte público?

Responder la pregunta de investigación supone las siguientes dificultades:

- El gran número de usuarios hace más difícil distinguir la movilidad entre uno y otro. En particular, Transantiago registra las transacciones de más de tres millones de usuarios por semana.
- Los registros de transporte público se concentran en posiciones específicas y altamente concurridas (paradas de buses y estaciones de metro).
- Los registros de transporte público son poco frecuentes comparados a otros registros de movilidad. En general se observan entre 0 a 6 registros por día, muy por debajo de los registros de telefonía o servicios GPS.
- No es trivial establecer la similitud requerida para determinar que dos movilidades pertenecen a un mismo usuario, debido a la variabilidad inherente al comportamiento humano.
- Existe poca literatura respecto al reconocimiento de usuarios por medio de su movilidad.

En la literatura se encontraron tres investigaciones que solucionan el problema de reconocer usuarios mediante su movilidad en bases de datos temporalmente independientes. Estos trabajos están expuestos detalladamente en la Sección 2.3. Si bien los tres trabajos presentan metodologías adaptables a cualquier tipo de registros de movilidad, ninguno evalúa el problema utilizando datos de transporte público. Además, las tasas de reconocimiento de usuarios presentadas varían entre un 20 % a un 80 %, donde cada investigación utilizó distintas bases de datos, por lo que no es claro concluir que método propuesto es mejor. Por otra parte, el número de usuarios utilizados varía según cada investigación. El trabajo que reportó mejor tasa de reconocimiento (80 %) fue sobre una base de datos de 100 usuarios. Finalmente, en ninguno de los trabajos se investigó la posibilidad de reducir la tasa de error imponiendo un

umbral de similitud mínima al momento de emparejar usuarios.

Considerando el problema descrito, sus dificultades y las soluciones actuales, la hipótesis que se busca confirmar con esta investigación es: los usuarios de transporte público mantienen patrones de movilidad estables en el tiempo, los cuáles pueden ser utilizados para reconocer a los usuarios en bases de datos de movilidad temporalmente independientes.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general del presente trabajo es evaluar si es posible reconocer usuarios en una base de datos de transacciones de transporte público anonimizada, utilizando los patrones de movilidad de cada usuario previamente extraídos de una base de datos temporalmente independiente. La factibilidad de reconocer usuarios se examina desde dos enfoques: midiendo la estabilidad temporal de los patrones de movilidad de los usuarios y midiendo la capacidad de distinguir a un usuario de otro. Ambos enfoques son evaluados según el rendimiento de tres algoritmos de caracterización y comparación de usuarios, los cuales utilizan distintos mecanismos para obtener y comparar los patrones de movilidad de cada usuario.

1.2.2. Objetivos específicos

- Implementación y adaptación para datos de transporte público del algoritmo de caracterización y comparación de movilidad propuesto por De Mulder et al. (2008).
- Implementación y adaptación para datos de transporte público del algoritmo de caracterización y comparación de movilidad propuesto por Yuan y Raubal (2014).
- Diseño e implementación de algoritmo de caracterización y comparación de movilidad basado en la extracción de características de la movilidad en transporte público.
- Evaluación y comparación de la estabilidad temporal de la caracterización de la movilidad de cada algoritmo implementado.
- Evaluación y comparación de la tasa de reconocimiento de usuarios de cada algoritmo implementado.
- Concluir sobre la factibilidad de reconocer usuarios en bases de datos independientes y determinar factores que afectan la tasa de reconocimiento.

1.3. Metodología

Para desarrollar este trabajo, se aplicó una metodología basada en el proceso Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). A continuación se describe de manera general cada fase de la metodología:

1. **Comprensión del área del problema**

   Se realiza una investigación preliminar sobre el contexto del problema, y sobre el estado del arte de las posibles soluciones.

2. **Comprensión de los datos**

   Se exploran las bases de datos de transacciones de transporte público, a entender el preprocesamiento que se realiza actualmente, y a buscar herramientas compatibles con el tipo y volumen de datos de este proyecto.

   En esta tesis se trabajó con dos bases de datos de transacciones, una con datos de dos semanas del sistema de transporte público de Santiago, Chile, y otra con datos de dos años del sistema de transporte público de Gatineau, Canadá.

3. **Preparación de los datos**

   Se preprocesan los datos para asegurar su disponibilidad en la siguiente fase. El preprocesamiento incluye limpieza de datos, manejo de datos faltantes, selección de registros y variables a utilizar, transformación de tablas y registros. Esta fase se desarrolla en conjunto con la Modelación, y se realizan modificaciones según los resultados de cada iteración.

4. **Modelación**

   Se implementan los tres algoritmos propuestos para caracterización de tarjetas, a diseñar los procesos de medición, a ejecutar los procesos sobre diferentes escenarios y a preparar los datos para la evaluación. Estas tareas se llevan a cabo en las siguientes etapas:

   (a) Etapa 1: Implementación de los algoritmos de caracterización y reconocimiento de usuarios

   (b) Etapa 2: Diseño de la medición de la variabilidad de los perfiles de movilidad en el tiempo

   (c) Etapa 3: Diseño de la medición de la capacidad de distinguir los perfiles de movilidad

   (d) Etapa 4: Ejecución de las mediciones diseñadas en las etapas 2 y 3 sobre diferentes escenarios

   (e) Etapa 5: Postprocesamiento

5. **Evaluación**

   Las etapas 2 y 3 de la Modelación corresponden a procesos independientes con distintos objetivos, por tanto sus resultados son evaluados de manera separada.

   La evaluación de la etapa 2 consiste en calcular métricas de variabilidad sobre la similitud del comportamiento de cada usuario en diferentes periodos de tiempo. Se estudia como varía la distribución de estas métricas al cambiar entre los tres algoritmos implementados, y a su vez como varían las distribuciones de estas métricas variando parámetros en cada algoritmo.

   La evaluación de la etapa 3 consiste en medir las tasas de identificación y de error de cada algoritmo bajo diferentes escenarios. Para medir estas tasas se asume que las tarjeta son portadas por los mismos usuarios en dos cortes temporales independientes. Por lo anterior, la tasa de identificación corresponde al número de tarjetas emparejadas consigo mismas luego de asociar la movilidad de dos cortes temporales independientes.

6. **Despliegue**
   Se ordenan y presentan los resultados de la investigación.

1.4. Limitaciones

Con respecto a la relación entre el objetivo de la tesis y las bases de datos utilizadas existen limitaciones que conllevan dejar de lado, omitir o a asumir como poco probables ciertos comportamientos de usuarios de transporte público. En primer lugar, esta tesis está enfocada en evaluar la capacidad de reconocer usuarios frecuentes de transporte público, ya que los usuarios con bajo número de transacciones no permiten la construcción de un perfil de movilidad característico. Por otra parte, el único atributo disponible para distinguir a un usuario es el identificador de la tarjeta. Por tanto, se asume que dos transacciones asociadas al mismo identificador de tarjeta fueron realizadas por el mismo usuario portador. Es decir, se asume la relación uno a uno entre un usuario y una tarjeta. De esta forma se omite la existencia de usuarios que han extraviado y reemplazado su tarjeta, usuarios que acostumbran utilizar más de una tarjeta, o tarjetas que son utilizadas por más de un usuario.

1.5. Estructura de la tesis

Esta tesis está organizada de la siguiente manera:

El Capítulo 2 corresponde a una revisión bibliográfica del estado del arte en caracterización de la movilidad, comparación de la movilidad y reconocimiento de usuarios a través de registros de movilidad. También se presenta una revisión de la literatura del uso y manejo de datos de tarjetas inteligentes en transporte público.

En el Capítulo 3 se exponen los tres algoritmos de caracterización y comparación de la movilidad que fueron evaluados en esta tesis. El Capítulo 4 presenta la metodología del trabajo realizado, donde se describe la comprensión de los datos, la preparación de los datos, la modelación y los métodos de evaluación de los experimentos realizados.

El Capítulo 5 presenta los resultados de los tres algoritmos en el experimento de medición de la variabilidad de los perfiles de movilidad. El Capítulo 6 presenta los resultados de los tres algoritmos en el experimento de medición de la identificabilidad de los perfiles de movilidad.

Finalmente, en el Capítulo 7 se exponen las conclusiones de este trabajo, se discuten las limitaciones encontradas y se proveen observaciones para el trabajo futuro.

1.6. Glosario

A continuación se presenta un glosario de términos recurrentes en este trabajo, conformado principalmente por palabras que tienen un significado particular en esta tesis, en comparación a su uso común.

- **Parada**: Estación de metro o paradero de bus perteneciente al sistema de transporte urbano. En las bases de datos utilizadas, las paradas se representan con un identificador, nombre y ubicación geográfica.
- **Tarjeta inteligente**: También es referida como *smart card* o tarjeta. Corresponde a la tarjeta utilizada en numerosos sistemas de transporte como mecanismo de pago y monedero electrónico.
- **Transacción**: Registro de la tarjeta inteligente asociado a la acción de pagar la tarifa del servicio de transporte público al momento de iniciar un viaje o un transbordo.
- **Tabla de transacciones**: Corresponde a la tabla que almacena la información asociada a cada transacción de una tarjeta inteligente en una base de datos de un sistema de transporte. Los atributos mínimos que componen una tabla de transacciones son:
  - Identificador de la tarjeta inteligente
  - Marca temporal
  - Parada de subida
  - Modo de transporte
  - Servicio

  Además otros atributos se pueden encontrar presentes:
  - Latitud y longitud de parada de subida
  - Marca temporal de bajada
  - Parada de bajada
  - Latitud y longitud de parada de bajada
  - Tipo de tarjeta
  - Número de viaje
  - Número de etapa de viaje

  Los registros de una Tabla de transacciones se encuentran ordenados por el atributo *Marca temporal*.
- **Perfil de movilidad** : Se refiere a cualquier estructura de datos asociada una tabla de transacciones que conserve información relevante de la movilidad asociada a la tarjeta.
- **Trayectoria**: Conjunto de puntos espaciotemporales que representan la movilidad de un objeto. En esta tesis el objeto es el usuario de transporte público y cada trayectoria es construida a partir de una tabla de transacciones, las posiciones corresponden a paradas y los tiempos a las marcas temporales de las transacciones.

Capítulo 2

Revisión Bibliográfica

La masificación de tecnologías que permiten registrar y compartir la ubicación de distintos objetos ha incentivado la investigación y desarrollo de nuevas aplicaciones en torno a la movilidad. En particular, el estudio de la movilidad humana se ha visto potenciado por la acumulación de registros de telefonía, registros de tarjetas inteligentes de transporte público, y mensajes geolocalizados en redes sociales (Yue et al., 2014).

Para comprender, modelar y reconocer el movimiento de los usuarios de transporte público, es necesario investigar en dos áreas generales: transporte público y movilidad humana. En relación a la movilidad humana, es preciso entender las características del movimiento humano, estudiar cómo extraer patrones de movilidad y cómo comparar y reconocer el movimiento de los usuarios. En relación al transporte público, resulta importante comprender características generales de los usuarios, las propiedades de los datos almacenados por los sistemas de transporte, cómo han sido utilizados hasta ahora y cuáles son sus potencialidades.

Cualquier acercamiento al estudio de la movilidad de un objeto requiere la construcción de un modelo que caracterice la trayectoria observada. Existe una variedad de estructuras de datos utilizadas para representar la movilidad. Cada estructura pone énfasis en ciertos atributos de la movilidad, por lo que la elección de una estructura dependerá de los objetivos de la investigación que se esté llevando a cabo. Las primeras tres secciones de este capítulo presentan distintas aproximaciones a la caracterización de la movilidad humana.

La primera sección revisa estudios de caracterización de la movilidad a través de variables descriptivas, modelo utilizado comúnmente para describir y agrupar perfiles de usuarios. En la segunda sección se presentan diversos modelos de movilidad propuestos con el objetivo de establecer relaciones de similitud entre los usuarios. La tercera sección presenta la escasa investigación existente relacionada al reconocimiento de usuarios a través de su movilidad.

Finalmente, en la última sección de este capítulo se exponen conceptos propios de la movilidad en transporte público. También se presenta investigación relacionada al uso de datos de tarjetas inteligentes de sistemas de transporte, en particular trabajos que han permitido enriquecer la información disponible.

2.1. Caracterización de la movilidad humana a través de variables descriptivas

La caracterización de la movilidad a través de variables descriptivas es utilizada frecuentemente en investigaciones cuyo principal objetivo es describir o agrupar usuarios. En la primera parte de esta sección se presentan variables encontradas en la literatura, con el objetivo de mostrar los descriptores de movilidad comúnmente utilizados. En la segunda parte de esta sección se discuten comportamientos recurrentes de la movilidad humana observados a través de conjuntos de variables descriptivas. Esta discusión abarca tres aspectos de la movilidad humana: regularidad, variabilidad intrapersonal y variabilidad interpersonal.

2.1.1. Variables descriptivas de registros de movilidad

A continuación se presentan las variables descriptivas de registros de movilidad encontradas en la literatura. Estas variables han sido utilizadas sobre registros de telefonía y transporte en distintos contextos espaciales. La clasificación empleada está basada en su mayoría en las categorías presentadas por Ortega-Tong (2013), quien divide los tipos de variables en: temporales, espaciales, sociodemográficas y de modo de transporte. A esta clasificación, se agregó la categoría espaciotemporal y se agregaron también algunas variables a las diferentes categorías. Debido a que es posible encontrar una descripción detallada de las categorías en el trabajo de Ortega-Tong, se presentan aquí de manera concisa los tipos de variables, las variables descriptivas pertenecientes a cada tipo y referencias a investigaciones en que fueron utilizadas. Por el mismo motivo, se omiten las definiciones de variables con nombres auto-explicativos.

Variables Temporales

Las variables temporales buscan patrones de uso de tiempo en los usuarios. Se pueden diferenciar dos parámetros temporales: frecuencia de viaje y tiempo de inicio de viaje. La frecuencia de viaje describe la periodicidad del comportamiento. Mientras que el tiempo de inicio de viajes está asociado al tipo de actividad que se desempeña rutinariamente. Las siguientes variables han sido utilizadas para medir parámetros temporales:

Frecuencia de viaje

- Número de viajes promedio por día. (Ortega-Tong, 2013)
- Número de días en que se realizaron viajes. (Ortega-Tong, 2013; Ma et al., 2013)
- Moda del número de viajes por día. (Valenzuela, 2011)
- Número de días en que ocurre la moda del número de viajes por día. (Valenzuela, 2011)

Tiempo de inicio de viajes

- Hora de inicio promedio del primer viaje (calculada separadamente para semana y fin de semana). (Ortega-Tong, 2013)
- Hora de inicio promedio del último viaje (calculada separadamente para semana y fin de semana). (Ortega-Tong, 2013)
- Número de horas de inicio similares de primer viaje. (Ma et al., 2013)

Variables Espaciales

Las variables espaciales miden parámetros asociados al uso del espacio en la ciudad o zona de estudio. Se pueden dividir en dos dimensiones. Por un lado, la frecuencia de uso de parada describe la periodicidad con que se visitan ciertos lugares. Por otro lado, la distancia de viaje describe la topología de la trayectoria. A continuación se listan las variables espaciales encontradas en la literatura.

Frecuencia de uso de parada

- Número total de paradas visitadas. (Richardson Corvalán, 2014)
- Número de paradas nuevas (en comparación a otro periodo de observación). (Morency et al., 2007)
- Frecuencia de visita a cada parada. (Morency et al., 2007)
- Porcentaje de primeras paradas diferentes: Corresponde al porcentaje de elementos distintos del conjunto de paradas donde se inicia el primer viaje de cada día del período observado. (Ortega-Tong, 2013)
- Porcentaje de últimas paradas diferentes: Equivalente a la variable anterior, pero utilizando el conjunto de paradas donde se inicia el último viaje de cada día del período observado. (Ortega-Tong, 2013)
- Entropía aleatoria: Captura el grado de predictibilidad de la posición del usuario considerando que cada ubicación visitada previamente tiene la misma probabilidad de ser visitada en el futuro. Queda definida por la siguiente fórmula:

  $$E_a = log_2(N),$$

  con E_a la entropía aleatoria y N el número de ubicaciones visitadas por cada usuario. (Song et al., 2010)
- Entropía temporalmente no correlacionada: Captura el grado de predictibilidad de la posición del usuario considerando la frecuencia con que cada ubicación fue visitada previamente. Queda definida por la siguiente fórmula:

  $$E_tnc = -\sum_{j=1}^{N} p(j) log_2(p(j)),$$

  con E_tnc la entropía temporalmente no correlacionada, N el número de ubicaciones visitadas por cada usuario, y $p(j)$ la probabilidad histórica de visitar la ubicación j. (Song et al., 2010)

Distancia de viaje

- Distancia total viajada. (Ortega-Tong, 2013)
- Radio de giro: Corresponde a la desviación estándar de las posiciones de las paradas visitadas por un usuario, respecto al centro de masa del total de paradas. Queda definido por la siguiente fórmula:

$$r_g = \sqrt{\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(r_{\mathrm{i}} - r_{cm})^2}$$

, donde r_g es el radio de giro, n es el número de ubicaciones visitadas, r_{i} es la posición de la ubicación i y r_{cm} es el centro de masa de las posiciones visitadas. (González et al., 2008; De Montjoye et al., 2013; Richardson Corvalán, 2014)

Variables Espaciotemporales

Las variables espaciotemporales son aquellas que miden parámetros espaciales y temporales a la vez. Debido a la complejidad que supone medir la variación simultánea de dos parámetros, este tipo de variable es menos utilizado que los anteriores. A continuación se presentan variables usadas en la literatura:

- Grado de retorno: Corresponde a la probabilidad de volver a un lugar determinado en función del tiempo transcurrido desde la última visita. González et al. (2008); Richardson Corvalán (2014)
- Entropía: Captura el grado de predictibilidad de la posición del usuario considerando la frecuencia con que cada ubicación fue visitada previamente y también, el orden y tiempo gastado en cada ubicación. (Song et al., 2010)
- Regularidad: Probabilidad de encontrar a un usuario en su ubicación más visitada a una hora determinada. (Song et al., 2010)

Variables Sociodemográficas

Este tipo de variables busca asociar características de los usuarios a comportamientos de viaje. Si bien por medio de encuestas es posible acceder a información detallada de las personas, usualmente los servicios de transporte público solo tienen acceso a características limitadas, como las siguientes:

- Tipo de tarifa contratada por el usuario. (Ortega-Tong, 2013)
- Tipo de tarjeta (estudiante, adulto,adulto mayor, funcionario del servicio, etc.). (Ortega-Tong, 2013)

Variables de Actividad

Las variables de actividad describen los intervalos de tiempo entre viajes, con el objetivo de describir indirectamente el propósito de cada viaje. Asociando valores semánticos a los lugares de origen y destino es posible determinar el propósito, esto se puede lograr a través de encuestas o estimaciones que consideren el uso de suelo. Sin embargo, usualmente no es posible contar con la información de propósito y el parámetro medible se limita a la duración de la actividad.

Duración de la actividad

- Duración promedio de la actividad principal (calculada separadamente para semana y fin de semana). (Ortega-Tong, 2013)
- Duración promedio de la actividad más corta (calculada separadamente para semana y fin de semana). (Ortega-Tong, 2013)

Variables de Modo de transporte

Existen factores que pueden determinar la preferencia de un usuario por algún método en particular, por ejemplo: la distancia entre origen y destino, la accesibilidad, costo del viaje, tiempo de viaje. Las siguientes variables describen la preferencia del usuario por utilizar cierto tipo de transporte.

- Porcentaje de días exclusivos de viajes en bus. (Ortega-Tong, 2013)
- Porcentaje de días exclusivos de viajes en metro. (Ortega-Tong, 2013)

2.1.2. Características de la movilidad humana medidas a través de variables descriptivas de la movilidad

Regularidad

Existen numerosos trabajos que dan cuenta de la regularidad del movimiento humano. A continuación, se presentan tres trabajos representativos que describen de manera complementaria diferentes aspectos relevantes de la regularidad.

González et al. (2008) muestran que los modelos comúnmente utilizados para simular la movilidad humana, basados en trayectorias aleatorias, no capturan las características generales del movimiento humano. Para mostrar lo anterior, utilizan variables como el radio de giro y el grado de retorno. El radio de giro describe la distancia recorrida de un usuario en relación al centroide de la trayectoria. La probabilidad de retorno se calcula con respecto a un lugar, y corresponde a la probabilidad de volver al lugar después de t tiempo. En relación al radio de giro, los usuarios muestran un crecimiento logarítmico, mucho más lento que el obtenido de

modelos aleatorios. En relación a la probabilidad de retorno, la distribución temporal de los usuarios muestra máximos locales cada 24 horas en ciertas ubicaciones, a diferencia de la distribución temporal monótona decreciente de la caminata aleatoria. Además, González et al. muestran que la probabilidad de retorno tiene una alta regularidad en las pocas ubicaciones donde se gasta la mayor parte del tiempo.

Hasan et al. (2013) utilizan registros de transacciones de metro del sistema de transporte de Londres, para hacer un *ranking* de las paradas de los usuarios según frecuencia de visita, para encontrar patrones espaciotemporales del comportamiento de las personas. Al analizar las características espaciales, encontraron que el lugar más visitado de las personas estaba distribuido a lo largo de toda la ciudad observada. En cambio, el segundo lugar más visitado estaba notoriamente concentrado en el centro de la ciudad. Por otro lado, al analizar el tiempo de permanencia, se encontró que la distribución temporal de permanencia en el lugar más visitado tiene un *peak* notorio a las 14 horas, y en el segundo lugar más visitado tiene un *peak* a las 9 horas. Lo anterior sugiere que el primer y segundo lugar más visitados corresponden respectivamente al hogar y trabajo de los usuarios, además de mostrar que en estas ubicaciones los usuarios pasan la mayor parte del tiempo, confirmando lo observado por González et al. .

Song et al. (2010) se proponen cuantificar los límites de la predictibilidad de la movilidad humana. Para lograr su objetivo utilizan tres tipos de entropía como variables descriptivas. Concluyen que la movilidad humana puede ser predicha en promedio un 93 %, i.e. solo un 7 % de las ubicaciones visitadas por los usuarios son aparentemente aleatorias. También encuentran que la predictibilidad está fuertemente relacionada a la hora del día y observan, que usuarios que recorren mayores distancias poseen una predictibilidad relativa mayor que usuarios que se mueven en un radio más pequeño.

Variabilidad interpersonal

Si bien es fácil ver que la movilidad humana está dominada por actividades recurrentes que determinan la regularidad de las personas, esta regularidad no es homogénea en la población y está determinada por las características propias del individuo, la ciudad y el servicio en el cual se registre la movilidad. Existe una amplia variedad de estudios que apuntan a encontrar perfiles de usuario que capturen los distintos tipos de regularidades observables en bases de datos de movilidad. A continuación, se presentan como ejemplo dos investigaciones que agrupan usuarios según su comportamiento en el transporte público.

Utilizando técnicas de minería de datos, Agard et al. (2006) logran reconocer diferentes tipos de usuarios de transporte público en la ciudad de Gatineau, Canadá. Utilizan una descripción semanal de los viajes registrados a través de la tarjeta inteligente del transporte público de Gatineau. Esta descripción consiste en un vector de variables binarias que indican si una tarjeta presentó viajes en cada jornada de cada día de la semana. Utilizando *clustering* jerárquico y el algoritmo *K-means*, encuentran y analizan las características de cuatro perfiles de usuarios. Dos de los perfiles muestran, en jornadas diferentes, alta regularidad de viajes. El tercer perfil caracteriza a usuarios poco frecuentes. El cuarto perfil se compone de usuarios sin una regularidad temporal evidente. A través de un análisis de la composición de

dichos *clusters* es posible comprender y describir mejor el comportamiento de los usuarios de transporte público.

Ortega-Tong (2013) realiza una clasificación de los usuarios de transporte público de Londres utilizando cinco tipos de variables descriptivas: temporales, espaciales, de actividad, demográficas y modo de transporte. Cada usuario fue representado a través de un vector de variables descriptivas, y utilizando un método de *clustering* se encontraron ocho tipos de usuarios con distintos tipos de patrones de viaje. Cuatro de los *clusters* corresponden a usuarios regulares, que utilizan 5 días o más el transporte público, con diferencias en la frecuencia, horarios y modo de uso del sistema de transporte. Los otros cuatro *clusters* corresponden a usuarios ocasionales, pudiendo distinguir entre ellos a turistas y viajeros de negocios.

Variabilidad intrapersonal

En la década de 1980, diversos estudios cuestionaron la representatividad de los modelos de movilidad basados en observaciones de un día. En este contexto, Huff y Hanson (1986) discuten la relación entre regularidad y variabilidad del comportamiento de los usuarios, y definen métricas para medir ambos fenómenos. En su investigación observaron una alta regularidad en pocos lugares (hogar, trabajo y compras), sin embargo también una alta variabilidad entre los días. Luego extrajeron patrones que describieran los tipos de día de los usuarios. Finalmente, encontraron que los usuarios tenían más de un patrón de viaje diario y que los patrones diarios de un mismo usuario eran notoriamente diferentes.

Siguiendo con la discusión sobre la necesidad de información multi-día, Jones y Clarke (1988) ponen énfasis en aclarar que toda medida de variabilidad tiene asociado un parámetro sobre el cual se identifica una variación, y la elección de ese parámetro determinará también cuán variable es percibido un comportamiento. En su trabajo definen distintas métricas de variabilidad y luego evidencian cómo cada medida puede llevar a diferentes conclusiones.

Schlich y Axhausen (2003) profundizaron las conclusiones de Jones y Clarke, presentando la diferencia en la variabilidad obtenida de tres medidas propuestas previamente en la literatura. Schlich y Axhausen miden la variabilidad del comportamiento de los usuarios día a día utilizando una encuesta con los registros de viajes de seis semanas. Concluyen, en primer lugar, que la variabilidad observada es menor cuando se utilizan métricas que miden el uso del tiempo en comparación a las métricas basadas en los viajes. En segundo lugar, observan que la variabilidad aumenta según la complejidad del comportamiento capturado por la métrica. En tercer lugar, concluyen que entre mayor el período observado mayor es la variación diaria promedio. Como una de las limitaciones de su trabajo, reconocen que la variabilidad puede ser medida en más niveles que día a día y que los comportamientos habituales pueden tener otros ciclos.

A diferencia de los estudios basados en encuestas presentados anteriormente, Pendyala et al. (2001) utiliza registros GPS de automóviles para medir la variabilidad del comportamiento día a día. Utilizaron un método propuesto en la literatura para medir la variabilidad intra-personal durante un periodo de siete días, utilizando como parámetros un conjunto de diez variables descriptivas. De su análisis destacan tres conclusiones: la variabilidad observa-

da depende del tipo de día (semana o fin de semana), la variabilidad aumenta al ampliar el número de días observado, y la variabilidad observada con datos GPS es mayor a la reportada en estudios anteriores.

Morency et al. (2007) se proponen medir la variabilidad espacial y temporal del uso de transporte público utilizando registros de tarjetas inteligentes. Sus resultados evidencian tanto la variabilidad interpersonal como intrapersonal. Por ejemplo, el promedio de nuevas paradas visitadas a la semana es de 0.7, sin embargo las tarjetas de tipo *Estudiante* visitan nuevas paradas con una tasa semanal de 0.92 a diferencia de la tasa 0.33 observada en los *Adulto interzona*.

2.2. Comparación de usuarios

En esta sección se presenta investigación cuyo objetivo es medir la similitud entre la movilidad de las personas. Para establecer relaciones de similitud a partir de registros de movilidad, por lo general se crean estructuras de datos que permitan representar la movilidad de las personas y luego se establecen métricas de distancia entre las estructuras. La investigación expuesta a continuación fue agrupada en cuatro categorías según el tipo de estructura de datos y tipo de comparación utilizada.

2.2.1. Matching de patrones

Una forma de enfrentar el problema de medir la similitud de los usuarios es crear una representación de la movilidad que permita cuantificar el grado de *matching* o sobreposición de ambas representaciones.

Li et al. (2008) proponen una metodología llamada *Hierarchical Graph-Based similarity measurment* (HGSM), cuyo objetivo es medir la similitud de dos trayectorias representando la movilidad mediante un grafo jerárquico. Para construir el grafo de un usuario, se agrupan los lugares visitados utilizando *clustering* jerárquico, lo que genera un árbol de posiciones visitadas. Luego, se añaden aristas entre las hojas del árbol según la trayectoria del usuario. La estructura jerárquica permite que la trayectoria de un usuario pueda ser representada en distintos niveles de agrupación. Finalmente, la similitud de dos usuarios es calculada según el grado de sobreposición de los grafos de cada usuario, utilizando dos factores: el largo de las secuencias que hacen *match* y el nivel del grafo en que se encuentran las secuencias.

Lee y Chung (2011) también utilizan una representación jerárquica de la movilidad, pero en vez de utilizar niveles de agrupación geoespacial, utilizan niveles de etiquetas semánticas de los lugares. La movilidad de un usuario queda representada a través de un árbol semántico de los lugares visitados, donde cada nodo tiene asociado un peso que corresponde a la razón entre el número de visitas al lugar y las visitas totales. Para calcular la similitud, seleccionan las k principales ubicaciones de un usuario y su árbol asociado, luego, calculan el puntaje relativo al peso de los nodos compartidos a través de un algoritmo de propagación de puntaje. Un aspecto interesante de este algoritmo es que no utiliza la posición espacial de las ubicaciones

y tampoco la secuencia de visitas, es decir comprende la movilidad de los usuarios como un conjunto de actividades realizadas, no como una trayectoria geoespacial.

Yuan y Raubal (2014) proponen una métrica de similitud de trayectorias basada en el algoritmo de distancia de edición. En este trabajo la representación de la movilidad de un usuario es la secuencia temporal de posiciones visitadas. Luego, para medir la similitud entre dos secuencias se calcula el costo de las operaciones requeridas para transformar una secuencia en otra. Las operaciones permitidas sobre los elementos de las secuencias son: insertar, eliminar y reemplazar. Los costos de cada operación se calculan como factores de la modificación temporal y espacial que significan. El costo final se calcula con programación dinámica, minimizando el costo requerido para transformar una secuencia en otra. Un aspecto interesante de la métrica propuesta por Yuan y Raubal es que la definición de costos permite ajustar el peso de los factores espaciales y temporales. De esta forma, la métrica se puede implementar desde una perspectiva espacial, temporal o espaciotemporal.

2.2.2. Similitud de conjunto de patrones

Una metodología muy utilizada para comparar movilidad consiste en extraer por cada usuario un conjunto de patrones de comportamiento, y luego medir la similitud entre los patrones de los usuarios. En esta línea de investigación, cada patrón corresponde a secuencias espaciales o espaciotemporales que se repiten en la trayectoria de un usuario.

Ying et al. (2010) proponen una medida de similitud de trayectorias llamada *Maximal Semantic Trajectory Pattern Similarity* (MSTP-Similarity). En primer lugar, realizan una transformación de registros de telefonía a trayectorias semánticas. Sobre las trayectorias utilizan un algoritmo de minería de patrones de secuencias, llamado *Prefix-Span*, que extrae fragmentos recurrentes de las trayectorias. Cada trayectoria queda asociada a un conjunto de patrones. Para calcular la similitud entre dos patrones se basan en el algoritmo *Longest Common Subsequence*. Finalmente, la función de similitud entre dos usuarios queda definida por la suma ponderada de la similitud entre cada par de patrones de los dos usuarios.

Liu y Schneider (2012) proponen una medida de similitud de trayectorias basada en características geográficas y semánticas. En cuanto a la trayectoria, asumen que es un conjunto de tuplas de cuatro elementos: coordenadas x e y, marca temporal y etiqueta semántica. La componente geográfica de la medida de similitud es una función de la distancia entre los centros de masa y la similitud coseno de las trayectorias. La componente semántica es una adaptación del algoritmo *Longest Common Subsequence*, similar a la propuesta de Ying et al., con la diferencia de que la medida propuesta por Liu y Schneider es simétrica.

Chen et al. Chen et al. (2013) proponen una versión modificada de la medida de similitud de Ying et al., llamada *Maximal Trajectory Pattern Similarity* (MTP-Similarity). En primer lugar, generalizan la medida de trayectorias semánticas a trayectorias espaciotemporales. Luego transforman las trayectorias de registros GPS en trayectorias de *RoIs* (Regiones de Interés, en inglés).[1] Sobre la trayectoria de *RoIs* utilizan el algoritmo de minería de patrones de trayectorias propuesto por Gianotti et al. . A diferencia de Ying et al., no calculan la

[1]El término *RoIs* fue acuñado en la investigación de Gianotti et al. Giannotti et al. (2007), y hace

similitud entre cada par de patrones de ambos usuarios, sino que calculan la similitud entre los patrones más similares de ambos usuarios.

Chen et al. (2014) definen principios que una métrica de similitud de trayectorias debiese seguir y proponen una métrica que los cumple. A diferencia de las métricas *MSTP-Similarity* y *MTP-Similarity*, la medida que proponen no compara la subsecuencia común más larga de patrones, sino que utiliza todos los patrones comunes de dos usuarios. La medida de similitud tiene dos componentes: La primera componente calcula la importancia relativa de cada patrón en común. La segunda componente calcula la diferencia de la frecuencia de cada patrón común en cada trayectoria.

2.2.3. Comparación de matrices espaciotemporales

Thakur et al. (2010) proponen una metodología para definir, modelar y analizar la similitud de la movilidad de usuarios. Definen como perfil de movilidad la *Matriz de asociación*, una matriz en que cada columna representa una posición y cada fila representa periodos de tiempo. Cada elemento de la matriz representa el porcentaje de tiempo gastado en cada ubicación. Luego, la similitud de dos usuarios se calcula como una función ponderada del producto punto de los vectores propios de cada Matriz de asociación, generando un indicador de similitud en el rango [0,1], donde 0 es baja similitud y 1 es alta similitud. Este modelo permite ajustar la resolución de la representación de la movilidad de los usuarios, tanto espacialmente como temporalmente; en particular en este trabajo se utilizó una columna por cada punto de acceso a una red local inalámbrica, y una fila por cada día del periodo registrado.

Lv et al. Lv et al. (2013) plantean un método para capturar y comparar la movilidad de los usuarios. En su trabajo representan la movilidad de los usuarios a través de *actividades de rutina*. Una actividad de rutina corresponde a una matriz en la que las filas representan lugares visitados, las columnas horas del día, y cada celda corresponde a la probabilidad de encontrar al usuario en un lugar a determinada hora. Un usuario puede ser caracterizado por más de una matriz, esto ocurre cuando el usuario posee comportamientos regulares en diferentes periodos, por ejemplo entre semana y fin de semana. Luego, diseñan una métrica de similitud entre usuarios en la que se comparan las actividades de rutina de cada uno, midiendo la similitud entre las matrices desde dos perspectivas: la similitud de la distribución temporal de visita de los pares de lugares que más se parecen, y la similitud de la probabilidad de encontrarse en un lugar en un determinado periodo de tiempo, i.e comparando las filas (distribución temporal) y las columnas (distribución espacial).

---

referencia a regiones o zonas en las que un usuario desempeña actividades; calcular estas regiones es de particular importancia cuando los registros de movilidad con los que se trabaja poseen alta granularidad. La forma de extraer las *RoIs* varía en la literatura: se utilizan diferentes algoritmos de *clustering*, dependiendo principalmente del tipo de registros que se utilicen.

2.2.4. Comparación cuantitativa

Wang et al. [42] acuñaron el concepto *mobile homophily*, usado después por Bapierre et al. [4], para referirse a la similitud de usuarios en relación a su movilidad. Para medir la similitud de los usuarios definen un conjunto de medidas que cuantifican el grado de sobreposición entre dos trayectorias. Las medidas definidas consideran aspectos espaciales y espaciotemporales como: la distancia, la probabilidad de visitar las mismas ubicaciones o la probabilidad de visitar las mismas ubicaciones al mismo tiempo. También definen medidas que consideran factores como la popularidad de los lugares o la influencia de la hora en la probabilidad de encontrarse en una misma ubicación. Estas medidas fueron comparadas con datos de redes sociales, mostrando que existe una fuerte dependencia entre la similitud de la movilidad y la cercanía social de las personas.

2.3. Reconocimiento de usuarios

A diferencia de los tópicos de las secciones anteriores, el problema específico de reconocer usuarios mediante la observación de su movilidad no ha sido investigado ampliamente. Si bien existe bastante investigación relacionada a de-anonimizar bases de datos de movilidad, problema que bien podría ser llamado reconocimiento de usuarios, usualmente las soluciones propuestas involucran utilizar información pública adicional y directamente relacionada con el periodo anonimizado. El problema que se aborda en esta tesis es el de emparejar la movilidad de usuarios en dos periodos de tiempo diferentes, sin contar con más información que la registrada mediante la tarjeta inteligente. Por lo anterior, las investigaciones presentadas a continuación también poseen la limitación de emparejar información de usuarios proveniente de *datasets* independientes.

2.3.1. Solución propuesta por De Mulder et al.

De Mulder et al. (2008) evalúan dos soluciones al problema de identificar usuarios en una base de datos de telefonía móvil anonimizada. En su trabajo, el problema consiste en emparejar los registros de movilidad de usuarios anónimos con perfiles de movilidad previamente construidos. Para lograr esto, separan la base de datos en dos cortes temporales. El primer periodo lo utilizan para construir *Location Profiles* de cada usuario. El segundo periodo lo utilizan para observar la movilidad de cada usuario, e intentar hacer un *match* con algún perfil del primer periodo. En relación al segundo periodo, implementan dos métodos para emparejar datos de movilidad con *Location Profiles*, de los cuales el más exitoso presenta un 80 % de identificaciones correctas. De esta forma demuestran que para anonimizar bases de datos de movilidad no basta con remover los identificadores.

El perfil de movilidad o *Location Profile* diseñado por De Mulder et al., consiste principalmente en una *Matriz de Probabilidad de Transición (MPT)*. La matriz MPT es un matriz de $n \times n$ en la que cada elemento almacena la probabilidad del usuario de trasladarse de una ubicación a otra. Se crea a partir del conjunto de n ubicaciones que corresponden a las celdas

telefónicas visitadas por el usuario.

El proceso de identificación con mayor tasa de identificación implementado por De Mulder et al. consiste en calcular un indicador de la afinidad entre una secuencia de celdas telefónicas y una matriz MPT. Se evaluó este método con un *dataset* de registros de telefonía de 100 usuarios, si bien anónimos, distinguibles por un identificador. Utilizaron un mes de registros para extraer la matriz MPT de cada usuario y un mes para obtener la trayectoria de celdas o ubicaciones. Luego, para cada usuario presente en los registros del segundo mes, seleccionaron la matriz MPT que maximizara el indicador de afinidad. Los resultados de este experimento señalan que en promedio un 80 % de los usuarios puede ser identificado correctamente. Otro resultado interesante es que teniendo los perfiles de movilidad previamente construidos, bastaba una hora de registros de movilidad para reconocer en promedio al 45 % de los usuarios. De Mulder et al. concluyen que los datos de movilidad debiesen ser considerados como información privada y almacenada con las protecciones correspondientes.

2.3.2. Solución propuesta por Naini et al.

Naini et al. (2016) implementan una solución generalizada al problema de emparejar los registros de un usuario en dos bases de datos independientes. Su solución se basa en representar el comportamiento de un usuario a través de un histograma. En el caso de la movilidad a través de la ciudad, el histograma se construye como el porcentaje relativo de visitas a una ubicación en relación al total de lugares visitados. En el escenario planteado por Naini et al. se tienen dos *dataset*: un *dataset* anonimizado de histogramas que almacenan el comportamiento de un grupo de usuarios, y otro *dataset* de histogramas con los registros de los mismos usuarios, pero con información independiente (ya sea distinto periodo o distinta fuente de datos). Luego, el problema consiste en emparejar histogramas correspondientes al mismo usuario.

Para emparejar histogramas Naini et al. diseñan una métrica de distancia entre histogramas. Luego diseñan una solución a modo de grafo bipartito completo, donde los nodos corresponden a los histogramas de ambos *dataset*, y los arcos tienen asociado un peso que corresponde a la distancia entre los histogramas. La solución que proponen fue implementada mediante un algoritmo de *Mínimo peso máximo emparejamiento*, donde el óptimo se encuentra entre que cada histograma se empareje con el más similar y la mayor cantidad de histogramas se emparejen con el histograma más similar disponible. Utilizaron este método sobre dos *datasets* de movilidad: un *dataset* de registros de llamadas telefónicas de dos semanas de 50.000 usuarios, otro *dataset* de registros GPS de cinco años de 182 usuarios. Ambos *datasets* fueron divididos en dos periodos. Luego de remover aquellos usuarios presentes solo en un periodo, se ejecutó el algoritmo anteriormente descrito, obteniendo una tasa de identificación de 21.1 % para el *dataset* de telefonía y un 58.4 % para el de registros GPS.

La riqueza del trabajo de Naini et al. radica en el análisis de los factores que afectan al porcentaje de identificación. En relación a los experimentos de movilidad urbana, evaluaron el efecto de variar: el número de usuarios, la resolución espacial, el periodo de tiempo registrado, la composición del *dataset* (cuántos usuarios están presentes en ambos *datasets*). En la Figura 2.1 A, es posible observar cómo disminuye significativamente el porcentaje de

usuarios identificados correctamente al aumentar el número de usuarios. Con 1.000 usuarios el porcentaje de identificados es de 78 %, mientras que con 47.000 es 21,1 %. En la Figura 2.1 B, es posible observar cómo aumenta el porcentaje de identificación al variar el periodo de registro de un día a una semana. En cuanto a la resolución espacial, el efecto depende del tipo de datos; por ejemplo, con datos GPS un disminución moderada de la resolución espacial puede aumentar la precisión. Finalmente, en relación a la composición del *dataset* los resultados señalan que entre mayor sea la cantidad de usuarios compartidos en ambos *datasets*, entonces mayor es la tasa de identificación.

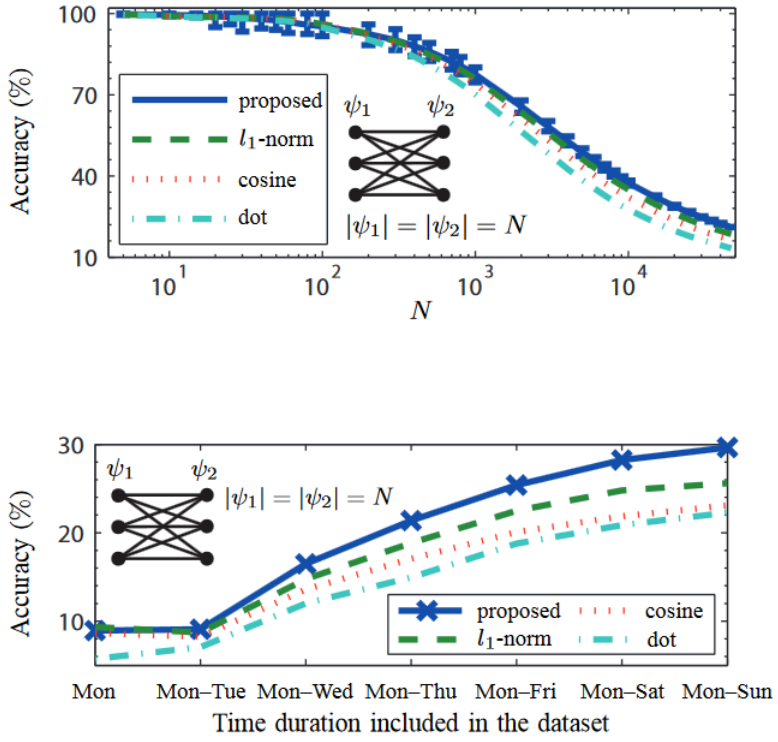


Figura 2.1: Precisión promedio del *matching* de dos conjuntos de histogramas, ψ_1 y ψ_2, compuestos por la misma cantidad de usurarios ($|\psi_1| = |\psi_2| = N$). La leyenda indica el tipo de distancia utilizada para comparar los histogramas. El gráfico A muestra como varía la precisión al variar el número de usuarios. El gráfico B muestra como varía la precisión al variar la duración del periodo de los registros. Fuente: Naini et al. (2016)

Otro aspecto interesante de su investigación es que evalúan dos enfoques: uno secuencial, donde cada histograma se empareja con el histograma más parecido; y otro paralelo, donde se busca el óptimo para todos los usuarios, minimizando la distancia entre los histogramas y maximizando el número de óptimas identificaciones. Sus resultados señalan que al resolver el problema en paralelo se obtienen mejores resultados, y se evita que un usuario sea emparejado con más de un calce.

2.3.3. Solución propuesta por Gambs et al.

Gambs et al. (2014) evalúan la efectividad de ataques de de-anonimización de una base de datos geolocalizados utilizando perfiles construidos con registros de movilidad observada independientemente. El perfil de movilidad que construyen es muy similar al utilizado por De Mulder et al., basado en una MPT. La principal diferencia es que Gambs et al. calculan los Puntos de Interés (PoIs) de cada usuario, para utilizarlos como las ubicaciones que componen la matriz MPT. Lo anterior se justifica porque Gambs et al. utiliza bases de datos de registros GPS de alta frecuencia, por tanto dispersos espacialmente. Los PoIs corresponden puntos de acumulación de registros GPS calculados mediante un método de *clustering* llamado *Density-Joinable Cluster*. Finalmente, el problema se reduce a lograr identificar a un usuario en una base de datos de matrices MPT anónimas, dado una matriz MPT construida con movilidad previamente observada de aquel usuario.

Para poder emparejar matrices MPT, Gambs et al. definen cuatro métricas de distancia entre matrices MPT. Con estas métricas construyen cinco algoritmos de-anonimizadores, los cuales utilizan una o varias de las métricas definidas. La eficacia de estos algoritmos es evaluada utilizando cinco *datasets* de registros GPS de movilidad, el más grande, con 185 usuarios. Las tasas de identificación obtenidas varían entre un 5 y un 45 %, dependiendo del de-anonimizador y del *dataset* utilizado.

La Figura 2.2 muestra los resultados obtenidos por los distintos de-anonimizadores sobre el *dataset* Geolife, *dataset* que también fue utilizado por Naini et al.. En esta Figura, es posible observar que existe una gran diferencia en la eficacia de los diferentes de-anonimizadores, con un máximo cercano al 45 %. También es posible observar que hay una mejora considerable de la tasa de identificación al disminuir la frecuencia temporal de los registros GPS de 5 a 10 segundos, y que la tasa de identificación no varía significativamente al disminuir el muestreo a dos minutos.

Un aspecto interesante de la investigación de Gambs et al., es su similitud con el método de menor éxito de De Mulder et al., tanto en procedimiento como en resultados. Esto sugiere que al utilizar matrices MPT para reconocimiento de la movilidad de usuarios, es conveniente utilizar el método de mayor éxito de De Mulder et al..

La Tabla 2.1 muestra un resumen del rendimiento de los algoritmos exhibidos en esta sección. Se puede ver que los rendimientos de los distintos algoritmos varían entre un 20 % y un 80 %. Este amplio rango se debe, no solo a los distintos algoritmos, sino también a las distintas características de los *datasets* utilizados. Por lo anterior, no es posible obtener conclusiones cuantitativas. Es posible apreciar que para *datasets* de registros GPS, los resultados de Naini et al. son mejores que los resultados de Gambs et al. . Por otro lado, los resultados de *datasets* de registros de antenas de telefonía muestran un rendimiento muy dependiente del número de usuarios comparados. Es posible concluir que al comparar 100 usuarios el rendimiento del algoritmo de Naini et al. es superior al del algoritmo propuesto por De Mulder et al., sin embargo no existe evidencia de que esta relación se mantenga al ejecutar el algoritmo de De Mulder et al. con mayor número de usuarios.

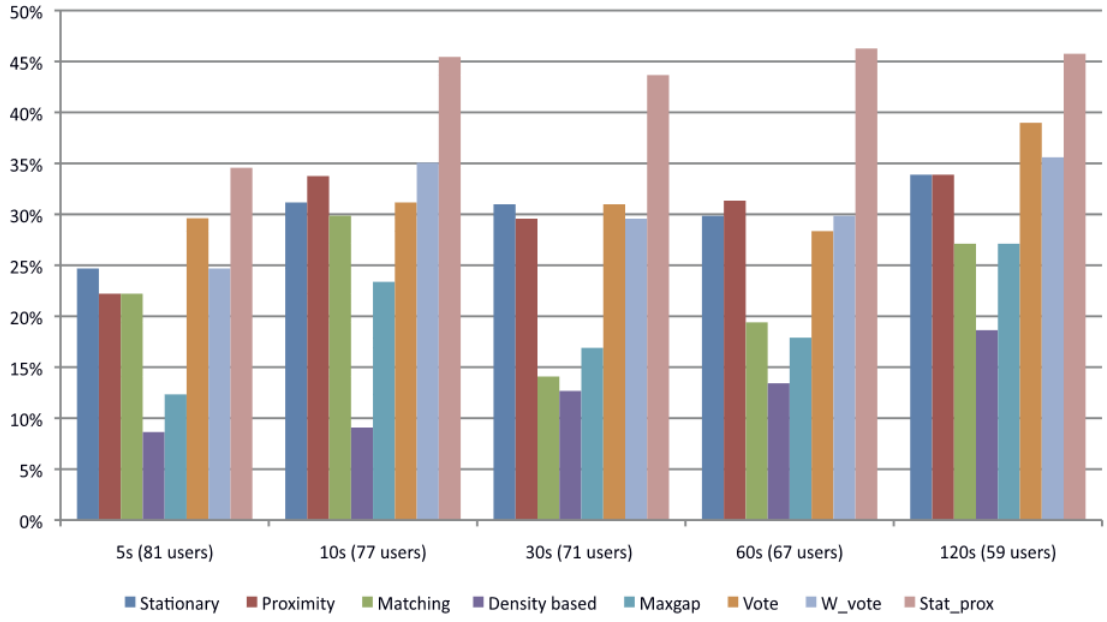


Figura 2.2: Tasa de identificación de los diferentes algoritmos de-anonimizadores propuestos por Gambs et al. (2014), sobre el *dataset* Geolife, variando la frecuencia temporal de los registros GPS.

Tabla 2.1: Tabla comparativa de rendimiento de los métodos propuestos en la literatura para reconocer usuarios a través de la comparación de su movilidad.

Propuesto por	Método basado en	Tipo de datos	Número de ubicaciones	Periodo de observación	Frecuencia de registros	Número de usuarios	Tasa de identificación
De Mulder et al.	Matriz MPT	Registros de antenas de telefonía	No se informa	2 meses	1 por hora	100	80 %
Naini et al.	Histograma	Registros de antenas de telefonía	1211	2 semanas	7,2 por día	100	90 %
Naini et al.	Histograma	Registros de antenas de telefonía	1211	2 semanas	7,2 por día	46986	21.1 %
Naini et al.	Histograma	GPS	1024	5 años, variable por cada usuario	30 por hora	154	58.4 %
Gambs et al.	Matriz MPT	GPS	1024	146 días	30 por hora	175	45 %

2.4. Sistemas de transporte público y tarjetas inteligentes

Desde que las tarjetas inteligentes comenzaron a ser utilizadas como medio de pago en transporte público, los proveedores de servicio han podido almacenar la información de las transacciones y utilizarlas como registros georeferenciados del uso del sistema. La accesibilidad a grandes volúmenes de datos ha potenciado investigaciones relacionadas al control del funcionamiento del servicio, comprensión del comportamiento de los usuarios, planificación del sistema y manejo de anomalías.

Robinson et al. (2014) proveen un contexto bastante completo del funcionamiento de las tarjetas inteligentes, sus beneficios y problemas comunes. Su descripción de un sistema de transporte general utiliza las siguientes entidades:

Etapa de viaje: Describe el movimiento de un pasajero en un solo vehículo, típicamente un bus o tren. El viaje comienza en la parada donde el pasajero toma el vehículo y termina en la parada donde el pasajero abandona el vehículo.

Parada: Describe una ubicación donde el pasajero puede abordar o descender de un vehículo del transporte público.

Viaje: Describe el movimiento desde una ubicación de origen hasta una ubicación de destino. El origen se asume que es el primer punto de parada en el que el pasajero ingresa a la red de transporte público. Así mismo, la ubicación de destino se asume que es la última parada desde la cual el pasajero sale de la red de transporte público.

Recorrido: Describe el movimiento de un vehículo de transporte público a través de una secuencia definida de puntos de parada.

Transbordo: Describe el movimiento de un pasajero entre un vehículo de transporte público y otro vehículo de transporte público. Esto puede involucrar que el pasajero camine entre diferentes puntos de parada, o bien, realice un cambio de vehículos en una misma parada.

De acuerdo a estas definiciones y desde la perspectiva de los registros de tarjetas inteligentes, un sistema de transporte público esta compuesto por vehículos con recorridos definidos a través de un número determinado de paradas. Los usuarios del sistema realizan viajes, los cuales pueden tener una o más etapas de viaje. Cada etapa corresponde a un viaje en un vehículo del sistema. Al abordar un vehículo en una parada, los usuarios realizan una transacción, la cual, implícitamente o explícitamente, está asociada a la parada utilizada. En algunos sistemas de transporte, el usuario también realiza una transacción al momento de descender del vehículo. Al descender del vehículo, el usuario puede realizar un transbordo, e iniciar una nueva etapa, o finalizar su viaje.

La información recolectada a través de las tarjetas inteligentes proviene de dos sistemas de registro: *Automatic Fare Collection System* (AFC) y *Automatic Vehicle Location System* (AVL). El sistema AFC registra las transacciones hechas por los usuarios en los vehículos que conforman el transporte público. El sistema AVL se compone de registros georeferenciados de la posición de los vehículos del sistema. Ya sea al momento de la transacción, o posterior a la recolección de datos, la información de los sistemas AFC y AVL se agrupa para que cada transacción registrada tenga asociada un tiempo, un servicio, un usuario, una tarifa y una posición.

Es importante notar que existen diferentes tipos de vehículos, diferentes tipos de usuarios, diferentes tipos de puntos de parada y diferentes tarifas. Finalmente, cada sistema de transporte público posee una configuración determinada, y los registros de tarjetas inteligentes responden a esta configuración.

Bagchi y White (2004) discuten sobre las potencialidades de las tarjetas inteligentes en transporte público. Destacan los beneficios de las tarjetas por sobre los medios de pago no digitales, y establecen un marco conceptual para los términos usuario, viaje y fuga de usuarios; conceptos comúnmente utilizados y que con los registros de las tarjetas inteligentes adquieren una nueva dimensión y pueden ser analizados con mayor profundidad. Por ejemplo, las tarjetas inteligentes permiten relacionar las transacciones pertenecientes a un usuario, luego, determinar la frecuencia con que los usuarios utilizan el transporte público. También es posible distinguir las etapas de viaje de los usuarios, y luego estimar el tiempo de viaje total.

Una de las potencialidades que se ha explotado de las tarjetas inteligentes es enriquecer la información registrada por los sistemas AFC y AVL. Como se mencionó anteriormente, no todos los sistemas de transporte público requieren registrar la bajada de un vehículo. Por lo anterior, el problema de estimar la bajada de los viajes de los usuarios cobra especial relevancia al intentar observar los viajes de los usuarios.

Trépanier et al. (2007) propone y evalúa un método de estimación de bajada, cuyo objetivo es identificar la parada de destino para cada viaje de un usuario. Su modelo considera los objetos: usuario, ruta, viaje y paradas, y además define el concepto *Vanishing Route*, que corresponde a la secuencia de paradas perteneciente a la ruta de un vehículo del transporte público tal que la primera parada corresponde al paradero donde el usuario comienza el viaje. Teniendo las rutas que utilizó un usuario a lo largo del día, es posible extraer sus *Vanishing Routes*, y luego, estimar la bajada seleccionando la parada de una *Vanishing Route* que esté más cerca de la primera parada de la siguiente *Vanishing Route*.

La Figura 2.3 resume el modelo propuesto por Trépanier et al. . Este modelo utiliza la restricción de que la distancia entre la parada de bajada estimada y la siguiente parada de abordaje sea menor que la distancia caminable. Además, como este método utiliza la parada de subida del siguiente viaje del día, propone un método para estimar la bajada del último viaje del día, en el cuál se consideran como referencias la primera parada del día, la primera parada del día siguiente, o la parada del viaje del día que utilice la misma ruta.

Munizaga y Palma (2012) plantean una metodología para construir matrices origen destino a partir de los viajes realizados por los usuarios de transporte público. Las matrices origen destino muestran el número de viajes realizados desde cada origen y cada destino del sistema (en este caso paraderos y estaciones de metro). A diferencia de otros métodos, la metodología de Munizaga y Palma permite ver las matrices con diferentes grados de agrupación tanto espacialmente como temporalmente.

Una parte importante del trabajo de Munizaga y Palma corresponde a la reconstrucción de los viajes de los pasajeros. En primer lugar implementan un algoritmo de estimación de bajada para viajes abordados en bus, metro y zonas paga. En el caso de la estimación de bajada en buses, se minimiza una función de costo temporal en la que se considera tanto el tiempo de bajada estimado como el tiempo asociado a caminar entre la bajada estimada y el siguiente paradero de origen. Para los viajes en Metro se utiliza una función similar a la del bus, con la diferencia de que se hace necesario estimar la hora de bajada y la ruta utilizada. Finalmente para Zonas Pagas antes de utilizar la estimación de bajada de buses, es necesario estimar cual fue el servicio utilizado por el usuario.

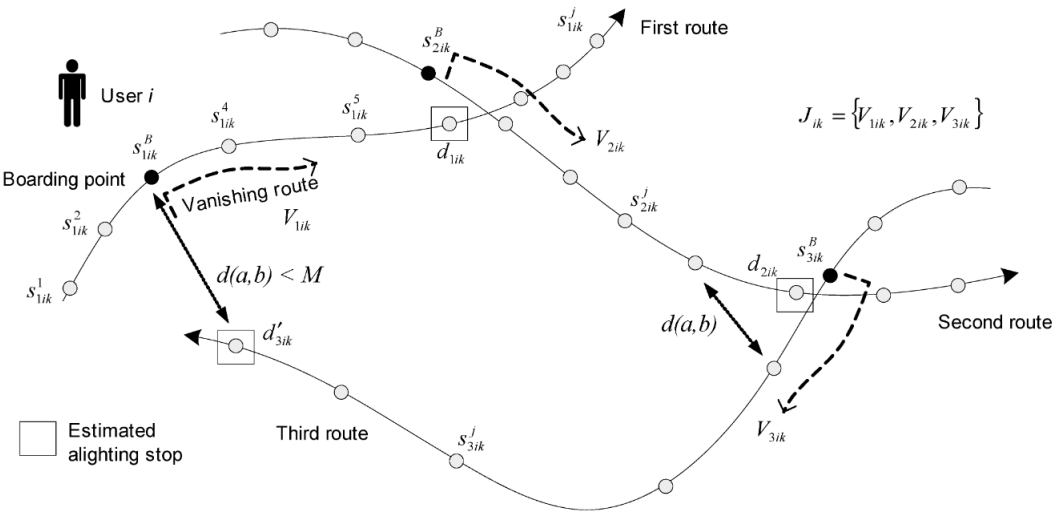


Figura 2.3: Modelo de estimación de bajadas de viajes en Transporte Público. Fuente: Trépanier et al. (2007)

Una vez estimadas las bajadas de cada etapa, Munizaga y Palma llevan a cabo un proceso de asociación de etapas para reconstruir los viajes de los usuarios. Un viaje corresponde a un desplazamiento entre un origen y un destino en el que se desempeñará una actividad. Una etapa corresponde al desplazamiento a través de un vehículo del transporte público. Debido a que un viaje puede tener una o más etapas, para construir la matriz origen destino de los viajes, es necesario asociar las etapas que corresponden a un mismo viaje. Para lograr lo anterior, se utiliza la premisa de que si el tiempo entre dos transacciones supera los 30 minutos, es porque se desempeñó una actividad, luego, las transacciones que marcan el comienzo e inicio de dicha actividad, corresponden al destino y origen de viajes diferentes. Una vez divididos los viajes de los usuarios, es posible construir una matriz origen destino, a la cual se le aplican factores de corrección, debido a viajes que permanecen sin origen o destino estimado.

Una vez reconstruidos los viajes de los usuarios, surgen otras formas de enriquecer los datos. Por ejemplo, se han propuesto métodos para estimar el propósito de los viajes de los usuarios (Devillaine et al., 2012; Kusakabe y Asakura, 2014). Otros esfuerzos se han dedicado a clasificar tipos de usuarios (Agard et al., 2006),(Ortega-Tong, 2013) o a cruzar los datos de tarjetas inteligentes con otras fuentes de información (Spurr et al., 2015). Pelletier et al. (2011) realizan una exhaustiva revisión de la literatura sobre el uso de tarjetas inteligentes.

Capítulo 3

Algoritmos de caracterización y comparación de movilidad

En este capítulo se describen los tres métodos implementados en esta tesis para caracterizar y comparar la movilidad de usuarios de transporte público. Los métodos han sido adaptados e implementados para comparar dos registros de movilidad, teniendo como entrada dos tablas de transacciones. Cada método tiene diferentes procedimientos para representar y comparar la movilidad, por lo que la salida de cada algoritmo varía de significado y de dominio. Para facilitar la comprensión, se ha separado la descripción de los métodos en dos etapas: Construcción del perfil de movilidad y Comparación de perfiles de movilidad. Es decir, cómo cada algoritmo representa la movilidad, y cómo compara aquellas representaciones. Además se incluye un análisis general de cada método. Para una mejor comprensión de los algoritmos, en el Anexo B se adjuntan ejemplos de cada algoritmo con datos reales.

3.1. Algoritmo basado en la matriz de probabilidad de transición

El algoritmo basado en la matriz de probabilidad de transición (TPM, por su sigla en inglés)[1], es una adaptación del método propuesto por De Mulder et al., cuya investigación tenía por objetivo reconocer usuarios en dos cortes temporales a partir del registro de celdas telefónicas visitadas. De Mulder et al. diseñaron un método para crear perfiles de movilidad o *Location Profiles*; y dos métodos para comparar perfiles de movilidad: *Identification process based on Markovian model* e *Identification process based on sequences of cell-ID's*. En esta sección se explica el método que obtuvo mejor rendimiento[2]: *Identification process based on sequence of cell-ID's*, utilizando una abstracción adaptada para registros de transacciones de transporte público.

[1]En adelante, para referirse a este algoritmo se utilizará "el algoritmo TPM", para referirse a la matriz se utilizará "la matriz TPM"

[2]ver Capítulo 2, Sección 2.3.1

3.1.1. Construcción del perfil de movilidad

Sea la entrada del algoritmo TPM: T_a y T_b, dos tablas de transacciones de n y m registros con los atributos *Identificador de tarjeta,Marca temporal,Parada de subida*; donde T_a y T_b corresponden a las transacciones de las tarjetas t_a y t_b respectivamente.

El algoritmo TPM utiliza dos estructuras como perfil de movilidad: la secuencia de posiciones y la matriz TPM. La secuencia de posiciones es calculada para las dos tablas de transacciones, en cambio la matriz TPM es calculada solo para la tabla T_a.

Secuencia de posiciones Una secuencia de posiciones corresponde a una secuencia temporalmente ordenada de ubicaciones visitadas. En el caso de un sistema de transporte público, cada ubicación corresponde a una parada.

La secuencia de posiciones asociada a la tabla T_a, se define como $S_a : [s_a^1, s_a^2, ..., s_a^{n-1}]$, $s_a^{\mathrm{i}} \neq s_a^{\mathrm{i}+1} \forall \mathrm{i} \in [1, n-1]$, y corresponde a los n valores del atributo *Parada de subida* ordenados según el atributo *Marca temporal*. En el caso de existir valores consecutivos equivalentes se deja solo un elemento.

De la misma forma se define $S_b : [s_b^1, s_b^2, ..., s_b^m]$, asociada a la tabla T_b.

En el caso de que el atributo *Parada de bajada* esté disponible, entonces la secuencia se construye iterando intercaladamente sobre los elementos de los atributos *Parada de subida* y *Parada de bajada* ordenados según el atributo *Marca temporal*. Es importante notar que cualquier valor asociado al atributo *Parada de subida* o *Parada de bajada* corresponden a una parada del sistema de transporte que se esté analizando, por tanto con un dominio definido por el conjunto de paradas de aquel sistema.

Matriz de probabilidad de transición Una matriz TPM corresponde a una matriz donde cada fila y columna representa una ubicación, y cada elemento corresponde a la probabilidad de transitar de una ubicación a otra.

Sea $S_a : [s_a^1, s_a^2, ..., s_a^n]$ la secuencia de posiciones de T_a, con $n \geq 2$. Sea $L : [l_1, l_2 ... l_u]$ el conjunto mínimo de las u ubicaciones que componen S_a. Llámese $l_{\mathrm{i}} \rightarrow l_j$ una transición de una ubicación l_{i} a una ubicación l_j. En total, pueden definirse $n-1$ transiciones entre posiciones de S_a.

La matriz TPM P se define como una matriz de $u \times u$, tal que:

$$P[\mathrm{i}, j] = Pr(l_{\mathrm{i}} \rightarrow l_j), \mathrm{i}, j \in [0, u],$$
$$\text{con } Pr(l_{\mathrm{i}} \rightarrow l_j) = \frac{Contar(l_{\mathrm{i}} \rightarrow l_j)}{\sum_{r=1}^{u} Contar(l_{\mathrm{i}} \rightarrow l_r)},$$

donde el método $Contar(t)$ cuenta la cantidad de veces que ocurre la transición t en S_a.

De lo anterior se desprende que la fila i de la matriz P corresponde a las probabilidades de desplazarse de la ubicación l_{i} a cada ubicación de L, y la suma de estas probabilidades es 1.

3.1.2. Comparación de perfiles de movilidad

La comparación de dos perfiles de movilidad del algoritmo TPM consiste en calcular un indicador de la probabilidad con la que una secuencia de posiciones haya sido generada por un usuario con una matriz TPM predeterminada.

Sea P una matriz TPM de T_a y S_b una secuencia de posiciones de T_b, luego el indicador de la similitud ($sim_{a,b}$) entre P y S, se calcula según la siguiente fórmula:

$$sim_{a,b} = \sum_{j=1}^{n-1} \log_{10} Pr(S_b^j \rightarrow S_b^{j+1}).$$

Cuando las transiciones de la secuencia de posiciones S_b existen en la secuencia de posiciones S_a que dio origen a P, entonces la probabilidad de dicha transición existe en la matriz P. Sin embargo, si la transición no fue realizada entonces la probabilidad es 0 y luego, el logaritmo de la probabilidad es $-\infty$, lo cual haría que el indicador de similitud fuese $-\infty$ independientemente del resto de las transiciones. Por lo anterior se definen dos parámetros para este algoritmo: p_0 que corresponde a la probabilidad asociada a viajes entre ubicaciones que existen en la matriz P pero que tienen probabilidad 0, y p_{nan} que corresponde a la probabilidad asociada a viajes entre ubicaciones que no fueron observadas en el corte temporal asociado a la matriz P. Los valores de p_0 y p_{nan} deben pertenecer al intervalo $[0, 1]$, y deben ser ajustados considerando que equivalen a la importancia que se le asignará a la existencia de nuevos viajes entre los periodos observados, i.e. entre menores sean los valores p_0 y p_{nan}, menor será la similitud total entre las tablas de transacciones que presenten viajes nuevos.

La Tabla 4.1 presenta el pseudócodigo del algoritmo TPM. La *salida* de este algoritmo corresponde al indicador de similitud entre una matriz TPM y una secuencia de posiciones. Este indicador pertenece al rango $(-\infty, 0]$, donde un valor 0 corresponde a máxima similitud. La cota inferior de este indicador queda definida por los valores asignados a p_{nan} y p_0 y por el número de transiciones de la secuencia de posiciones no definidas en la matriz TPM.

La *salida* de este algoritmo corresponde a la distancia de edición entre dos trayectorias. Esta distancia pertenece al rango $[0, \infty]$, donde distancia 0 indica máxima similitud. Por otro lado, no existe una cota superior definida ya que la distancia dependerá del costo de las operaciones de transformación que a su vez no tienen un límite definido y dependen de la distancia geotemporal entre las trayectorias.

Tabla 3.1: Algoritmo TPM.

Algoritmo 1:	TPM
Input:	tablas de transacciones T_a, T_b
Output:	Similitud $sim_{a,b}$, con $sim_{a,b} \in (\infty, 0]$

```
1: funcion TPM(T_a, T_b) :
2:     S_a ← obtenerSecuenciaPosiciones(T_a)
3:     S_b ← obtenerSecuenciaPosiciones(T_b)
4:     tpm_a ← obtenerTPM(S_a)
5:     sim_{a,b} = 0
6:     for j in [0...largo(S_b) − 1] :
7:         pr_{j,j+1} ← obtenerProbabilidad Transicion(tpm_a, S_b^j, S_b^{j+1})
8:         sim_{a,b} ←+= log_10(pr_{j,j+1})
9:     retornar sim_{a,b}
```

3.1.3. Análisis del algoritmo TPM

Complejidad

A continuación se presenta un análisis asintótico del peor caso del número de comparaciones del algoritmo TPM. Se describe el costo de los diferentes pasos de este algoritmo, dejando con costo 1 operaciones como asignaciones y cómputos aritméticos. Los parámetros de la función de costo son el largo de las tablas T_a y T_b. Se utiliza como referencia las líneas del pseudocódigo de la Sección 3.1.

1. Línea 2 y 3: Obtener la secuencia de posiciones requiere recorrer una vez las tablas T_a y T_b. Costo: n y m respectivamente.
2. Línea 4: Obtener la matriz TPM considera diferentes pasos:
   (a) Extraer el conjunto mínimo de ubicaciones que componen la secuencia (o eliminar los elementos repetidos de la secuencia). El costo de esta operación depende del número de elementos diferentes en la secuencia. En el peor de los casos todos los elementos son diferentes y el costo es n^2 comparaciones.
   (b) Crear la matriz de $u \times u$, con u el largo del conjunto mínimo de ubicaciones. Costo: 1.
   (c) Recorrer los $n-1$ pares origen-destino de la secuencia y por cada par de ubicaciones buscar sus índice en el conjunto mínimo de ubicaciones. Con estos índices actualizar el contador del elemento respectivo de la matriz. Costo: $(n-1)\times(2n+1)$.
   (d) Sumar los elementos de cada fila de la matriz y dividir cada fila por la suma total. Costo: $n(n+1)$.
3. Línea 5: Asignación, costo 1.
4. Línea 6: Iteración de costo $m-1 \times Costo$(Línea 7) + $Costo$(Línea 8)

5. Línea 7: Obtener probabilidad de transición requiere buscar los índices del par de posiciones S_b^j y S_b^{j+1} en el conjunto mínimo de ubicaciones y luego acceder con estos índices en la matriz. Costo: $2n + 1$
6. Línea 8: Calcular logaritmo y sumar el resultado a $sim_{a,b}$, costo 1.

A continuación se suman todos los costos y se utiliza la notación asintótica para obtener la complejidad.

$$\begin{aligned} C(n,m) &= n + m + n^2 + 1 + (n-1)(2n+1) + n(n+1) + 1 + (m-1)(2n+1+1) \\ &= n + m + n^2 + 1 + 2n^2 - 2n + n - 1 + n^2 + n + 1 + 2nm - 2n + 2m - 2 \\ &= 4n^2 + 2nm - n + 3m - 1 \\ &= O(n^2) + O(nm) + O(m) \\ &= O(n^2) + O(nm) \end{aligned}$$

El número de comparaciones del algoritmo TPM es $O(n^2) + O(nm)$, donde predominará uno de los dos términos dependiendo del valor de m.

Características generales y limitaciones

El algoritmo TPM es un algoritmo que no considera la ubicación geográfica de las posiciones, por lo que calificaría como una comparación de movilidad de tipo *location-based*. Por lo anterior, una de las limitaciones de este algoritmo es la alta sensibilidad a los cambios de paradas de una secuencia de posiciones, aun cuando el cambio no sea significativo desde una perspectiva espacial.

El algoritmo TPM mide principalmente patrones espaciales de pares origen-destino, la temporalidad está implícita en el orden de la secuencia. Lo anterior permite al algoritmo comparar tablas de transacciones de distintos periodos, por ejemplo: construir una matriz TPM con los registros de un mes y comparar su afinidad con la secuencia de posiciones de una semana.

3.2. Algoritmo basado en el método de distancia de edición espaciotemporal (EDM)

El algoritmo basado en el método de distancia de edición espaciotemporal (EDM, por su sigla en inglés) fue propuesto por Yuan y Raubal para medir la similitud de trayectorias de usuarios de telefonía. Este algoritmo está basado en el método de distancia de edición de *strings* propuesto por Wagner y Fischer (1974). En términos generales, la distancia de edición espaciotemporal entre dos trayectorias corresponde al costo de transformar una trayectoria en otra.

3.2.1. Construcción del perfil de movilidad

Sea la entrada del algoritmo EDM: T_a y T_b, dos tablas de transacciones de n y m registros con los atributos mínimos *Identificador de tarjeta,Marca temporal,Posición Parada de subida*; donde T_a y T_b corresponden a las transacciones de las tarjetas t_a y t_b respectivamente.

El método EDM utiliza como perfil de movilidad la trayectoria asociada a cada una de las Tabla de transacciones.

La trayectoria t asociada a una Tabla de transacciones T se define como un conjunto de tuplas ordenadas temporalmente $t = [p^1(temp_1, lat_1, long_1), p^2(temp_2, lat_2, long_2), ..., p^u(temp_u, lat_u, long_u)]$, con u el número de transacciones de la tabla T, donde cada elemento p_i con $\mathrm{i} \in [1, u]$ es una posición espaciotemporal tal que $temp_\mathrm{i}$ pertenece al atributo *Marca temporal*, y $lat_\mathrm{i}, long_\mathrm{i}$ pertenece al atributo *Posición Parada de subida* asociado a cada elemento $temp_\mathrm{i}$.

En el caso de que el atributo *Posición Parada de bajada* y *Marca temporal de bajada* estén disponibles, entonces la secuencia se construye iterando intercaladamente sobre los elementos de los atributos *Posición Parada de subida* y *Posición Parada de bajada* ordenados según los atributos *Marca temporal* y *Marca temporal de bajada*.

3.2.2. Comparación de perfiles de movilidad

La comparación de dos perfiles de movilidad (i.e. dos trayectorias) del algoritmo EDM consiste en calcular un indicador del costo de transformación de una trayectoria a otra. Sean dos trayectorias $t_a = [p_a^1, ...p_a^n]$ y $t_b = [p_b^1, ...p_b^m]$, de acuerdo con el algoritmo propuesto por Wagner y Fischer, hay tres formas de transformar t_a en t_b:

- p_a^n se elimina y el resto $p_a^1, ...p_a^{n-1}$ se transforma a $p_b^1, ...p_b^m$
- Se transforma $p_a^1, ...p_a^n$ a $p_b^1, ...p_b^{m-1}$ y se inserta p_b^m al final
- p_a^n se reemplaza con p_b^m y el resto $p_a^1, ...p_a^{n-1}$ se transforma a $p_b^1, ...p_b^{m-1}$

Cada una de las tres operaciones: eliminar, insertar y reemplazar, tiene un costo asociado. Al minimizar el costo total de transformación es posible obtener la mejor combinación de operaciones para transformar una trayectoria en otra, costo denominado *distancia de edición*. La función recursiva para calcular la distancia de edición queda definida como sigue:

$$\begin{aligned} DistanciaEdicion_{a,b}[\mathrm{i}, j] = min(&DistanciaEdicion[\mathrm{i}-1, j] + Costo[eliminar(p_a^\mathrm{i})], \\ &DistanciaEdicion[\mathrm{i}, j-1] + Costo[insertar(p_b^j)], \\ &DistanciaEdicion[\mathrm{i}-1, j-1] + Costo[reemplazar(p_a^\mathrm{i}, p_b^j)]) \end{aligned}$$

El método de distancia de edición espaciotemporal propuesto por Yuan y Raubal considera que los costos de transformación de una trayectoria pueden ser cuantificados a través del impacto de las transformaciones en el centroide de la trayectoria a modificar. El centroide se

calcula como el promedio de las posiciones espaciotemporales. Luego, al agregar o eliminar una posición que esté lejos del centroide esta operación tendrá mayor impacto que agregar o eliminar una posición que esté cercana al centroide. Además, Yuan y Raubal definen una constante $c \in [0, 1]$ para regular la influencia entre el costo espacial y temporal. Cuando $c = 0$ la distancia es completamente espacial, cuando $c = 1$ la distancia es completamente temporal. Las funciones de costo de las operaciones son las siguientes:

$$\text{Cost[Delete } (p_{1i})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} x_{1k}}{n-1}\right)\right]^2 + \left[\left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} y_{1k}}{n-1}\right)\right]^2\right\} + c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} t_{1k}}{n-1}\right)\right]^2} \quad (1)$$

$$\text{Cost [Insert } (p_{2j})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right) - \left(\frac{\sum_{k=1}^{n} x_{1k} + x_{2j}}{n+1}\right)\right]^2 + \left[\left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right) - \left(\frac{\sum_{k=1}^{n} y_{1k} + y_{2j}}{n+1}\right)\right]^2\right\} + c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right) - \left(\frac{\sum_{k=1}^{n} t_{1k} + t_{2j}}{n+1}\right)\right]^2} \quad (2)$$

$$\text{Cost [Replace } (p_{1i}, p_{2j})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} x_{1k} + x_{2j}}{n}\right)\right]^2 + \left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} y_{1k} + y_{2j}}{n}\right)\right]\right\}^2 + c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right) - \left(\frac{\sum_{k=1,k\neq i}^{n} t_{1k} + t_{2j}}{n}\right)\right]^2} \quad (3)$$

Figura 3.1: Costos de las operaciones: eliminación, inserción y reemplazo, utilizados en el algoritmo EDM. (1) Desplazamiento del centroide de la trayectoria t_1 luego de eliminar p_1^i. (2) Desplazamiento del centroide de la trayectoria t_1 luego de insertar p_2^j. (3) Desplazamiento del centroide de la trayectoria t_1 luego de reemplazar p_1^i por p_2^j. Fuente: Yuan y Raubal (2014)

La Tabla 3.2 presenta el pseudocódigo del Algoritmo EDM. La *salida* de este algoritmo corresponde a la distancia de edición entre dos trayectorias. Esta distancia pertenece al rango $[0, \infty]$, donde distancia 0 indica máxima similitud. Por otro lado, no existe una cota superior definida ya que la distancia dependerá del costo de las operaciones de transformación que a su vez no tienen un límite definido y dependen de la distancia geotemporal entre las trayectorias.

3.2.3. Análisis del algoritmo EDM

Complejidad

A continuación se presenta un análisis asintótico del peor caso del número de comparaciones del algoritmo EDM. Se describe el costo de los diferentes pasos de este algoritmo, dejando

Tabla 3.2: Algoritmo EDM.

Algoritmo 2: EDM

Input: tablas de transacciones T_a, T_b

Output: Distancia $\mathrm{d}_{a,b}$, con $\mathrm{d}_{a,b} \in [0, \infty)$

```
1:  funcion EDM(T_a, T_b) :
2:      t_a ← extraerTrayectoria(T_a)
3:      t_b ← extraerTrayectoria(T_b)
4:      D ← Matriz(largo(t_a) + 1, largo(t_b) + 1)
5:      for i in [0...largo(t_a) − 1] :
6:          D[i + 1, 0] ← D[i, 0] + eliminar(t_a^i)
7:      for j in [0...largo(t_b − 1)] :
8:          D[0, j + 1] ← D[0, j] + insertar(t_b^j)
9:      for i in[1..largo(t_a)]
10:         for j in[1..largo(t_b)]
11:             eliminar_{i,j} ← D[i − 1, j] + eliminar(t_a^i)
12:             insertar_{i,j} ← D[i, j − 1] + insertar(t_b^j)
13:             reemplazar_{i,j} ← D[i − 1, j − 1] + reemplazar(t_a^i, t_b^j)
14:             D[i, j] ← minimo(eliminar_{i,j}, insertar_{i,j}, reemplazar_{i,j})
15:     d_{a,b} ← D[largo(t_a), largo(t_b)]
16:     retornar d_{a,b}
```

con costo 1 operaciones como asignaciones, cómputos aritméticos o cualquier operación de costo constante. Los parámetros de la función de costo son el largo de las T_a y T_b. Se utiliza como referencia las líneas del pseudocódigo de la 3.2.

1. Línea 2 y 3: Extraer las trayectorias de las tablas T_a y T_b. Costo: n y m respectivamente.
2. Línea 4: Crear la matriz de $n \times m$. Costo: 1.
3. Línea 5: Iteración de costo $n - 1 \times Costo(Línea\ 6)$
4. Línea 6: El costo de esta línea queda descrito por las distintas operaciones que la componen:
   (a) Acceso a la Matriz D. Costo 1.
   (b) La operación $eliminar(t_a^{\mathrm{i}})$ según su fórmula de la Figura 3.1 requiere realizar 6 sumas iterando sobre el largo de la trayectoria t_a más otras operaciones aritméticas de costo constante. Costo $6n$.
   (c) Asignación a la Matriz D. Costo 1
5. Línea 7: Iteración de costo $m - 1 \times Costo(Línea\ 8)$
6. Línea 8: El costo de esta línea queda descrito por las distintas operaciones que la componen:
   (a) Acceso a la Matriz D. Costo 1.
   (b) La operación $insertar(t_b^{\mathrm{i}})$ según su fórmula de la Figura 3.1 requiere realizar 6 sumas iterando sobre el largo de la trayectoria t_a más otras operaciones aritméticas de costo constante. Costo $6n$.

(c) Asignación a la Matriz D. Costo 1

7. Línea 9: Iteración de costo $n-1 \times Costo(\textit{Línea 10})$
8. Línea 10: Iteración de costo $m-1 \times Costo(\textit{Línea 11-14})$
9. Línea 11: Equivalente al costo de la Línea 6.
10. Línea 12: Equivalente al costo de la Línea 8.
11. Línea 13: El costo de esta línea queda descrito por las distintas operaciones que la componen:
    (a) Acceso a la Matriz D. Costo 1.
    (b) La operación $reemplazar(t_a^i, t_b^j)$ según su fórmula de la Figura 3.1 requiere realizar 6 sumas iterando sobre el largo de la trayectoria t_a más otras operaciones aritméticas de costo constante. Costo $6n$.
    (c) Asignación a la Matriz D. Costo 1
12. Línea 14: Encontrar el mínimo entre tres valores es de costo constante. Costo 1.
13. Línea 15: Asignación de costo 1

A continuación se suman todos los costos y se utiliza la notación asintótica para obtener la complejidad general.

$$\begin{aligned}
C(n,m) &= n+m+1+(n-1)(1+6n+1)+(m-1)(1+6n+1)+ \\
&\quad (n-1)(m-1)(1+6n+1+1+6n+1+1+6n+1)+1+1 \\
&= n+m+n-1+6n^2-6n+n-1+m-1+6mn-6n-1+m-1+ \\
&\quad (nm-m-n+1)(18n+6)+3 \\
&= 6n^2+6mn-9n+3m-2+18n^2m-12mn-18n^2-12n-6m+6 \\
&= 18n^2m-12n^2-6mn-21n-3m+4 \\
&= O(n^2m)-O(n^2)-O(mn)-O(n)-O(m) \\
&= O(n^2m)
\end{aligned}$$

El número de comparaciones del algoritmo EDM es $O(n^2m)$.

Características generales y limitaciones

El algoritmo EDM es un algoritmo donde el movimiento es considerado como un conjunto de puntos espaciotemporales que describen una trayectoria. Esto permite que cualquier par de trayectorias puede ser comparado, es decir, no es necesario que las dos trayectorias tengan alguna posición en común para calcular la distancia. Por otro lado, este paradigma no establece relaciones de importancia entre las posiciones ni entre las secuencias de lugares visitados más frecuentes.

Si bien este método se llama distancia de edición, la medida presentada no cumple las condiciones para ser una métrica. A diferencia de las versiones de este método con operaciones de costo constante, la versión espaciotemporal tiene funciones de costo que hacen que no siempre se cumpla la igualdad $distanciaEdicion_{a,b} = distanciaEdicion_{b,a}$.

Por otro lado, es necesario que el tamaño del intervalo de tiempo en que se observan las trayectorias que se estén comparando sea similar, de otro modo la diferencia de largo de las trayectorias afecta significativamente la distancia.

3.3. Algoritmo basado en Regiones de Interés y un vector de características (RoIs-FV)

A diferencia de los métodos presentados anteriormente, este método fue diseñado e implementado en este trabajo. La motivación para proponer este algoritmo se encuentra en la variedad de trabajos presentes en la literatura sobre descripción y agrupación de usuarios a través de variables descriptivas de registros de movilidad (ver Sección 2.1). De manera similar a los trabajos de clasificación de usuarios, este método define una distancia utilizando técnicas de minería de datos para establecer relaciones de similitud entre registros de movilidad. En particular, en este método se emplea *Clustering jerárquico* para encontrar las zonas más importantes de una trayectoria, y se aplica una metodología de *Extracción de características* para describir la movilidad registrada en las tablas de transacciones.

3.3.1. Construcción del perfil de movilidad

Sea la entrada del algoritmo RoIs-FV: T_a y T_b, dos tablas de transacciones de n y m registros con los atributos mínimos *Identificador de tarjeta,Marca temporal, Parada de subida, Posición Parada de subida, Número de viaje, Número de etapa, Tipo de transporte, Tipo de tarjeta*; donde T_a y T_b corresponden a las transacciones de las tarjetas t_a y t_b respectivamente.

El perfil de movilidad de este método está compuesto por dos estructuras de datos: Regiones de Interés y un vector de características. A continuación se describe el significado y proceso de extracción de ambas estructuras para la tabla T_a. El proceso de extracción del perfil de movilidad asociado a la tabla T_b es equivalente.

Regiones de Interés

Las Regiones de Interés (RoIs, por su sigla en inglés) representan las ubicaciones más importantes de un usuario. La importancia de una ubicación está asociada a la frecuencia de visita durante el periodo de observación. Considerando que para trasladarse a una zona de destino pueden existir diferentes rutas, las RoIs se obtienen agrupando registros de paradas cercanas y seleccionando las zonas que concentran un número significativo de transacciones.

Una región de interés corresponde a una zona definida por un centroide y un radio r. Para extraer las regiones de interés de una Tabla de transacciones se utiliza *clustering* jerárquico aglomerativo de la siguiente forma:

En primer lugar es necesario transformar la tabla T_a para obtener las posiciones visitadas junto a su porcentaje de visita. Para lograr esto se extrae la secuencia de posiciones asociada a la tabla T_a, la cual se define como $S_a : [s_a^1, s_a^2, ..., s_a^n]$, y corresponde a los n valores del atributo *Posición Parada de subida* ordenados según el atributo *Marca temporal*. Sea $P_a : [p_a^1, p_a^2 ... p_a^u]$ el conjunto mínimo de las u posiciones que componen S_a. Luego, se construye una lista L_a tal que cada elemento de la lista $l_a^{\mathrm{i}}, \mathrm{i} \in [0, u]$ corresponde a la tupla $(p_a^{\mathrm{i}}, porcentaje\_de\_visitas_{\mathrm{i}})$, donde el primer elemento corresponde al elemento $\mathrm{i} - esimo$ de P_a, y el segundo elemento corresponde a la razón entre el número de veces que p_a^{i} aparece en S_a y n, el número total de elementos de S_a.

Antes de comenzar el *clustering* se precisa definir una tabla de distancia entre las distintas posiciones de parada de la lista P_a. Para medir la distancia entre dos posiciones se utilizó la fórmula de Vincenty [3] que considera la forma elipsoidal de la Tierra y permite obtener una distancia en metros. En el estado inicial del algoritmo de *clustering* aglomerativo todas las posiciones de P_a son consideradas *clusters* independientes, donde el centroide de cada *cluster* queda definido por la única observación que lo compone.

El algoritmo de agrupación itera sobre los siguientes pasos:

1. Si hay un solo *cluster* se detiene el algoritmo. Si no, se encuentran los *clusters* más cercanos.
2. Si los *clusters* más cercanos están a menos de r metros, se agrupan en un *cluster*. Si no, se detiene el algoritmo.
3. Se calcula el centroide del nuevo *cluster* como el promedio de las posiciones de los *clusters* que lo integran.
4. Se calculan las distancias entre el nuevo *cluster* y los *clusters* que no se modificaron.

En el paso 2 es necesario definir el parámetro r. En este trabajo se utilizó un radio de 500 metros, de esta manera el diámetro de cada RoI abarca el promedio de distancia caminada observado en la literatura (Tirachini, 2015). En el paso 3, la distancia entre un *cluster* c_k y el nuevo *cluster* $c_{\mathrm{i}\cup j}$ (formado de la unión de los *clusters* c_{i} y c_j), se utilizó el promedio de las distancias de los *clusters* c_{i} y c_j con c_k, como muestra la siguiente fórmula:

$$dist(c_{\mathrm{i}\cup j}, c_k) = (dist(c_{\mathrm{i}}, c_k) + dist(c_j, c_k)/2)$$

Cuando se detiene el algoritmo de *clustering* todas las posiciones de parada han sido agrupadas en zonas circulares de 500 metros de radio. Entonces, utilizando la lista L_a se procede a sumar los porcentaje de visita correspondientes a las posiciones de paradas que componen cada *cluster*.

Finalmente, se reconocen como RoIs aquellos *clusters* que reúnen el mayor porcentaje de visitas. Para esto se utiliza un parámetro v el cual marca el porcentaje de visitas total que deben reunir las RoIs. Se define un conjunto vacío de RoIs, y se itera agregando el *cluster* con mayor porcentaje de visitas al conjunto de RoIs hasta que la suma total de los porcentaje de visitas del conjunto de RoIs se igual o mayor que el parámetro v.

[3]Esta distancia fue desarrollada por Thaddeus Vincenty en 1975 Vincenty (1975), y se utilizó la implementación de la librería *geopy* (https://geopy.readthedocs.io/en/1.10.0/)

Vector de características

Un vector de características es un conjunto de variables descriptivas que caracterizan diferentes aspectos de algún fenómeno estudiado, en este caso la movilidad de un usuario. La extracción de características es un área bastante desarrollada en minería de datos, y es posible encontrar una descripción de la metodología y técnicas relacionadas en (Guyon y Elisseeff, 2006). Para distinguir las características que mejor describen un fenómeno observado se requiere dominio del área del problema. Por lo anterior, la principal fuente de características de este método han sido variables descriptivas utilizadas previamente en la literatura (ver Sección 2.1). Se seleccionaron variables cuya definición estuviese disponible y fuese compatible con los datos de transporte público. Durante el diseño de este método también se agregaron variables descriptivas complementarias. A continuación la tabla 3.3 presenta el conjunto de características utilizado.

Tabla 3.3: Variables descriptivas que componen el vector de características.

Tipo de Característica	**Característica**	**Tipo de dato**	**Unidad**	**Dominio**
Temporal	Hora de inicio promedio primer viaje	Ordinal continuo	segundos	[0,86400]
	Hora de inicio promedio último viaje	Ordinal continuo	segundos	[0,86400]
	Número de días con viajes	Ordinal discreto	-	N
	Moda del número de viajes por día	Ordinal discreto	-	N
	Frecuencia de la moda del número de viajes por día	Ordinal discreto	-	N
	Promedio de número de viajes por día	Ordinal continuo	-	N
Espacial	Distancia viajada	Ordinal continuo	metros	$R+$
	Mínima distancia viajada promedio	Ordinal continuo	metros	$R+$
	Máxima distancia viajada diaria promedio	Ordinal continuo	metros	$R+$
	Radio de giro	Ordinal continuo	metros	$R+$
	Entropía temporalmente no correlacionada	Ordinal continuo	-	$R+$
	Entropía aleatoria	Ordinal continuo	-	$R+$
	Porcentaje de primeras paradas diferentes	Ordinal continuo	-	[0.0,100.0]
	Porcentaje de últimas paradas diferentes	Ordinal continuo	-	[0.0,100.0]
Demográfica	Tipo de tarjeta	Nominal	-	Depende del Sistema de transporte
Actividad	Promedio de tiempo de actividad más corta por día	Ordinal continuo	segundos	$R+$
	Promedio de tiempo de actividad más larga por día	Ordinal continuo	segundos	$R+$
Modo de transporte	Número de etapas por viaje más frecuente	Ordinal discreto	-	N
	Porcentaje de días con viajes exclusivos en bus	Ordinal continuo	-	[0.0,100.0]
	Porcentaje de días con viajes exclusivos en metro	Ordinal continuo	-	[0.0,100.0]
	Porcentaje de viajes en bus	Ordinal continuo	-	[0.0,100.0]

De las 21 características expuestas en la tabla 3.3, las siguientes características se calcularon separado para los días laborales y no laborales (semana y fin de semana) del periodo de observación:

- Hora de inicio promedio primer viaje
- Hora de inicio promedio último viaje
- Promedio de tiempo de actividad más larga por día
- Promedio de tiempo de actividad más corta por día
- Porcentaje de primeras paradas diferentes
- Porcentaje de últimas paradas diferentes
- Promedio de viajes por día

Por lo anterior el conjunto total de características es de 28 variables, con las cuales se construye un vector $v_a : [v_a^1, v_a^2, ..., v_a^{28}]$ que representa el comportamiento asociado a la Tabla de transacciones T_a.

La tabla 3.3 también muestra que las variables poseen diferentes tipo de datos y dominios. Considerando que los vectores de características serán utilizados para comparar la movilidad de distintos usuarios, es necesario realizar un preprocesamiento sobre los vectores de manera que los valores sean comparables. El preprocesamiento consiste en llevar todas las características a un dominio común. Para lograr lo anterior es necesario realizar los siguientes procesos:

- Manejo de datos faltantes
- Transformación de datos categóricos
- Detección de *outliers*
- Normalizar características

3.3.2. Comparación de perfiles

Sean dos perfiles de movilidad $P_1 : (RoIs_{p1}, v_{p1})$ y $P_2 : (RoIs_{p2}, v_{p2})$, cada uno compuesto por un conjunto de RoIs y un vector de características. La comparación de dos perfiles de movilidad del algoritmo RoIs-FV consta de dos etapas consecutivas: 1. Comparar regiones de interés y 2. Comparar vectores de características.

Comparar regiones de interés

Esta etapa consiste en determinar si dos perfiles de movilidad comparten un mínimo de lugares importantes. Es una etapa de salida binaria, con resultados *Positivo* o *Negativo*. Se define el parámetro i como la cantidad de regiones de interés compartidas requeridas. Luego, el resultado de esta etapa es positivo cuando dos perfiles comparten al menos i regiones de interés. El resultado de esta etapa es negativo cuando los perfiles comparten menos de i regiones de interés.

Sean dos conjuntos de regiones de interés: $RoIs_{p1} : [r_{p1}^1 ... r_{p1}^s]$ y $RoIs_{p2} : [r_{p2}^1 ... r_{p2}^t]$, asociados a P_1 y P_2 respectivamente. Para calcular el número de regiones compartidas entre P_1 y P_2 se itera sobre el conjunto $RoIs_{p1}$, y se evalúa la distancia entre cada elemento r_{p1}^{i} con cada elemento de $RoIs_{p2}$. La distancia se mide entre los centroides de cada región utilizando la fórmula de distancia *vincenty* en metros. Si la distancia entre dos pares de RoIs es menor que 500 metros, se dice que ambos perfiles comparten esas regiones de interés.

Luego de iterar por todos los elementos de $RoIs_{p1}$, si la cantidad de regiones compartidas es igual o mayor que i, el resultado de esta etapa es *Positivo*, entonces se procede a la segunda etapa. En caso contrario, el resultado de esta etapa es *Negativo*, entonces el algoritmo se detiene y la comparación de los dos perfiles resulta en el estado *No comparables*.

Comparar vectores de características

Esta etapa consiste en calcular la disimilitud entre dos perfiles de movilidad midiendo la distancia entre los vectores de características que los componen. La función de distancia utilizada es fácilmente intercambiable y puede tener distintos niveles de complejidad.

Sean dos vectores de características v_{p1} y v_{p2}, asociados a P_1 y P_2 respectivamente. La distancia entre v_{p1} y v_{p2} queda determinada por la medida de disimilitud que se utilice, la cual queda representada por el parámetro $f\mathrm{dis}t$.

$$\mathrm{d}(P_1, P_2) = \begin{cases} f\mathrm{dis}t(v_{p1}, v_{p2}), & \text{si } RoIsCompartidos(P_1, P_2) \geq \mathrm{i} \\ \textbf{No comparables}, & \text{si } RoIsCompartidos(P_1, P_2) < \mathrm{i} \end{cases}$$

A continuación se presentan las medidas de disimilitud que fueron utilizadas en la evaluación de este algoritmo.

Medidas de disimilitud Sean dos vectores $X_1 : [x_1^1, x_1^2, ...x_1^n]$ y $X_2 : [x_2^1, x_2^2, ...x_2^n]$, ambos de largo n.

Euclidiana

$$e\mathit{ucl}\mathrm{id}i\mathit{ana}(X_1, X_2) = \sqrt{\sum_{\mathrm{i}=1}^{n} (x_1^{\mathrm{i}} - x_2^{\mathrm{i}})^2} \tag{3.1}$$

Manhattan

$$manhattan(X_1, X_2) = \sum_{\mathrm{i}=1}^{n} |x_1^{\mathrm{i}} - x_2^{\mathrm{i}}| \tag{3.2}$$

Bray-Curtis

$$braycurtis(X_1, X_2) = \frac{\sum_{\mathrm{i}=1}^{n} |x_1^{\mathrm{i}} - x_2^{\mathrm{i}}|}{\sum_{\mathrm{i}=1}^{n} |x_1^{\mathrm{i}} + x_2^{\mathrm{i}}|} \tag{3.3}$$

Chebyshev

$$chebyshev(X_1, X_2) = max_{\mathrm{i}} |x_1^{\mathrm{i}} - x_2^{\mathrm{i}}| \tag{3.4}$$

Canberra

$$canberra(X_1, X_2) = \sum_{\mathrm{i}=1}^{n} \frac{|x_1^{\mathrm{i}} - x_2^{\mathrm{i}}|}{|x_1^{\mathrm{i}}| + |x_2^{\mathrm{i}}|} \tag{3.5}$$

Hamming

$$hamming(X_1, X_2) = \sum_{\mathrm{i}=1}^{n} \mathrm{eq}(x_1^{\mathrm{i}}, x_2^{\mathrm{i}}), \mathrm{eq}(a, b) = \begin{cases} 1, \textbf{si } a = b \\ 0, \textbf{si } a \neq b \end{cases} \tag{3.6}$$

La Tabla 3.4 presenta el pseudocódigo del Algoritmo RoIs-FV. La *salida* de este algoritmo corresponde a la distancia entre dos vectores de características. Esta distancia pertenece al rango $[0, \infty]$, donde distancia 0 indica máxima similitud. Por otro lado, no existe una cota superior definida ya que la distancia dependerá de la función $f\mathrm{dis}t$ utilizada.

Tabla 3.4: Algoritmo RoIs-FV.

Algoritmo 1:	RoIs-FV
Input:	T_1, T_2: tablas de transacciones, $porcentaje\_visita$: Porcentaje mínimo de transacciones que deben agrupar las RoIs, min_RoIs: Mínimo número de RoIs que deben compartir dos perfiles para ser comparables, $f\mathrm{dis}t$: Función de distancia.
Output:	Disimilitud $\mathrm{d}_{1,2}$, con $0 \leq \mathrm{d}_{1,2}$. Valor -1 indica que las tablas no son comparables.

```
1:  funcion RoIsFV(T1, T2, radio, porcentaje_visita, min_RoIs, fdist) :
2:      RoIs1 ← extraerRoIs(T1, radio, porcentaje_visita)
3:      RoIs2 ← extraerRoIs(T2, radio, porcentaje_visita)
4:      if compararRoIs(RoIs1, RoIs2, min_RoIs) :
5:          v1 ← extraerVector(T1)
6:          v2 ← extraerVector(T2)
7:          d1,2 ← fdist(v1, v2)
8:          retornar d1,2
9:      else :
10:         retornar −1
```

3.3.3. Análisis del algoritmo RoIs-FV

Complejidad

A continuación se presenta un análisis asintótico del peor caso del número de comparaciones del algoritmo RoIs-FV. Se describe el costo de los diferentes pasos de este algoritmo, dejando con costo 1s operaciones como asignaciones y cómputos aritméticos. Los parámetros de la función de costo son el largo de las T_a y T_b. Se utiliza como referencia las líneas del pseudocódigo de la 3.4.

1. Línea 2: El costo asociado a extraer las regiones de interés para la tabla T_1 se calcula en diferentes pasos:
   (a) Extraer la secuencia de posiciones de la tabla T_1. Costo n.
   (b) Contar las ocurrencias de cada posición en la secuencia de posiciones. En el peor de los casos todas las posiciones son diferentes. Costo $\frac{n(n+1)}{2}$.
   (c) Según los pasos del algoritmo de *Clustering* jerarquico descrito en la Sección 3.3.1, el peor de los casos sería agrupar todas las posiciones sin detenerse. Entonces el costo recae en buscar el par de clusters más cercano y calcular las nuevas distancias entre los clusters. Costo $\frac{n^2(n+1)}{2} + \frac{n(n+1)}{2}$
   (d) Buscar los *clusters* que reúnen el porcentaje $porcentaje_visita$ de las transacciones requiere en el peor de los casos buscar $n-1$ veces en un arreglo en que cada iteración se reduce su tamaño de n a 1. Costo: $\frac{n(n+1)}{2}$

   El costo total de la Línea 2 es $n + (\frac{n^2(n+1)}{2}) + 3(\frac{n(n+1)}{2})$.
2. Línea 3: La operación es simétrica a la de la línea 2, considerando la tabla T_b. Costo: $m + (\frac{m^2(m+1)}{2}) + 3(\frac{m(m+1)}{2})$
3. Línea 4: El costo de la operación $comparar RoIs$ se calcula midiendo la distancia entre cada par de $RoIs$ de dos conjuntos de regiones. En el peor caso se tienen que hacer todas las comparaciones y dado que el calculo de la distancia entre los centroides es constante; el costo de esta operación es: nm.
4. Línea 5: El costo de la operación $extraer Vector$ dependerá del costo de las variables descriptivas. En el caso del vector utilizado en esta tesis las distintas variables tienen distintos costos. La variable más costosa es el *Radio de Giro* que requiere n^2 comparaciones. Por tanto la cota máxima de la complejidad del vector de 28 variables es $28n^2$.
5. Línea 6: La operación es simétrica a la de la línea 5, considerando la tabla T_b. Costo: $28m^2$.
6. Línea 7: Todas las medidas de disimilitud presentadas anteriormente son constantes en el largo de T_a y T_b, solo dependen del largo del vector de características que en este trabajo es 28.

A continuación se suman todos los costos y se utiliza la notación asintótica para obtener la complejidad.

$$
\begin{aligned}
C(n,m) &= n + \frac{n^2(n+1)}{2} + 3\frac{n(n+1)}{2} + m + \frac{m^2(m+1)}{2} + 3\frac{m(m+1)}{2} + nm + \\
&\quad 28n^2 + 28m^2 + 28 \\
&= \frac{n^3}{2} + \frac{n^2}{2} + \frac{3n^2}{2} + \frac{3n}{2} + \frac{m^3}{2} + \frac{m^2}{2} + \frac{3m^2}{2} + \frac{3m}{2} + 28n^2 + 28m^2 + \\
&\quad nm + n + m + 28 \\
&= \frac{n^3}{2} + \frac{m^3}{2} + 30n^2 + 30m^2 + nm + n + m + 28 \\
&= O(n^3) + O(m^3) + O(n^2) + O(m^2) + O(nm) + O(n) + O(m) + O(1) \\
&= O(n^3 + m^3)
\end{aligned}
$$

El número de comparaciones del algoritmo TPM queda descrito como $O(n^3) + O(m^3)$, donde predominará la complejidad de uno de los dos términos dependiendo del valor de m.

Características generales y limitaciones

El algoritmo RoIs-FV es un algoritmo donde los registros de las transacciones son representados como un conjunto de regiones de interés y un conjunto de características descriptivas de la movilidad. Las regiones de interés permiten establecer un mínimo de similitud geoespacial. En la Sección 2.1.2 se discute sobre la regularidad de viajes en torno a dos ubicaciones, esto sugiere que estos dos lugares caracterizan a los usuarios. Por tanto, para reconocer a un usuario no es necesario compararlo con todos, solo con aquellos que comparten las regiones importantes. Sin embargo, si el usuario cambia significativamente sus regiones importantes entonces este algoritmo no podrá reconocerlo.

Por otro lado, es necesario que el tamaño del intervalo de tiempo sea equivalente entre las trayectorias que se estén comparando, de otro modo existirán grandes diferencias entre los valores de las variables descriptivas, afectando la normalización de los valores y la distancia final entre los vectores. Por ejemplo si se compara la distancia recorrida, entre una Tabla de una semana de transacciones y otra Tabla dos semanas de transacciones, es esperable que exista una gran diferencia en la cantidad de metros recorridos.

El conjunto de características extraídas se puede extender o encoger dependiendo de los datos disponibles asociados a las transacciones. La función de distancia también puede ser reemplazada por una más compleja que asocie pesos constantes o variables a las características.

Capítulo 4

Metodología

En este capítulo se describe la metodología desarrollada para evaluar la factibilidad de reconocer usuarios a través de la observación de los registros en el sistema de transporte público. Como se mencionó anteriormente, en esta tesis se utilizaron dos bases de datos de transacciones de transporte público. En primer lugar, una base de datos de transacciones de Transantiago, el sistema de transporte público de Santiago, Chile. En segundo lugar, una base de datos de transacciones de la Société de transport de l'Outaouais (STO), sistema de transporte público de Gatineau, Canadá. Las diferencias de estas bases de datos permitieron que el problema de esta tesis se abordara desde dos perspectivas:

1. Cuán estables en el tiempo son las representaciones de la movilidad de los usuarios.
2. Cuán reconocibles son las representaciones de la movilidad de los usuarios.

En esta tesis se aplicó una metodología de trabajo basada en el proceso Cross Industry Standard Process for Data Mining (CRISP-DM). La descripción de cada fase del proceso se encuentra divida en las siguientes secciones: Comprensión de los datos, Preparación de los datos, Modelación y Evaluación.

4.1. Comprensión de los datos

A continuación se describen las dos bases de datos con las que se trabajó en esta tesis. La descripción aquí expuesta es más bien específica a cada base de datos y al transporte público de la ciudad respectiva. Para una descripción general sobre registros de tarjetas inteligentes es posible dirigirse a la Revisión Bibliográfica, Sección 2.4.

4.1.1. Base de datos de Transantiago, Santiago, Chile

Santiago es la capital de Chile, ubicada en la Región Metropolitana. El área metropolitana de Santiago está formada por 37 comunas, cuya distribución geopolítica se muestra en la

Figura 4.1[1]. De acuerdo a la proyección del Instituto Nacional de Estadísticas de Chile, basada en el Censo del 2002, la población para el 2015 de la Región Metropolitana era de 7.092.988 habitantes. Según el mismo censo, el año 2002 la Provincia de Santiago contaba con 4.728.443 habitantes y una densidad de 2.304,83 hab/km^2, siendo la provincia más poblada del país.

Figura 4.1: Comunas de Santiago.

Transantiago es el sistema de transporte público que opera en Santiago desde Febrero del 2007. Es un sistema integrado de buses y metro. En la actualidad, ocho empresas integran el sistema: Metro de Santiago y siete empresas operadoras de buses. La flota de buses es de 6.500 buses con recorridos que cubren más de 11.000 paradas. El metro está formado por cinco líneas con un total de 108 estaciones.

La tarjeta inteligente Bip! es el único medio de pago de Transantiago. Las transacciones se realizan cada vez que el usuario ingresa a un bus o a una estación de metro. La salida del sistema no se registra. Los buses tienen una tarifa plana, y el metro tiene tres tarifas dependiendo del horario. El sistema de pago permite realizar hasta dos transbordos en un plazo de dos horas. En caso de que el usuario realice un transbordo entre servicios de distinta tarifa, se le carga la diferencia.

La base de datos utilizada contiene los registros de dos periodos de transacciones de Transantiago, la semana del 14/04/2013 al 21/04/2013, y la semana del 23/09/2013 al 29/09/2013. La base de datos fue generada por el software de análisis de datos de transporte público (ADATRAP), basado en la investigación de Munizaga y Palma y desarrollado por el departamento de Ingeniería Civil, división de Transporte de la Universidad de Chile. El software ADATRAP enriquece la información con la bajada estimada y el tiempo de la bajada estimada, y además agrupa las transacciones de cada usuario en etapas y viajes.

[1]Fuente Osmar Valdebenito [CC BY-SA 2.5 (http://creativecommons.org/licenses/by-sa/2.5)], via Wikimedia Commons

La base de datos está compuesta por dos tablas de registros de transacciones:

1. Tabla etapas semana abril 2013 (14/04/2013 al 21/04/2013)
2. Tabla etapas semana septiembre 2013 (23/09/2013 al 29/09/2013)

Las dos tablas están compuestas por los mismos atributos, cada registro de las tablas corresponde a la información asociada a una transacción en Transantiago. A continuación se expone el nombre, tipo y descripción de cada atributo.

- **tiempo_subida** (*timestamp*): Fecha y hora en la que se realiza la transacción.
- **id** (*integer*): Identificador de tarjeta Bip! con la que se realiza la transacción. Este identificador es anónimo, no está asociado de ninguna forma a información personal.
- **x_subida** (*float*): Coordenada X de la transacción en el sistema UTM. Todas las transacciones se encuentran en la zona 19H del modelo WGS84.
- **y_subida** (*float*): Coordenada Y de la transacción en el sistema UTM.
- **tipo_transporte** (*varchar*): Corresponde al tipo de servicio en el que se realiza la transacción. Puede tomar los valores: Metro, Bus o Zona Paga.
- **serviciosentidovariante** (*varchar*): Servicio en el que se realizó la transacción. En caso de bus corresponde al recorrido y sentido. En caso de metro corresponde a la línea de metro.
- **tipo_dia** (*varchar*): Laboral, Sábado o Domingo.
- **nviaje** (*integer*): Número de viaje asociado a la tarjeta.
- **netapa** (*integer*): Número de etapa asociado al número de viaje de la transacción.
- **x_bajada** (*float*): Coordenada X de la bajada estimada en el sistema UTM.
- **y_bajada** (*float*): Coordenada Y de la bajada estimada en el sistema UTM.
- **par_subida** (*varchar*): Identificador del paradero donde se realiza la transacción.
- **par_bajada** (*varchar*): Identificador del paradero de la bajada estimada.
- **zona_subida** (*integer*): Código de área de la zonificación EOD 777 asociado al paradero de subida.
- **zona_bajada** (*integer*): Código de área de la zonificación EOD 777 asociado al paradero de bajada.
- **adulto** (*integer*): Código de tipo de tarjeta. Existen diversos tipos de tarjeta de los cuales destacan tres: escolar, adulto, anciano.
- **tiempo_bajada** (*timestamp*): Fecha y hora en la que se realiza la bajada estimada.

Tabla 4.1: Tabla descriptiva de las observaciones de la base de datos de Santiago.

	Tabla etapas Abril 2013	Tabla etapas Septiembre 2013	Total
N transacciones	35.984.008	33.525.838	69.509.846
Por tipo de día			
Laboral	29.060.075	28.347.306	57.407.381
Sábado	3.215.871	3.290.652	6.506.523
Domingo	3.708.062	1.887.880	5.595.942
Por tipo de transporte			
Bus	22.573.522	20.680.660	43.254.182
Metro	13.410.486	12.845.178	26.255.664
N tarjetas	3.419.039	3.321.592	5.033.380
N paradas	11.281	11.299	11.345

La Tabla 4.1 muestra un resumen descriptivo de las transacciones observadas en la base de datos de Santiago. En esta tabla es posible notar que en la semana de abril 2013 se realizaron más transacciones, lo cual se justifica ya que el periodo de esta tabla es un día más largo que el de la tabla de septiembre. La Figura 4.2 muestra el histograma de transacciones observadas durante la semana de septiembre 2013 durante los días de la semana y fin de semana, agrupadas según el horario de cada transacción por cada 15 minutos.

La Figura 4.2 muestra el número de transacciones según hora del día. Se puede ver claramente los *peaks* mañana y tarde de demanda del transporte público de Santiago durante los días laborales. En cambio durante los días de fin de semana se observa una distribución más uniforme y de mucho menor frecuencia que los días laborales. En la semana de abril de 2013 se observa un comportamiento similar.

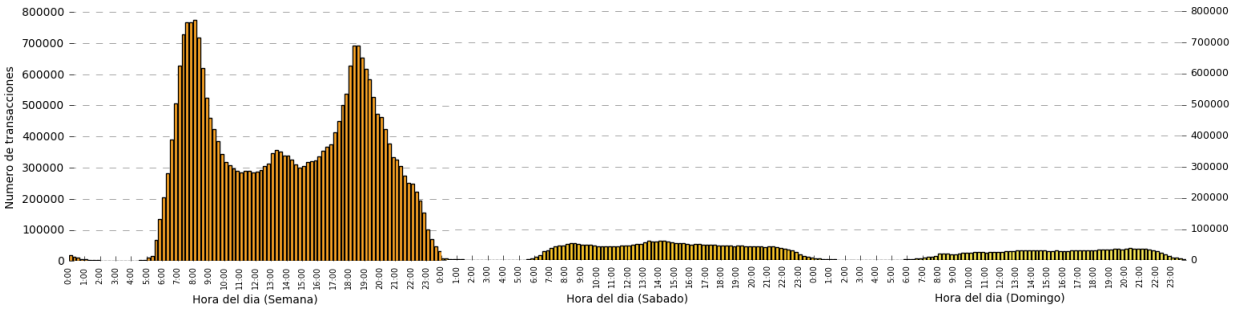


Figura 4.2: Histograma de transacciones de la semana de septiembre 2013 en día de semana, sábado y domingo, con rangos temporales de 15 minutos.

En cuanto a la distribución geográfica de los registros de la Base de datos de Santiago, la Figura 4.3 ilustra las posiciones de las transacciones de una muestra aleatoria de 5.000 usuarios, separada en dos conjuntos temporales: una semana de abril 2013 y otra de septiembre 2013. Los puntos verdes corresponden a paradas de abordaje y los puntos azules a paradas de destino. Es posible observar que la distribución de paradas es equivalente en ambos períodos, y cubre toda el área metropolitana de Santiago previamente graficada en la Figura 4.1.

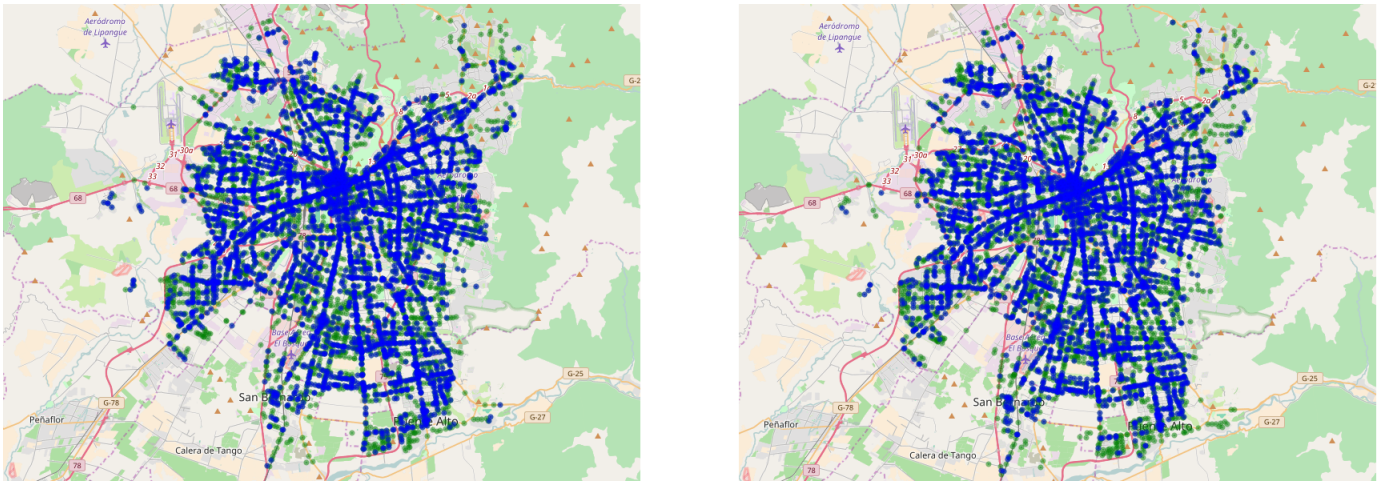


Figura 4.3: Distribución geográfica de las transacciones de 5.000 usuarios. (A) semana de abril 2013. (B) semana de septiembre 2013.

4.1.2. Base de datos de Société de transport de l'Outaouais, Gatineau, Canadá

Gatineau es una ciudad ubicada en la región Outaouais, provincia de Quebec, Canadá. Se encuentra a orillas del río Ottawa, río que marca el límite con la capital canadiense Ottawa, con la cual forman la conurbación Ottawa-Gatineau. Según el censo del gobierno canadiense del año 2016, Gatineau tiene una población de 276245 habitantes y una densidad de 773.7/km2. Gatineau está formada por cinco municipalidades: Aylmer, Buckingham, Gatineau, Hull y Masson-Angers. La Figura 4.4 muestra su distribución geopolítica.

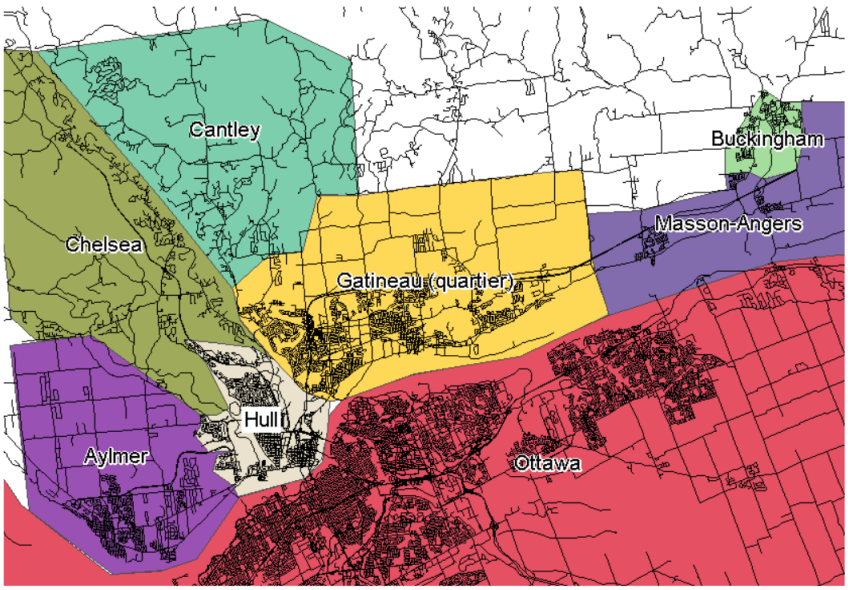


Figura 4.4: Distribución geográfica de las municipalidades de Gatineau y sus alrededores.

La Société de transport de l'Outaouais (STO) es el sistema de transporte público que opera en Gatineau desde 1971. Es un sistema formado únicamente por buses, con una flota actual de 305 buses distribuidos en 68 rutas. Gatineau también cuenta con la presencia del sistema de transporte de Ottawa, operado por la empresa OC Transpo, en las rutas que conectan Gatineau-Ottawa en ambos sentidos. Estas rutas son compartidas por la STO y OC Transpo, cuyos sistemas de pago mediante tarjeta inteligentes se encuentran integrados.

La tarjeta inteligente Multi es el medio de pago integrado de la STO, con la cual se realiza el pago de aproximadamente 80 % de los viajes registrados. El resto de las transacciones se realizan mediante boletos adquiridos previamente o pago en el mismo bus. Las transacciones se realizan cada vez que el usuario ingresa a un bus. La salida del sistema no se registra. El sistema de pago permite realizar un transbordo sin costo adicional para el usuario en un plazo de dos horas. Existen diferentes tarifas para cada medio de pago, siendo más conveniente pagar con tarjeta. Además es posible realizar abonos que permiten al usuario realizar viajes ilimitados en periodos de un día, tres días, mes o año. Por otro lado, distintos tipos de usuarios poseen distintos tipos de tarjetas, los cuales permiten acceder a tarifas diferenciadas. En general, los usuarios pueden ser agrupados en cuatro tipos: adulto, estudiante, infante y adulto mayor.

La base de datos utilizada contiene los registros de dos años de transacciones del sistema de la STO, correspondientes a los años 2012 y 2013. La base de datos fue generada utilizando la metodología propuesta por Trépanier et al. para estimar la bajada de cada transacción.

La base de datos está compuesta por dos tablas de registros de transacciones:

1. Tabla Resultat Destination STO 2012
2. Tabla Resultat Destination STO 2013

Las dos tablas están compuestas por los mismos atributos, cada registro de las tablas corresponde a la información asociada a una transacción en la red de la STO. A continuación se expone el nombre, tipo y descripción de cada atributo.

- **carteId** (*integer*): Identificador de tarjeta Multi con la que se realiza la transacción. Cada tarjeta Multi está asociada al nombre del usuario, sin embargo este identificador fue anonimizado y no está asociado de ninguna forma a información personal.
- **timestamp** (*timestamp*): Fecha y hora en la que se realiza la transacción.
- **typeTransport** (*varchar*): Corresponde al tipo de servicio en el que se realiza la transacción. En el caso de Gatineau, solo toma el valor *Bus*.
- **numLigne** (*varchar*): Servicio en el que se realizó la transacción. En el caso de Gatineau, corresponde al recorrido y sentido del bus que se abordó.
- **direction** (*varchar*): Dirección en la que se abordó el servicio de transporte.
- **typeJour** (*varchar*): Tipo de día, puede tomar los valores: *SE* (semana), *SA* (sábado), *DI* (domingo), *F1* (feriado de semana santa, 29 de marzo 2013).
- **estPremier** (*integer*): Variable binaria que indica si la transacción corresponde al inicio de un viaje.

- **estDernier** (*integer*): Variable binaria que indica si la transacción corresponde al final de un viaje.
- **estSeul** (*integer*): Variable binaria que indica si la transacción corresponde a un viaje de una sola etapa.
- **stopId** (*integer*): Identificador del paradero donde se realiza la transacción.
- **destId** (*integer*): Identificador del paradero de bajada estimada.
- **distanceOD** (*float*): Distancia entre la parada de subida y la parada de bajada.

La Tabla 4.2 muestra un resumen descriptivo de las transacciones observadas en la base de datos de Gatineau. En esta tabla es posible notar que en el año 2013 se realizaron más transacciones, que aumentó el número de tarjetas observadas en 20.000 y que también aumentaron el número de paradas. Por otro lado, es posible notar que los usuarios de transporte público de Gatineau viajan principalmente durante la semana. Esto se ve reflejado en la Figura 4.5, que representa la distribución temporal de las transacciones del año 2012. La figura muestra el histograma de transacciones observadas del año 2012 durante los días de la semana y fin de semana, agrupadas según el horario de cada transacción por cada 15 minutos. La Figura 4.5 grafica el número de transacciones según la hora del día. Se puede ver claramente los *peaks* mañana y tarde de demanda del transporte público de Gatineau durante los días laborales. En cambio, durante los días de fin de semana se observa una distribución más uniforme y de mucho menor frecuencia que los días laborales. El año 2013 se observa un comportamiento similar.

Tabla 4.2: Tabla descriptiva de las observaciones de la base de datos de Gatineau.

	Tabla STO 2012	Tabla STO 2013	Total
N transacciones	10.618.519	11.143.343	21.761.862
Semana	9.923.571	10.360.776	20.284.347
Sábado	409.927	454.643	864.570
Domingo	285.021	319.179	604.200
Feriado	N.A.	8.745	8.745
N tarjetas	61.547	81.755	114.508
N paradas	3.254	3.612	3.653

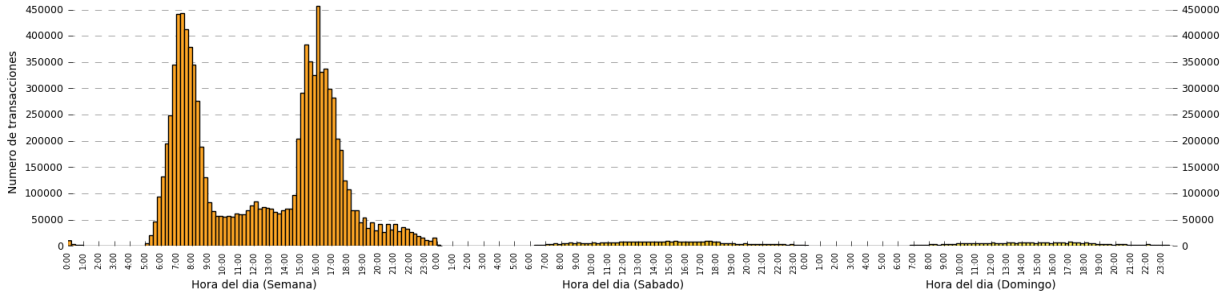


Figura 4.5: Histograma de transacciones del año 2012 en día de semana, sábado y domingo, con rangos temporales de 15 minutos.

En cuanto a la distribución geográfica de los registros de la Base de datos de Gatineau, la Figura 4.6 muestra las posiciones de las paradas del sistema STO. Al relacionar esta Figura con la distribución geográfica de las municipalidades de Gatineau, se observa que su distribución demarca el límite con Ottawa, y se produce una concentración de paradas el la municipalidad de Gatineau. Al comparar con la distribución de paradas de Santiago, se evidencia una gran diferencia entre ambas redes de transporte público, tanto en tamaño del sistema, como en la forma de la distribución geográfica.

Figura 4.6: Distribución geográfica de las paradas de buses del sistema de la STO, Gatineau.

4.2. Preparación de los datos

Según la metodología CRISP-DM, la preparación de los datos ocurre de manera paralela con la Modelación. Lo anterior se debe a que la Preparación de los datos varía según dos factores: el objetivo de cada iteración del Modelo y la base de datos que se esté utilizando. A continuación se presenta una descripción de los distintos procesos a los cuales los datos fueron sometidos por los distintos procesos de la Modelación.

4.2.1. Preprocesamiento base de datos Santiago

Debido a que algunos algoritmos de caracterización y reconocimiento de usuarios son sensibles al tamaño del período de observación, la primera operación que se realizó sobre los datos de Santiago fue equiparar los períodos de observación de las dos tablas de transacciones que componen la base de datos de Santiago. La tabla etapas semana abril 2013 (14/04/2013 al 21/04/2013) contiene las observaciones de ocho días a diferencia de la tabla etapas semana septiembre 2013, que abarca siete días. Se seleccionaron las transacciones realizadas en el

período 14/04/2013 al 20/04/2013 de la tabla etapas semana abril 2013. Luego, se seleccionaron las transacciones de aquellas tarjetas presentes en las dos tablas de transacciones (etapas abril y septiembre). Posteriormente, se seleccionaron aquellas tarjetas que tuvieran un mínimo de diez transacciones por semana. Las operaciones anteriores cumplen dos funciones: reducir el tamaño de la base de datos y seleccionar aquellas tarjetas de las cuales se tiene la información suficiente como para ser reconocidas. La base de datos resultante está compuesta por transacciones pertenecientes a tarjetas.

En el caso de la transacciones realizadas en metro, los valores de los atributos *par_subida* y *par_bajada* indican el nombre de la estación y línea de metro, e.g. Santa Ana L2. Esto hace que la misma estación pueda tener distinto nombre en caso de estaciones con combinación a otras líneas, e.g. Santa Ana L5. Por lo anterior se estandarizan los nombres de las paradas, removiendo la línea de las estaciones de metro de los valores de *par_subida* y *par_bajada*.

Las operaciones restantes que se aplican sobre los datos dependen del algoritmo con el que sean utilizados en la etapa de *Modelación*. Si bien algunas operaciones se repiten entre algoritmos, por claridad son descritas de manera separada paso a paso.

Algoritmo TPM

1. De los atributos que poseen las tablas de la base de datos de Santiago (descritas en la Sección 4.1.1), se seleccionan los siguientes atributos:
   (a) *id*
   (b) *tiempo_subida*
   (c) *par_subida*
   (d) *par_bajada*
   (e) *zona_subida*
   (f) *zona_bajada*
2. Se ordenan las transacciones según *id* y *tiempo_subida*.
3. Finalmente las tablas de transacciones se agrupan según el atributo *id*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

Algoritmo EDM

1. De los atributos que poseen las tablas de la base de datos de Santiago (descritas en la Sección 4.1.1), se seleccionan los siguientes atributos:
   (a) *id*
   (b) *tiempo_subida*
   (c) *par_subida*
   (d) *par_bajada*
   (e) *tiempo_bajada*
2. El algoritmo EDM utiliza las posiciones geográficas de cada parada. Para asociar el GPS de cada parada con su correspondiente identificador, se utiliza un diccionario que

contiene las posiciones con coordenadas latitud y longitud de cada paradero y estación de metro. Luego, se agregan los atributos [*lat_subida*,*long_subida*,*lat_bajada*,*long_bajada*] que corresponden a la latitud y longitud asociadas a la parada de subida y parada de bajada respectivamente.

3. Se remueven los atributos *par_subida* y *par_bajada*.
4. Se ordenan las transacciones según *id* y *tiempo_subida*.
5. Finalmente las tablas de transacciones se agrupan según el atributo *id*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

Algoritmo RoIs-FV

1. De los atributos que poseen las tablas de la base de datos de Santiago (descritas en la Sección 4.1.1), se seleccionan los siguientes atributos:
   (a) *id*
   (b) *tiempo_subida*
   (c) *par_subida*
   (d) *par_bajada*
   (e) *tipo_transporte*
   (f) *nviaje*
   (g) *netapa*
   (h) *adulto*
2. El algoritmo RoIs-FV utiliza las posiciones geográficas de cada parada para calcular las RoIs y diferentes características. Para asociar el GPS de cada parada con su correspondiente identificador, se utiliza un diccionario que contiene las posiciones con coordenadas latitud y longitud de cada paradero y estación de metro. Luego, se agregan los atributos [*lat_subida*,*long_subida*,*lat_bajada*,*long_bajada*] que corresponden a la latitud y longitud asociadas a la parada de subida y parada de bajada respectivamente.
3. Se agrega el atributo *weekday* el cual representa el día de la semana asociado a cada transacción. Se codifican los días de la semana de Lunes a Domingo con los números de 0 a 6 respectivamente.
4. Se agrega el atributo *date* que corresponde a la fecha de cada transacción.
5. Se ordenan las transacciones según *id* y *tiempo_subida*.
6. Se agrega el atributo *time_diff* cuyo valor es el tiempo que transcurre entre cada transacción y la anterior.
7. Finalmente las tablas de transacciones se agrupan según el atributo *id*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

4.2.2. Preprocesamiento base de datos Gatineau

La primera operación que se realizó sobre los datos de Gatineau fue unir los registros de las dos tablas de transacciones (2012 y 2013) en una sola tabla. La Figura 4.7 muestra el

histograma del número de tarjetas por el número de días entre la primera y última transacción de cada tarjeta observada durante el periodo 2012-2013. Es posible observar una tendencia decreciente de la cantidad de tarjetas cuanto más largo sea el periodo en que se registraron las transacciones. En particular, la mayoría de los usuarios no registró más de dos meses de transacciones y solo alrededor de 17.000 usuarios fueron observados en un periodo mayor a un año.

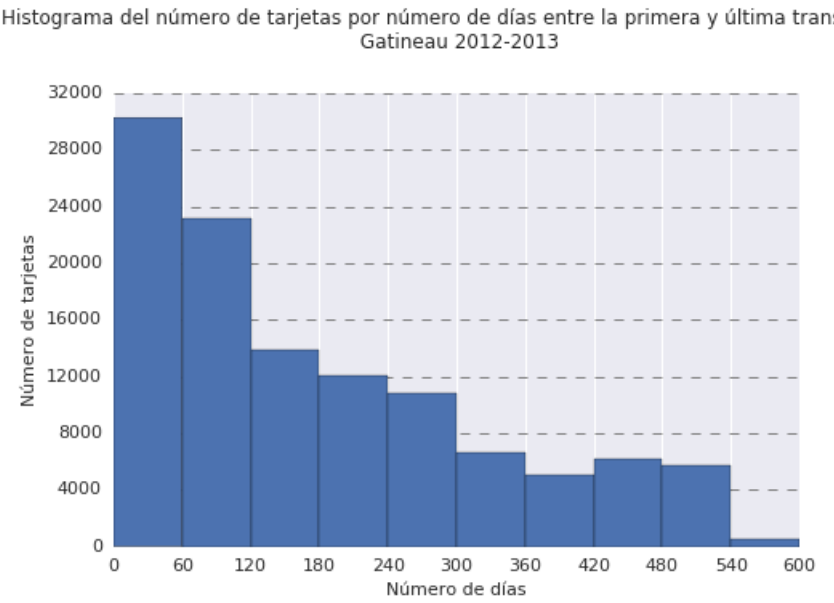


Figura 4.7: Histograma del número de tarjetas por número de días entre la primera y última transacción, Base de datos Gatineau 2012-2013.

A continuación se describe la preparación de los datos según las necesidades de cada algoritmo.

Algoritmo TPM

1. De los atributos que poseen las tablas de la base de datos de Gatineau (descritas en la Sección 4.1.2), se seleccionan los siguientes atributos:
   (a) *carteId*
   (b) *timestamp*
   (c) *stopId*
   (d) *destId*
2. Algunos experimentos con el algoritmo TPM varían el nivel de agregación espacial utilizado, i.e. en vez de utilizar los identificadores de paradas se utilizan identificadores de zonas. Se utiliza un diccionario que contiene los identificadores de paradas asociados a dos zonificaciones: Zona 400 y Zona 66, donde el número indica la cantidad de zonas en las que se agrupan las paradas. Luego se agregan los atributos [*stop_zona400_Id*,*dest_zona400_Id*,*stop_zona66_Id*,*dest_zona66_Id*]

3. Se ordenan las transacciones según *carteId* y *timestamp*.
4. Finalmente se agrupan los registros según el atributo *carteId*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

Algoritmo EDM

1. De los atributos que poseen las tablas de la base de datos de Gatineau (descritas en la Sección 4.1.2), se seleccionan los siguientes atributos:
   (a) *carteId*
   (b) *timestamp*
   (c) *stopId*
   (d) *destId*
2. El algoritmo EDM utiliza las posiciones geográficas de cada parada. Para asociar el GPS de cada parada con su correspondiente identificador, se utiliza un diccionario que contiene las posiciones con coordenadas latitud y longitud de cada paradero. Luego, se agregan los atributos [*lat_subida*,*long_subida*,*lat_bajada*,*long_bajada*] que corresponden a la latitud y longitud asociadas a la parada de subida y parada de bajada respectivamente.
3. Se remueven los atributos *stopId* y *destId*.
4. Se ordenan las transacciones según *carteId* y *timestamp*.
5. Finalmente se agrupan los registros según el atributo *carteId*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

Algoritmo RoIs-FV

1. De los atributos que poseen las tablas de la base de datos de Gatineau (descritas en la Sección 4.1.2), se seleccionan los siguientes atributos:
   (a) *carteId*
   (b) *timestamp*
   (c) *typeJour*
   (d) *estPremier*
   (e) *estDernier*
   (f) *estSeul*
   (g) *stopId*
   (h) *destId*
2. El algoritmo RoIs-FV utiliza las posiciones geográficas de cada parada para calcular las RoIs y diferentes características. Para asociar el GPS de cada parada con su correspondiente identificador, se utiliza un diccionario que contiene las posiciones con coordenadas latitud y longitud de cada paradero y estación de metro. Luego, se agregan los atributos [*lat_subida*,*long_subida*,*lat_bajada*,*long_bajada*] que corresponden a la latitud y longitud asociadas a la parada de subida y parada de bajada respectivamente.

3. Se agrega el atributo *date* derivado de *timestamp* que corresponde a la fecha de cada transacción.
4. Se ordenan las transacciones según *carteId* y *timestamp*.
5. Se agrega el atributo *time_diff* cuyo valor es el tiempo que transcurre entre cada transacción y la anterior.
6. Finalmente se agrupan los registros según el atributo *carteId*, es decir agrupando los datos en tablas de transacciones separadas por el identificador de tarjeta.

4.3. Modelación

En esta sección se describen los procesos mediante los cuales se mide la factibilidad de reconocer a un usuario mediante la observación de su movilidad en transporte público. La modelación fue divida en cinco etapas: en la primera etapa se implementan los algoritmos, en la segunda y tercera etapa se expone el diseño de los experimentos que se desarrollaron en esta tesis. En la cuarta etapa se describen los diferentes escenarios en que fueron ejecutados los experimentos de la segunda y tercera etapa. Finalmente, en la quinta etapa se describe el postprocesamiento que se realiza sobre los resultados para que sean comparables entre los distintos escenarios.

4.3.1. Etapa 1: Implementación de los algoritmos de caracterización y reconocimiento de usuarios

En esta etapa se implementan los tres algoritmos descritos en el capítulo 3: el algoritmo TPM, el algoritmo EDM y el algoritmo RoIs-FV. La implementación de estos algoritmos se hizo en el lenguaje de programación Python. Para cada algoritmo se construye un paquete con tres archivos:

1. Un módulo de preprocesamiento
2. Un módulo de extracción y comparación de perfiles
3. Un módulo de funciones auxiliares

4.3.2. Etapa 2: Diseño de la medición de la variabilidad de los perfiles de movilidad en el tiempo

El objetivo de esta etapa es capturar las variaciones del perfil de movilidad de los usuarios a lo largo del tiempo. Los tres algoritmos descritos en el capítulo 3 calculan la distancia o similitud de la movilidad registrada en dos tablas de transacciones. Si las dos tablas de transacciones almacenan registros de un mismo usuario, la distancia o similitud calculada por los algoritmos puede ser entendida como un indicador de la variación de la movilidad de un usuario entre los periodos a los que pertenecen las tablas comparadas. Luego, si se tienen

tablas de transacciones de un usuario en diferentes ventanas de tiempo, es posible analizar la variabilidad de la movilidad como una serie temporal.

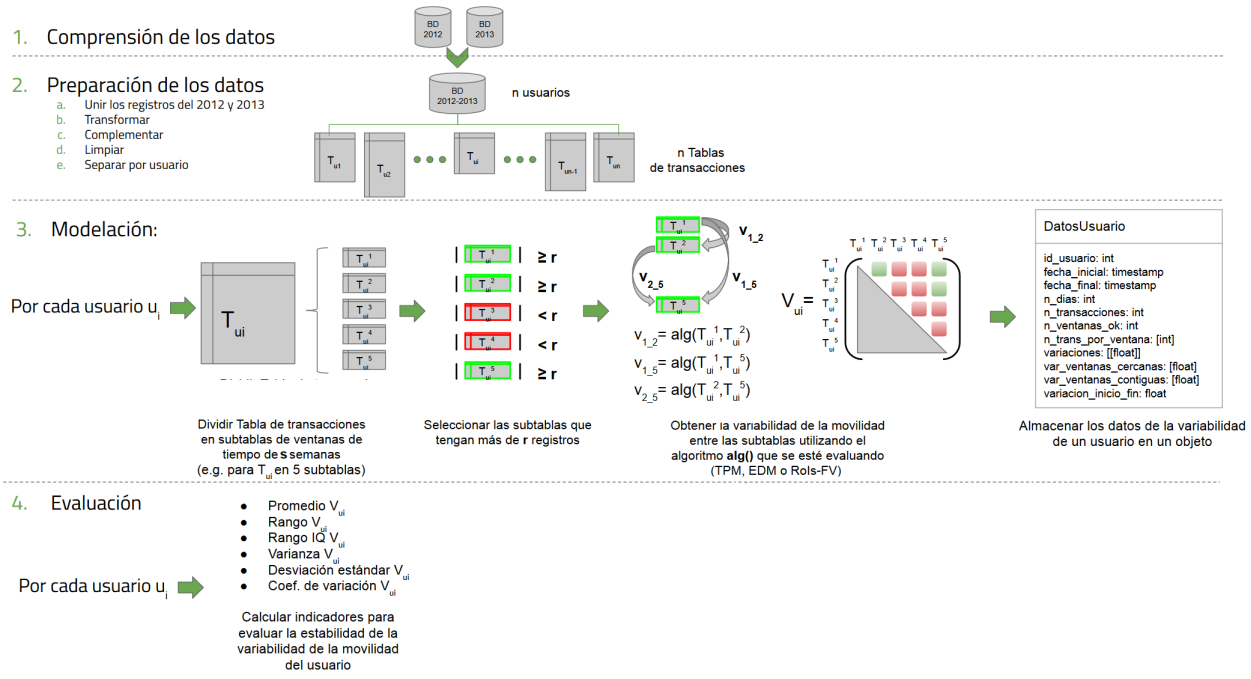


Figura 4.8: Descripción gráfica de la Etapa 2: proceso de medición de la variabilidad de los perfiles de movilidad en el tiempo.

La Figura 4.8 describe el proceso para obtener las variaciones de la movilidad. Este proceso se resume de la siguiente manera: En primer lugar, se preprocesan los registros de movilidad en transporte público y se agrupan por usuario. Posteriormente, los registros de transacciones de cada usuario se dividen en subtablas de transacciones asociadas a ventanas de tiempo. Luego, con un algoritmo de caracterización y comparación de movilidad se calcula la similitud entre las subtablas de transacciones de cada usuario, obteniendo la variación de la movilidad a lo largo de su periodo de actividad[2]. Finalmente se utilizan métricas para evaluar la variabilidad de los usuarios, descritas en la Sección 4.4.1. Se itera sobre este proceso variando el algoritmo de caracterización y comparación (TPM, EDM o RoIs-FV) y variando el tamaño de las ventanas de tiempo.

Al finalizar la preparación de datos, descrita en la sección anterior, se cuenta con las transacciones agrupadas por identificador de tarjeta, es decir, se cuenta con tablas de transacciones asociadas a cada usuario. A continuación, se describe paso a paso la transformación de una tabla de transacciones de un usuario a un objeto. Este último almacena la variación de la caracterización de la movilidad del usuario correspondiente, a lo largo de su periodo de actividad.

Sea una tabla de transacción T_u una tabla que contiene los registros de las transacciones asociadas a la tarjeta de identificador t durante los años 2012 y 2013. La tarjeta t está asociada al usuario U.

[2]Periodo de actividad hace referencia al intervalo de tiempo entre la primera y última transacción de cada usuario.

1. Se asignan valores a los siguientes parámetros:
   (a) **s** *(integer)*: número de semanas de la ventana de tiempo.
   (b) **r** *(integer)*: número mínimo de registros que debe tener una subtabla asociada a una ventana de tiempo para poder extraer un perfil de movilidad.
   (c) **algoritmo** *(función)*: algoritmo TPM, EDM o RoIs-FV.
2. Se divide el período de observación en $n = largo\_periodo/r$ ventanas de tiempo, con $largo\_periodo$ el número de semanas del período de actividad de la tarjeta t. Se utilizan semanas de calendario de Lunes a Domingo para que las ventanas sean comparables entre usuarios.
3. Se agrupan las transacciones de la tabla T_u en las n ventanas de tiempo definidas en el paso anterior, quedando T_u dividida en un conjunto de n Subtablas de transacciones $[T_u^0, T_u^1, ...T_u^n]$. De este conjunto, algunas Subtablas quedaran sin transacciones, otras con menos transacciones que r, y otras con suficientes transacciones para extraer un perfil de movilidad.
4. Por cada par de Subtablas $(T_u^{\mathrm{i}}, T_u^j)$, tal que $\mathrm{i} < j$, se calcula la variación de la movilidad según la siguiente formula:

$$\text{Variacion}(T_u^{\mathrm{i}}, T_u^j) = \begin{cases} \text{Sin transacciones, si } |T_u^{\mathrm{i}}| == 0 \vee |T_u^j| == 0 \\ \text{Sin transacciones suficientes, si } |T_u^{\mathrm{i}}| < r \vee |T_u^j| < r \\ algoritmo(T_u^{\mathrm{i}}, T_u^{\mathrm{i}+1}), \text{ en cualquier otro caso} \end{cases}$$

   El resultado las comparaciones se almacena en una matriz V de tamaño $n \times n$. Cada variación $Variacion(T_u^{\mathrm{i}}, T_u^j)$, con $\mathrm{i} < j$, se almacena en la celda $M[\mathrm{i}, j]$, i.e. se compara cada tabla con todas las tablas de ventanas posteriores y en particular, no se calcula la variación de una tabla consigo misma. Por lo anterior, la matriz resultante es una matriz triangular superior, donde la diagonal también toma valores nulos.
5. Se crea un objeto con los siguientes datos de cada usuario:
   (a) Identificador de usuario
   (b) Fecha inicial: Fecha de la primera transacción observada.
   (c) Fecha final: Fecha de la última transacción observada.
   (d) Número de días: Número de días entre la fecha inicial y la fecha final.
   (e) Número de transacciones: Número de transacciones total observadas.
   (f) Número de ventanas: Número de ventanas con transacciones suficientes.
   (g) Transacciones por ventana: Arreglo de número de transacciones por cada ventana.
   (h) Variaciones: Matriz V con las variaciones entre las tablas de transacciones asociadas a cada ventana de tiempo.
   (i) Variaciones de ventanas cercanas: Arreglo que contiene las variaciones entre las tablas de transacciones consecutivas más cercanas. En caso de que alguna tabla de transacciones no tenga las transacciones suficientes, se calcula la variación con la tabla de transacciones que tenga las transacciones suficientes asociada a la ventana de tiempo más cercana. Por ejemplo, en la Figura 4.8, donde de las cinco ventanas de tiempo disponibles, tres tienen las ventanas suficientes, las variaciones de las ventanas cercanas sería el arreglo $[v_{1_2}, v_{2_5}]$.

(j) Variaciones de ventanas contiguas: Arreglo que contiene solo las variaciones entre tablas de transacciones de ventanas de tiempo consecutivas que tengan transacciones suficientes. Corresponde a la diagonal de la submatriz $V[0 : n-1, 1 : n]$. Por ejemplo, en la Figura 4.8, las variaciones de las ventanas contiguas sería el arreglo $[v_{1_2}]$.

Una vez extraída la información de variabilidad de cada usuario, se procede a evaluar la variabilidad de los usuarios con las medidas descritas en la Sección 4.4.1.

4.3.3. Etapa 3: Diseño de la medición de la identificabilidad de los perfiles de movilidad

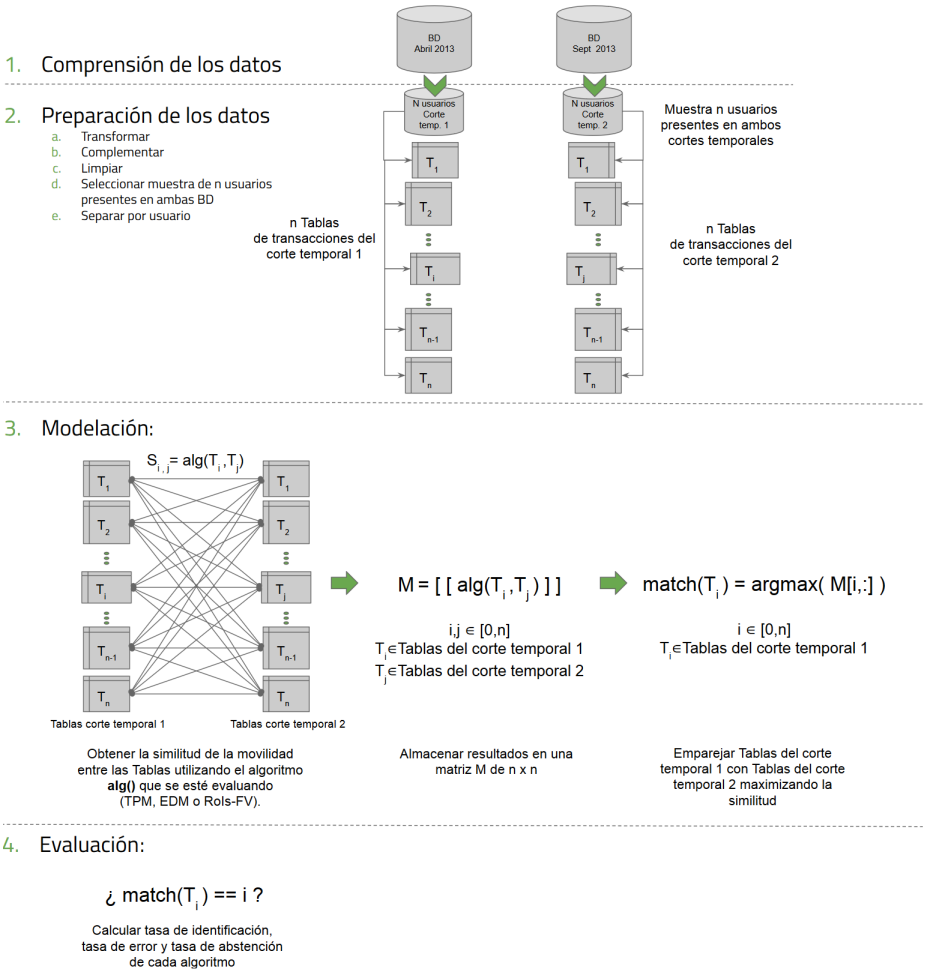


Figura 4.9: Descripción gráfica de la Etapa 3: proceso de medición de la identificabilidad de los perfiles de movilidad.

El objetivo de esta etapa es evaluar la factibilidad de hacer un *matching* de usuarios entre dos bases de datos temporalmente independientes. La Figura 4.9 describe el proceso para medir la capacidad de reconocer usuarios en dos cortes temporales. El emparejamiento es individual, es decir, para cada usuario registrado en la primera base de datos se busca el

usuario más similar en la segunda base de datos. La similitud se determina con los algoritmos presentados en el Capítulo 3.

A continuación se describen los pasos para obtener la similitud entre la movilidad registrada en las tarjetas y así evaluar cuan distinguible es el movimiento de los usuarios de transporte público.

1. En primer lugar se debe dividir los registros de la base de datos de movilidad en dos periodos independientes.
2. Se selecciona una muestra de registros de n tarjetas cuyos identificadores estén en la base de datos de ambos periodos.
3. Se agrupan los registros de ambas bases de datos según el identificador de tarjeta, generando dos grupos de tablas de transacciones con los registros de cada tarjeta. Se ordenan ambos grupos según el identificador de la tarjeta, de tal manera que los registros de la tarjeta i en el grupo uno coincida con la tarjeta i del grupo dos.
4. Por cada tabla de transacciones del conjunto de tablas del primer periodo se realiza una comparación de la movilidad con todas las tablas del segundo periodo. Esta comparación se realiza con cada algoritmo presentado en el Capitulo 3. Es importante notar que los algoritmos en ningún momento utilizan el identificador de la tarjeta.
5. Las comparaciones de cada algoritmo se almacena en una matriz M de $n \times n$, donde cada celda i, j representa la similitud (o distancia, dependiendo del algoritmo) entre la tarjeta i y la tarjeta j. En particular la diagonal corresponde a la similitud entre la movilidad de la misma tarjeta en los dos periodos de movilidad observados.
6. Por cada fila i de la matriz M se selecciona la celda $[\mathrm{i}, m]$ que almacena la similitud máxima, lo cual corresponde a hacer el *matching* entre la tarjeta i y la tarjeta m.

Una vez extraídos los *matching* de cada tarjeta, se procede a evaluar cuan identificables son los usuarios con las medidas descritas en la Sección 4.4.2.

4.3.4. Etapa 4: Ejecución de las mediciones diseñadas en las etapas 2 y 3 sobre diferentes escenarios

A continuación se explican detalles de la ejecución de las etapas 2 y 3.

Ejecución etapa 2

Para medir la variabilidad de la movilidad de los usuarios de transporte público se utilizó la base de datos de Gatineau. Esta base de datos posee dos características propicias: en primer lugar, contiene los registros de dos años de transacciones y en segundo lugar, las tarjetas inteligentes de la STO están asociadas a un nombre de usuario, y si bien el identificador de la tarjeta es anónimo es posible asumir sin inconvenientes la relación 1 a 1 entre una tarjeta y un usuario. Estas características permiten observar los cambios en la movilidad de los usuarios en largos periodos de tiempo.

De los parámetros descritos en el primer paso para obtener la variabilidad de un usuario (ver Sección 4.3.2)) se utilizan las siguientes configuraciones de Ventanas s y mínimo de transacciones r:

1. Ventanas de una semana con mínimo 8 transacciones
2. Ventanas de dos semanas con mínimo 16 transacciones
3. Ventanas de cuatro semanas con mínimo 32 transacciones

Al ejecutar la etapa 2 con estas configuraciones, se encontraron usuarios que no tienen dos ventanas con las transacciones mínimas, por lo que no se les puede asociar una métrica de variabilidad. Con el resto de los usuarios se procede según lo señalado.

Se divide el período de observación 2012-2013 (106 semanas) en $n = 106/v$ ventanas de tiempo. Como se mencionó previamente, se utilizan semanas de Lunes a Domingo. No está demás decir que el domingo 01 de enero del 2012 pertenece a la última semana del 2011, y a pesar de que esta semana tiene un solo día en el periodo de observación, queda codificada como la primera semana del 2012.

El proceso descrito en la etapa 2 se ejecuta con los tres algoritmos: TPM, EDM y RoIs-FV.

El algoritmo TPM es ejecutado con cuatro configuraciones diferentes las cuales varían en el nivel de agregación de las unidades espaciales. Se utilizaron las unidades:

1. Identificador de parada
2. Identificador de zonificación 1
3. Identificador de zonificación 2
4. Identificador de cluster de transacciones (utilizando clusterización de las transacciones en radios de 500m)

El algoritmo EDM es ejecutado con solo una configuración debido al tiempo de ejecución del algoritmo. La configuración utilizada es la versión netamente espacial del algoritmo, es decir, con el parámetro c fijado en 0.

El algoritmo RoIs-FV se ejecutó utilizando la siguiente configuración:

1. Número de RoIs mínimo (min_{RoIs}): 2
2. Porcentaje mínimo de transacciones que deben agrupar las RoIs ($porcentaje\_visita$): 70
3. Función de distancia ($f\mathrm{dist}$): manhattan
4. Variables descriptivas del vector de características: La Tabla 4.3 muestra las variables seleccionadas[3].

[3]Se utilizaron las variables del conjunto de características expuesto en la Sección 3.3.1 que fuesen compatibles con los datos de la base de datos de Gatineau.

Tabla 4.3: Variables descriptivas que componen el vector de características usado con la base de datos de Gatineau.

Tipo de Característica	Característica
Temporal	Hora de inicio promedio primer viaje
	Hora de inicio promedio último viaje
	Número de días con viajes
	Moda del número de viajes por día
	Frecuencia de la moda del número de viajes por día
	Promedio de viajes a la semana
Espacial	Distancia viajada
	Máxima distancia viajada promedio
	Mínima distancia viajada promedio
	Radio de giro
	Entroía temporalmente no correlacionada
	Entropía aleatoria
	Porcentaje de primeras paradas diferentes
	Porcentaje de últimas paradas diferentes
Demográfica	Tipo de tarjeta
Actividad	Promedio de tiempo de actividad más corta del día
	Promedio de tiempo de actividad más larga del día
Modo de transporte	Número de etapas por viaje más frecuente

Ejecución etapa 3

La medición de cuan identificables son los usuarios se llevó a cabo con la base de datos de Transantiago. Se utilizaron dos cortes temporales de una semana: una semana de Abril del 2013 y una semana de septiembre del 2013. Como se señaló en la Preparación de los datos, se seleccionaron los registros de tarjetas presentes en ambas semanas y que tuviesen más de 10 transacciones semanales. Del grupo resultante se tomó una muestra de 5.000 tarjetas para llevar a cabo el experimento.

El proceso descrito en la etapa 3 se ejecuta con los tres algoritmos: TPM, EDM y RoIs-FV.

El algoritmo TPM es ejecutado con tres configuraciones diferentes las cuales varían en el nivel de agregación de las unidades espaciales. Se utilizaron las unidades:

1. Identificador de parada
2. Identificador de zonificación 1

El algoritmo EDM es ejecutado con solo una configuración debido a la complejidad del algoritmo, esta es utilizando la versión netamente espacial del algoritmo, es decir, con el parámetro c fijado en 0.

El algoritmo RoIs-FV se ejecutó utilizando numerosas configuraciones variando los siguientes parámetros:

1. el número de RoIs mínimo (min_RoIs)
2. el porcentaje mínimo de transacciones que deben agrupar las RoIs ($porcentaje\_visita$)
3. la función de distancia ($fdist$), se utilizaron las distancias descritas en la Sección 3.3
4. las variables descriptivas del vector de características. Se utilizaron diferentes combinaciones de las características señaladas en la Sección 3.3.1. Finalmente se seleccionan dos conjuntos: el conjunto de todas las características y el conjunto mínimo de características que logra mejor resultados.

4.3.5. Etapa 5: Postprocesamiento

Para el análisis de algunos escenarios se requiere detectar valores fuera de rango y normalizar los datos para que sean comparables, en particular al comparar resultados de diferentes algoritmos.

Se utilizaron dos métodos de detección de valores fuera de rango:

- Z-score modificado (Iglewicz y Hoaglin, 1993)
- IQR (Tukey, 1977)

En general, en cada caso se seleccionó el método que minimizara el número de *outliers* y obtuviese una distribución balanceada. Para todos los valores fuera de rango se utilizó una máscara con el valor máximo. Posteriormente se procede a realizar una normalización *min-max*.

4.4. Evaluación

4.4.1. Evaluación etapa 2: Evaluar la estabilidad de los perfiles de movilidad en el tiempo

Para evaluar la estabilidad de la caracterización de la movilidad, se calculan indicadores sobre las distintas variaciones almacenadas en los objetos asociados a cada usuario. Los siguientes son los indicadores utilizados:

1. Promedio
$$\frac{1}{n}\sum_{\mathrm{i}=1}^{n} X_{\mathrm{i}}$$
2. Rango
$$max_{\mathrm{i}=0\ldots n}(X) - min_{\mathrm{i}=0\ldots n}(X)$$
3. Rango interquartil: Luego de dividir los valores en quartiles, se calcula la diferencia entre el valor mínimo del cuarto quartil y el valor máximo del primer quartil.

4. Varianza:

$$\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(X_{\mathrm{i}}-\bar{X})^2$$

5. Desviación estándar

$$\sqrt{\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(X_{\mathrm{i}}-\bar{X})^2}$$

6. Coeficiente de Variación: Razón entre el Promedio y la Desviación estándar.

Es necesario aclarar que la estabilidad que miden los algoritmos TPM, EDM y RoIs-FV se encargan de medir la similitud o diferencia (dependiendo del algoritmo) entre las distintas ventanas de movilidad de un usuario. En consecuencia, el indicador Promedio es un estimador de la medida de similitud de la movilidad en el periodo de actividad del usuario. En cambio, los otros indicadores (Rango, Rango IQ, Varianza, etc.), miden como varía aquella similitud capturada por el promedio. Por ejemplo, un usuario puede presentar similitud TPM 0,2 entre sus ventanas de manera constante, lo cual es baja similitud. Luego, el Promedio de su variación de la movilidad indicará que presenta baja similitud (0,2). Sin embargo, su Rango será pequeño (0, en este caso), ya que fue constantemente variable.

Los indicadores señalados anteriormente se utilizan para analizar el comportamiento de los usuarios bajo las diferentes configuraciones utilizadas, descritas en la Sección 4.3.4. En general se utilizaron dos enfoques: El primer enfoque consiste en comparar los resultados según la proximidad de las ventanas, i.e. comparar las Variaciones, las Variaciones de ventanas cercanas y las Variaciones de ventanas contiguas de cada usuario. El segundo enfoque consiste en comparar los resultados obtenidos al variar el tamaño de las ventanas entre una, dos y cuatro semanas.

4.4.2. Evaluación etapa 3: Evaluar la identificabilidad de los perfiles de movilidad

Para evaluar cuan identificables son los perfiles de movilidad se calcula la tasa de identificación, la tasa de error y la tasa de abstención a partir de las matriz M que almacena las similitudes de las comparaciones entre los perfiles del primer corte temporal con los del segundo corte temporal.

La tasa de identificación corresponde al porcentaje de tarjetas que fueron emparejadas consigo mismas. Se obtiene contando el número de filas de la matriz M en que el índice de la fila coincide con el índice de la columna con mayor similitud.

La tasa de error corresponde al porcentaje de tarjetas que fueron emparejadas con tarjetas distintas a sí mismas. Se obtiene contando el número de filas de la matriz M en que el índice de la fila no coincide con el índice de la columna con mayor similitud.

La tasa de abstención corresponde al porcentaje de tarjetas que por falta de información o por falta de similitud mínima no fueron emparejadas con ninguna. Se obtiene contando el

número de filas de la matriz M en que la columna con mayor similitud posee un valor fuera de rango.

Es importante notar que para el caso de Santiago, debido a la anonimidad de la tarjeta Bip!, no hay seguridad de que las tarjetas en periodos temporalmente independientes hayan sido usadas por un mismo usuario. Por lo anterior, para el porcentaje incierto de tarjetas que cambian de usuario portador hay dos casos posibles:

1. Caso 1: no son identificadas correctamente ya que existe otra tarjeta con movilidad más similar. En este caso las tarjetas pasan a formar parte de la tasa de error (a pesar de efectivamente corresponder a usuarios diferentes).
2. Caso 2: son identificadas correctamente a pesar de que hubo un cambio de usuario. En este caso las tarjetas pasan a formar parte de la tasa de identificación, lo cual es un error ya que no corresponde al mismo usuario.

Si solo se considera el Caso 1, entonces la tasa de identificación obtenida funciona como una cota mínima, ya que habrían tarjetas que no pueden ser reconocidas debido al cambio de movilidad que implica el cambio de portador de la tarjeta.

Por su parte, el Caso 2 es un caso límite e ineludible del problema que aborda esta tesis. Si una tarjeta tiene el mismo identificador en dos periodos, y además se mueve extremadamente similar, no es posible advertir un cambio de portador. Esta situación es similar al caso de dos tarjetas diferentes con movilidad exactamente igual, donde el identificador de la tarjeta es la única forma de distinguir la movilidad registrada.

Capítulo 5

Reconocimiento de usuarios de Transantiago

En este capítulo se presentan los resultados de los algoritmos TPM, EDM y RoIs-FV, obtenidos al medir la factibilidad de emparejar tarjetas con el mismo identificador entre dos cortes temporales. Los resultados presentados en este capítulo corresponden al experimento explicado en la Sección 4.3.3 evaluados según las tasas de identificación, de error y abstención presentadas en la Sección 4.4.2. Se utilizó una muestra de la base de datos de Transantiago, la cual consiste en dos semanas de transacciones de 5.000 usuarios, una semana de Abril del 2013 y otra semana de Septiembre del mismo año.

Las tres primeras secciones de este capítulo describen los resultados asociados a cada algoritmo bajo diferentes experimentos. La última sección corresponde a un análisis general de los resultados obtenidos por los tres algoritmos.

5.1. Algoritmo TPM

5.1.1. Análisis del rendimiento del algoritmo TPM variando la agregación espacial

El algoritmo TPM caracteriza la movilidad de los usuarios con una matriz TPM, compuesta por la probabilidad que tiene un usuario de viajar entre los diferentes orígenes y destinos visitados en un periodo de tiempo determinado. En esta sección se presenta el efecto de cambiar la resolución espacial de los orígenes y destinos, utilizando dos niveles de agregación: las paradas de buses y estaciones de metro (≈11.000 posiciones posibles), y una zonificación que divide a Santiago en 795 zonas según el uso de suelo.

La Figura 5.1 muestra la intersección de las tarjetas identificadas con el algoritmo TPM con los dos niveles de agregación espacial. En la figura se muestra el tamaño de los conjuntos de tarjetas identificadas con zonificación y con paradas de buses, y el tamaño de los

subconjuntos de los elementos comunes y disjuntos. Se observa que usar datos menos agregados mejora levemente el rendimiento de este algoritmo, ya que la tasa de identificación utilizando paradas es un 2.7 % mayor que utilizando las zonas. La Figura 5.1 muestra que la diferencia del rendimiento no se produce simplemente porque una configuración identifique las mismas y más tarjetas que otra, sino que hay una mayoría de tarjetas identificadas por ambas configuraciones, y dos grupos de tamaño similar que solo pueden ser identificados al utilizar paradas o zonas independientemente. Lo anterior se puede explicar por la existencia de dos tipos de usuarios: Por un lado, personas que usan paradas diferentes pero cercanas para realizar un mismo viaje, por lo cual no es fácil identificarlos al utilizar paradas, pero si con zonas. Por otro lado, personas que viajan entre zonas populares, por tanto la única forma de distinguirlas es utilizando un mayor nivel de desagregación espacial.

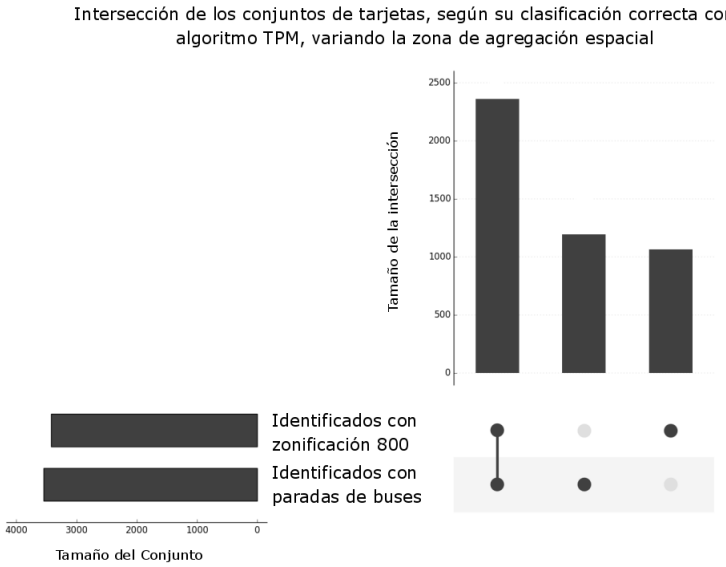


Figura 5.1: Intersección de los conjuntos de tarjetas identificadas correctamente con el algoritmo TPM utilizando distintos niveles de agregación espacial.

5.1.2. Análisis de la mejor configuración del algoritmo TPM

En esta sección analizan las tasas de identificación y de error del algoritmo TPM utilizando como nivel de agregación espacial las paradas de buses y estaciones de metro.

El algoritmo TPM reconoció un 66,72 % de la muestra de usuarios de Transantiago. Este rendimiento es aproximadamente un 15 % menor al reportado por De Mulder et al. sobre datos de llamadas telefónicas. Esta diferencia podía deberse al número de usuarios comparados (De Mulder et al. utilizaron un *dataset* de 100 usuarios), por tanto se llevó a cabo el ejercicio de obtener 1.000 grupos aleatorios de 100 usuarios de Transantiago. El rendimiento promedio de los 1.000 grupos fue de 85 %, es decir un 5 % más que al utilizar datos de telefonía. Este aumento en el rendimiento obtenido con los datos de Transantiago se puede explicar por el tipo de usuarios seleccionados en la muestra. El experimento de De Mulder et al. fue aplicado sobre estudiantes de un campus universitario, por tanto eran usuarios con patrones

de movilidad en común. En cambio, los datos de Transantiago corresponden a una muestra aleatoria de usuarios, cuyos viajes están distribuidos a través de toda la red de transporte público y no necesariamente tienen intersecciones. Incluso considerando lo anterior, resulta relevante notar que si bien los datos de transporte público son menos frecuentes que los datos de telefonía, registran información equivalente para caracterizar la movilidad de los usuarios.

La Figura 5.2a muestra el número de tarjetas reconocidas y no reconocidas según el indicador de similitud de la movilidad de cada tarjeta en los dos cortes temporales. Es posible observar que los usuarios identificados correctamente se distribuyen en todo el espectro del indicador de similitud. En cambio, los usuarios no identificados se concentran en el extremo izquierdo del espectro, con un máximo cuando la similitud es 0. Lo anterior señala que hay un 10 % de usuarios donde la movilidad de abril no comparte ningún viaje con la movilidad de Septiembre. La distribución relativamente homogénea de los usuarios identificados permite concluir que este algoritmo es altamente sensible a las diferencias de la movilidad, haciendo que un mismo usuario no sea tan similar consigo mismo, y al mismo tiempo diferenciándolo aún más del resto.

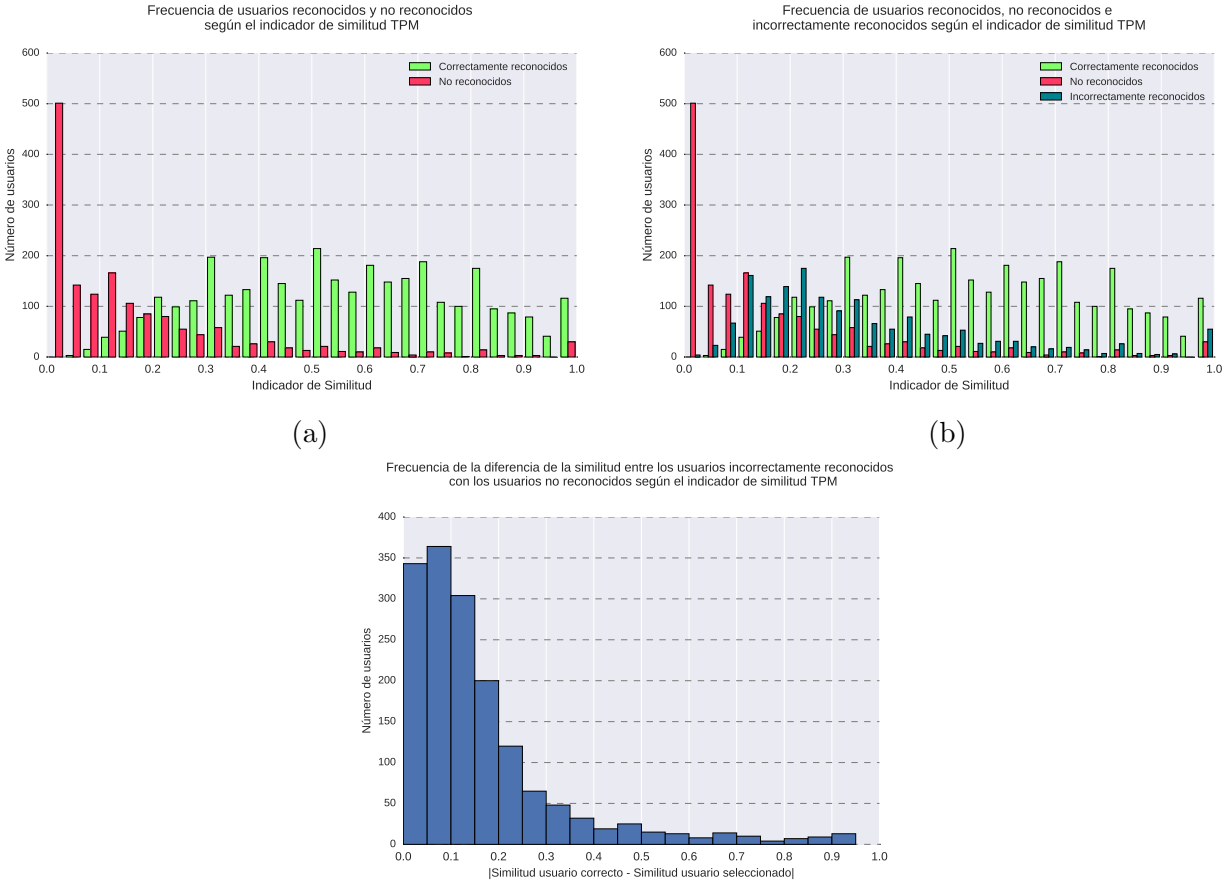


Figura 5.2: Resultados del reconocimiento de usuarios utilizando el algoritmo TPM sobre la base de datos de Transantiago.

La Figura 5.2b muestra la frecuencia del indicador de similitud de los usuarios reconoci-

dos, incorrectamente reconocidos y no reconocidos. Es decir, muestra los mismos datos de la Figura 5.2a más la frecuencia de la similitud de los usuarios incorrectamente reconocidos. Los usuarios incorrectamente reconocidos corresponden a aquellos usuarios cuya similitud fue mayor que la del usuario no reconocido. En la Figura 5.2b se ve que la distribución de los usuarios incorrectamente reconocidos se concentra principalmente en el rango de similitud [0.0,0.5], lo cual indica baja similitud. Lo anterior sugiere que sería posible establecer un umbral mínimo de similitud, y así evitar la mayoría de los usuarios incorrectamente reconocidos.

La Figura 5.2c presenta el histograma de la distancia entre los indicadores de similitud del usuario no reconocido y el usuario incorrectamente reconocido. Es decir, muestra cuánto más se parece el usuario incorrectamente reconocido que el usuario que debía ser reconocido. Se observa que la frecuencia de la distancia es decreciente con un máximo claro entre 0,0 y 0,05. Es decir, un número importante de usuarios no reconocidos fue "vencido"por otro usuario con una similitud muy cercana. En vista de lo observado, en un escenario con toda la base de datos de Transantiago, resultaría beneficioso permitir que se emparejen usuarios que no tienen máxima similitud, pero que tengan una similitud muy cercana a la máxima y además compartan el identificador de tarjeta.

5.2. Algoritmo EDM

Como se mencionó en la Sección 3.2, el algoritmo EDM permite regular la importancia de las dimensiones espaciales y temporales al calcular la distancia entre dos trayectorias. Sin embargo, la complejidad cúbica de este algoritmo fue una gran limitación al momento de evaluar su rendimiento. El tiempo de ejecución incrementa rápidamente con el número de usuarios que se esté comparando, llegando a 433,6 horas al comparar la movilidad de los 5.000 usuarios en ambos cortes temporales. Finalmente, se optó por paralelizar las comparaciones y evaluar una versión netamente espacial del algoritmo EDM.

Sobre los resultados almacenados en la matriz M (matriz que almacena la distancia entre los distintos usuarios), se procede a detectar *outliers* con el método Z-score modificado, el cual indica que los valores mayores a 24.942 serían valores atípicos. Entonces, se procede a enmascarar todos los valores atípicos con el valor límite 24.942, para luego aplicar una normalización minmax sobre las distancias obtenidas.

Utilizando solo la dimensión espacial de la trayectoria, y con los datos normalizados, este algoritmo reconoce a un 40,01 % de los usuarios.

La Figura 5.3a muestra el número de tarjetas reconocidas y no reconocidas según el indicador de similitud de la movilidad de cada tarjeta en los dos cortes temporales. Ambos grupos se concentran en el extremo izquierdo del espectro de distancia, sin embargo los usuarios reconocidos correctamente se ubican en un rango más pequeño y más cercano al 0. La mayoría de los usuarios no reconocidos muestran una distancia menor a 0,2, y, a diferencia de los resultados del algoritmo TPM, no se observan concentraciones de usuarios que hayan cambiado drásticamente su movilidad.

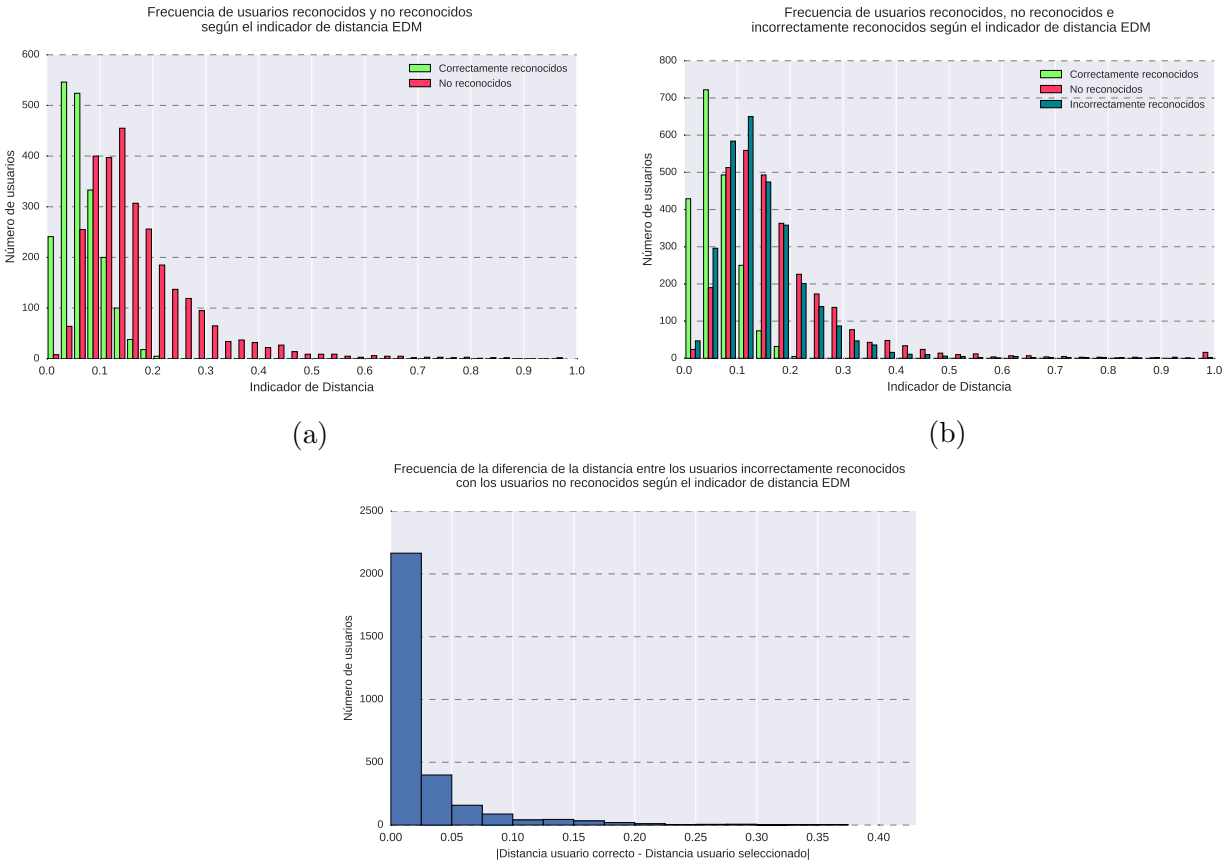


Figura 5.3: Resultados del reconocimiento de usuarios utilizando el algoritmo EDM sobre la base de datos de Transantiago.

La Figura 5.3b muestra la frecuencia del indicador de distancia de la movilidad de los usuarios reconocidos, incorrectamente reconocidos y no reconocidos. La distribución de los usuarios incorrectamente reconocidos es muy similar a la de los usuarios no reconocidos. De lo observado se concluye que este algoritmo tiende a sobrevalorar las similitudes entre las trayectorias, haciendo que un mismo usuario sea altamente parecido a sí mismo, y haciendo al mismo tiempo que se parezca a otros.

La Figura 5.3c muestra el histograma de la distancia entre los indicadores de similitud del usuario no reconocido y el usuario incorrectamente reconocido. Se observa que la frecuencia de la distancia se distribuye entre 0,0 y 0,2 principalmente y decrece rápidamente. El mismo fenómeno se observó con el algoritmo TPM, sin embargo con el algoritmo EDM es aun más marcada la concentración en el extremo izquierdo del espectro de distancia. En particular, el máximo entre 0,0 y 0,025 indica que más del 40 % de los usuarios no fue reconocido a causa de otro usuario con una distancia extremadamente similar.

El algoritmo EDM utiliza una medida basada en la distancia geográfica y permite calcular la distancia entre trayectorias sin coincidencias exactas. Estas características hacen que los usuarios tengan más usuarios similares, por tanto más difíciles de distinguir. Por ejemplo, dos usuarios que viajan entre los mismos orígenes y destinos geográficos, pueden parecer

altamente parecidos, aunque uno utilice metro y el otro bus. En cambio, con el algoritmo TPM si dos trayectorias no comparten ninguna parada la similitud es inmediatamente nula. Por lo anterior resulta comprensible que el rendimiento del algoritmo EDM sea inferior al del algoritmo TPM.

5.3. Algoritmo RoIs-FV

5.3.1. Análisis del rendimiento del algoritmo RoIs-FV variando parámetros

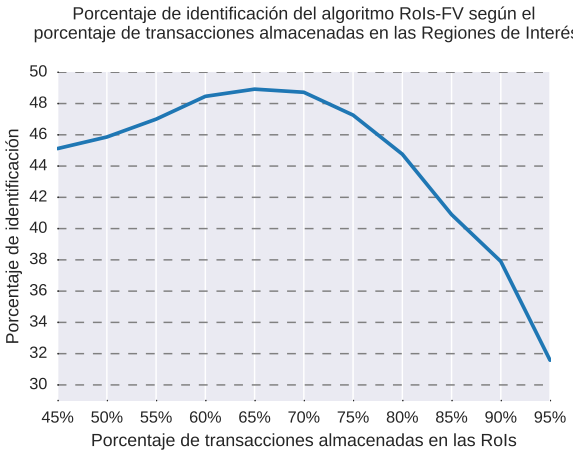
El algoritmo RoIs-FV fue diseñado como parte de esta tesis, por lo que se hicieron pruebas variando diferentes parámetros con el objetivo de maximizar el número de usuarios reconocidos. En esta sección se presenta los resultados de estas pruebas, que finalmente determinaron la mejor configuración.

Parámetros relativos a las RoIs

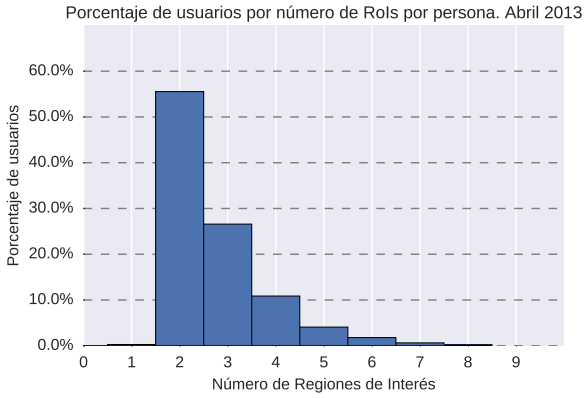
En primer lugar se definió el porcentaje mínimo de transacciones que debiesen almacenar las RoIs. La idea de extraer las RoIs es capturar al menos las dos ubicaciones principales de un usuario, probablemente cerca del hogar y del trabajo o lugar de estudio. Definir el porcentaje de transacciones mínimo determina el número de RoIs que se obtendrá por usuario. Si se selecciona un porcentaje pequeño, los usuarios tendrán menos RoIs que si se selecciona un porcentaje grande. Por un lado, obtener pocas RoIs puede no ser suficiente para caracterizar un usuario. Por otro lado, muchas RoIs pueden incluir zonas no tan importantes y aumentar la cantidad de usuarios con los que se comparara.

La Figura 5.4a muestra los resultados luego de ejecutar el algoritmo variando los valores del porcentaje de transacciones entre un 45 % y un 95 %. Se puede observar que el máximo se alcanza al utilizar un 70 % de las transacciones para obtener las RoIs de los usuarios. La Figura 5.4b presenta el número de RoIs por usuario utilizando un 70 % de las transacciones. De aquí se desprende que más de un 50 % de los usuarios realiza al menos un 70 % de sus transacciones en solo dos lugares importantes.

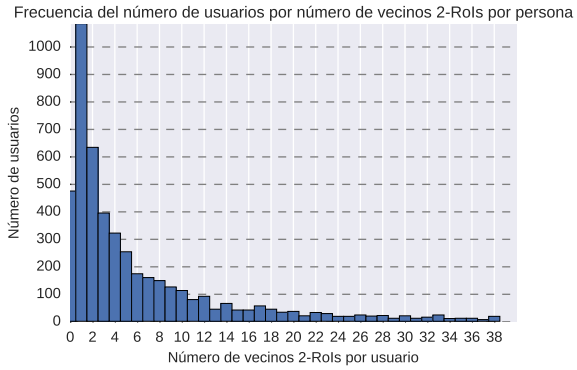
La Figura 5.4c muestra el número de vecinos de dos RoIs por usuario. Se consideran vecinos de dos RoIs a todo par de usuarios de distintos cortes temporales que comparte al menos dos RoIs. En este gráfico se observan dos fenómenos interesantes: en primer lugar, hay cerca de 500 usuarios que no comparten RoIs entre ambos cortes temporales, lo cual indica un cambio de comportamiento importante en casi un 10 % de los usuarios. En segundo lugar, se observa que la mayoría de los usuarios tiene menos de 6 vecinos de dos RoIs. Esto reduce notoriamente el número de comparaciones que hay que hacer, ya que cada usuario es comparado solo con sus vecinos de dos RoIs. Es importante notar que esta observación solo es válida para la muestra de 5.000 usuarios. Si el tamaño de la muestra aumenta, se espera encontrar más usuarios con vecinos de dos RoIs.



(a)



(b)



(c)

Figura 5.4: Características relativas a las RoIs de los usuarios de la base de datos de Transantiago.

Tabla 5.1: Conjunto mínimo de variables descriptivas utilizadas como vector de características en el algoritmo RoIs-FV.

Tipo de Característica	Característica
Temporal	Tiempo de inicio del primer viaje durante la semana laboral Tiempo de inicio del último viajo durante la semana laboral
Espacial	Promedio semanal de máxima distancia viajada Distancia viajada Entropía no correlacionada Radio de giro
Actividad	Promedio de la actividad más corta durante la semana laboral
Demográfica	Tipo de tarjeta
Modo de transporte	Porcentaje de viajes en bus

Como hay usuarios que comparten solo una Región de Interés, se analizó el rendimiento de este algoritmo variando la condición de número mínimo de RoIs compartidos en los periodos analizados. Al utilizar la condición de compartir una RoI el algoritmo reconoce solo a un 18,4 % de los usuarios. Bajo esta condición, todos los usuarios tienen una pareja potencial, i.e. hay al menos una tarjeta del segundo periodo con una RoI en común para todas las tarjetas del primer periodo. Por lo tanto, este método no se abstendrá en níngun caso. Sin embargo, solo un 95,51 % de las tarjetas comparten una RoI entre ambos periodos. Es decir, un 4,49 % de las tarjetas fueron emparejadas con un "impostor"por defecto.

Al utilizar como condición un mínimo de dos RoIs compartidas, este método reconoce un 44,34 % de los usuarios. Si bien el rendimiento mejora, el efecto del impostor aumenta. Se observa que un 90,52 % de los usuarios tienen una pareja potencial, es decir el método se abstendrá en un 9,48 % de las tarjetas. No obstante, solo un 65,85 % de las tarjetas comparten dos RoIs en dos periodos. Esto significa que en el 25 % de los casos este algoritmo no considerará la tarjeta correcta como posible *match*. Al seguir aumentando el número de RoIs mínimos compartidos se observó una disminución del rendimiento.

Parámetros relativos al vector de características

Se utilizó fuerza bruta para encontrar el conjunto mínimo de características que optimizara el rendimiento del algoritmo RoIs-FV. Se seleccionaron las nueve características presentadas en la Tabla 5.1, y al ejecutar el algoritmo con distancia Euclidiana, se obtuvo una tasa de reconocimiento de 50,56 %, en comparación con el 44,34 % obtenido al usar todas las características.

Por último, se evaluó el rendimiento variando la distancia utilizada para medir la disimilitud entre los vectores de características de los usuarios a comparar. Se utilizaron las distancias descritas en la Sección 3.3.2. Por cada distancia evaluada se calculó la tasa de reconocimiento, la tasa de abstención, la tasa de error y la razón entre la tasa de identificación y la tasa de error. Cada distancia fue evaluada en tres escenarios:

1. Utilizando todas las características disponibles del vector de características

2. Utilizando solo el conjunto mínimo de características
3. Utilizando el mejor de los escenarios anteriores, junto con un umbral óptimo de máxima distancia

El umbral óptimo de máxima distancia se definió como una forma de contrarrestar la falta de abstención en los casos donde las tarjetas no comparten 2 RoIs y un impostor es identificado. Con el umbral se busca definir un mínimo de similitud entre dos vectores. Luego, dos vectores se consideran como posible *match* solo si la distancia entre ellos es menor que el umbral. Se selecciona como umbral óptimo la distancia que maximice la razón entre la tasa de reconocimiento y la tasa de error.

Los resultados de cada distancia evaluada se presentan en la Tabla 5.2. Se agregan como punto base dos experimentos que no utilizan la distancia entre los vectores, sino que usan solo los RoIs. En el primer experimento, se emparejan tarjetas que tienen solo un *match* posible, i.e. se selecciona un *match* solo cuando una tarjeta tiene un solo vecino de dos RoIs. El segundo experimento consiste en emparejar las tarjetas de manera aleatoria entre todos los vecinos de segundo orden.

De la Tabla 5.2 se desprende que la distancia entre vectores de características efectivamente mejora la tasa de identificación. Si bien al emparejar usuarios con un solo *match* posible se logra una buena tasa de reconocimiento en comparación a la tasa de error, utilizando la distancia Bray-Curtis y un umbral 0,05 se logran mejores resultados.

En relación a los mejores resultados presentados en la Tabla 5.2, se observa que la distancia con mejor tasa de reconocimiento es la distancia Manhattan con el conjunto mínimo de variables, alcanzando un 51,94 %. La distancia con mejor razón entre reconocimiento y error es la distancia Canberra con el conjunto mínimo de variables, con una tasa de identificación de un 5,04 % y una tasa de error de un 0,38 %, y le sigue la distancia Manhattan con una tasa de identificación del 6,56 % y una tasa de error del 0,6 %.

En relación a los resultados generales de la Tabla 5.2, el conjunto mínimo de variables obtiene mejores resultados en la mayoría de las distancias, excepto con la distancia Hamming. En todos los experimentos el umbral óptimo de distancia máxima mejoró la razón entre tasa de identificación y error considerablemente.

Finalmente se concluye que la configuración con mejor rendimiento dependerá de cuánto valor se le asigna a la tasa de identificación versus la tasa de error.

Tabla 5.2: Rendimiento del algoritmo RoIs-FV cambiando la medida de distancia.

Distancia	**Experimento**	**Identificación (I)**	**Abstención**	**Error (E)**	**I/E**
Sin distancia	Un solo match de 2 RoIs	16,72	78,36	4,92	3,40
	Match aleatorio entre vecinos de 2 RoIs	26,92	9,22	63,86	0,42
Euclidiana	Todas las variables	46,38	9,22	46,18	1,00
	Conjunto mínimo de variables	51,32	9,22	39,96	1,28
	Umbral óptimo: 0,15	10,74	87,90	1,36	7,90
Manhattan	Todas las variables	48,72	9,22	42,06	1,16
	Conjunto mínimo de variables	52,76	9,22	38,02	1,39
	Umbral óptimo: 0,25	6,56	92,84	0,60	15,00
Bray-Curtis	Todas las variables	48,6	9,22	42,18	1,15
	Conjunto mínimo de variables	52,38	9,22	38,4	1,36
	Umbral óptimo: 0,05	20,12	77,66	2,22	9,06
Chebyshev	Todas las variables	40,52	9,22	50,26	0,81
	Conjunto mínimo de variables	48,34	9,22	42,44	1,14
	Umbral óptimo: 0,10	9,08	89,18	1,74	5,22
Canberra	Todas las variables	47,00	9,22	43,78	1,07
	Conjunto mínimo de variables	50,66	9,22	40,12	1,26
	Umbral óptimo: 0,30	7,16	92,54	0,30	23,87
Hamming	Todas las variables	48,24	9,22	42,54	1,13
	Conjunto mínimo de variables	48,06	9,22	42,72	1,13
	Umbral óptimo: 0,05	23,08	66,40	10,52	2,19

5.3.2. Análisis de la mejor configuración de RoIs-FV

En esta sección se analizan las tasas de identificación y de error del algoritmo RoIs-FV utilizando la siguiente configuración:

- Porcentaje mínimo de transacciones para extraer las RoIs: 70 %
- Número mínimo de RoIs compartidas: 2
- Características: Conjunto mínimo de características presentado en la Tabla 5.1
- Función de distancia: Manhattan

Luego se procede a diferenciar los valores atípicos para preparar los datos para la normalización. Utilizando el algoritmo IQR sobre todas las distancias almacenadas en la matriz M, se obtiene que los valores mayores a 3,80 se consideran *outliers*. Luego se procede a enmascarar todos los valores atípicos con el valor 3,8, y por último se utiliza una normalización min-max sobre todos los valores de la matriz M.

El algoritmo RoIs-FV en su mejor configuración reconoció un 51,94 % de la muestra de usuarios de Transantiago con una tasa de abstención del 9,48 %. Esta tasa de identifiación es aproximadamente un 15 % menor que el rendimiento obtenido con el algoritmo TPM y un 10 % mayor que la tasa del algoritmo EDM.

La Figura 5.5a muestra el número de tarjetas reconocidas y no reconocidas según el indicador de distancia de la movilidad de cada tarjeta en los dos cortes temporales. De este gráfico llama la atención el alto número de usuarios no reconocidos con distancia 1. Estos usuarios corresponden al 25 % de usuarios que no comparten 2 RoIs entre abril y septiembre, y que tienen un "impostor"que si los comparte. También se observa que los usuarios reconocidos correctamente se distribuyen en el rango [0,0-0,5], con mediana en 0,1. Los usuarios no reconocidos con distancia distinta de 1, si bien tienen distancia mayor a 0,1, también se encuentran entre valores menores a 0,5. Por tanto, todos los usuarios que comparten dos RoIs mantienen una movilidad similar, independiente de si fueron reconocidos o no.

La Figura 5.5b muestra la frecuencia del indicador de distancia de los usuarios reconocidos, incorrectamente reconocidos y no reconocidos. La distribución de los usuarios incorrectamente reconocidos se concentra en el rango [0.0-0.5], similar a los usuarios no reconocidos con distancia distinta de 1. Esto sugiere que al igual que el algoritmo EDM, el algoritmo RoIs-FV homogeniza las diferencias de la movilidad, haciendo que muchos usuarios sean similares.

La Figura 5.5c presenta el histograma de la distancia entre los indicadores de distancia entre el usuario no reconocido y el usuario incorrectamente reconocido. Se observa que la frecuencia de la distancia tiene dos máximos locales. El máximo en el rango [0,0-0,05] corresponde a usuarios no reconocidos que fueron "vencidos"por otro usuario con una similitud muy cercana. El segundo máximo en el rango [0,8-0,85] corresponde a los usuarios que no comparten 2 RoIs y que fueron "vencidos"por algún usuario que si compartía los 2 RoIs. Es importante recordar que se asignó valor 1 a la distancia de los usuarios que no compartían 2 RoIs, sin embargo este valor es arbitrario y solo pretende graficar baja similitud. Por lo anterior, de la Figura 5.5c solo se puede advertir la presencia de los dos grupos de usuarios no reconocidos, pero no relacionarlos cuantitativamente.

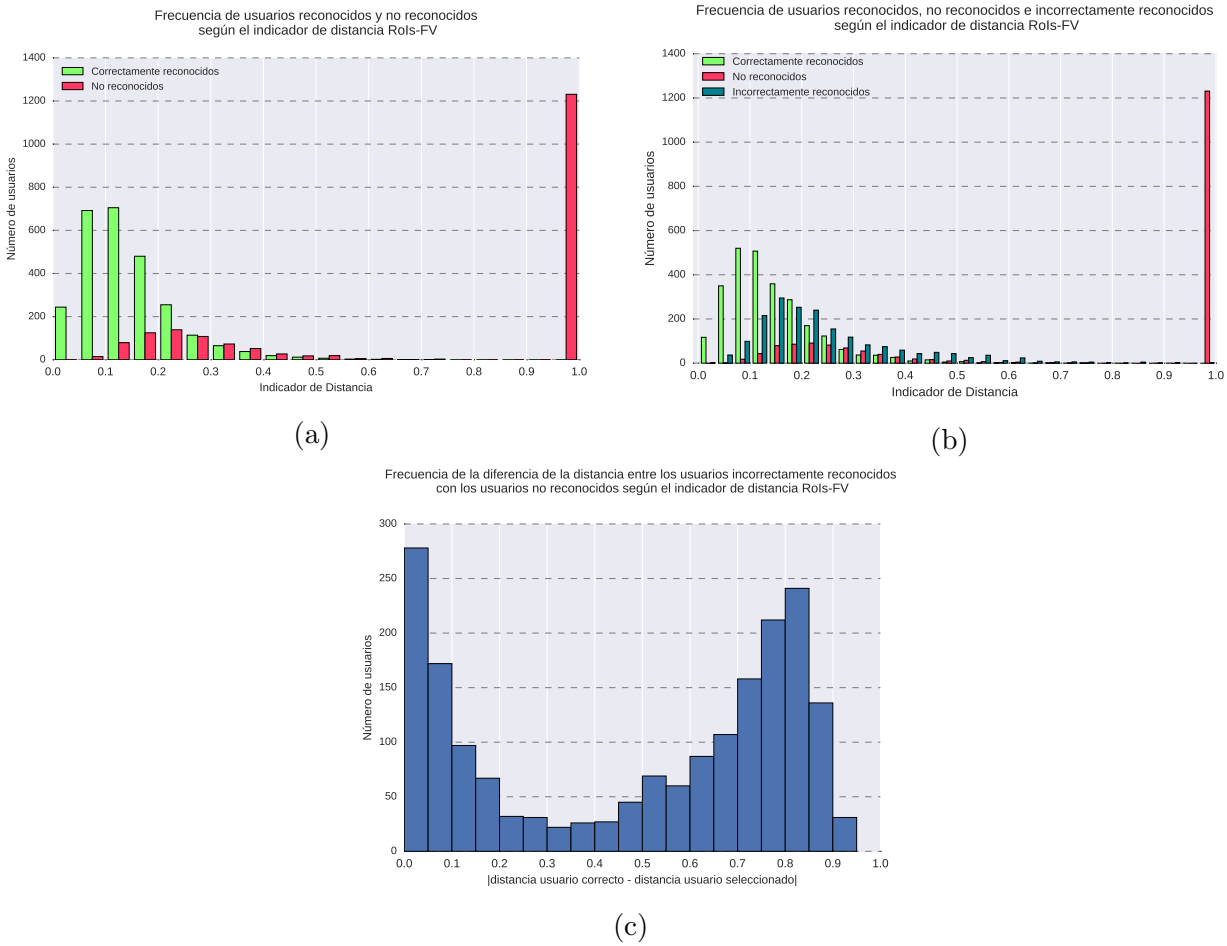


Figura 5.5: Resultados del reconocimiento de usuarios utilizando el algoritmo RoIs-FV sobre la base de datos de Transantiago.

5.4. Resultados generales

Luego de analizar los resultados de cada algoritmo, en esta sección se procede a comparar los resultados de los tres algoritmos en conjunto.

La Figura 5.6 muestra la intersección de los conjuntos de usuarios reconocidos correctamente e incorrectamente por los tres algoritmos ordenados de mayor a menor tamaño. En primer lugar es interesante notar el orden de los conjuntos en general: el algoritmo con mayor número de usuarios reconocidos correctamente es el algoritmo TPM, le siguen el algoritmo RoIs-FV y luego el algoritmo EDM. Los no reconocidos del algoritmo EDM son incluso más que los reconocidos correctamente por el algoritmo RoIs-FV.

En segundo lugar, de la Figura 5.6 se desprende que la mayor intersección corresponde a la de los usuarios correctamente reconocidos con los tres algoritmos. Le sigue el conjunto de usuarios no reconocidos por los tres algoritmos. De esto se obtiene que para casi el 50 % de los usuarios el resultado es el mismo según los tres algoritmos. Lo anterior es interesante, ya que señala la presencia de usuarios cuya movilidad es suficientemente estable y única bajo tres medidas diferentes para ser distinguible del resto de los usuarios. De la misma forma, el

gráfico señala la presencia de usuarios cuya movilidad es confundida con la de otros usuarios por los tres algoritmos, lo cual indica la presencia de usuarios cuyo cambio de comportamiento y similitud con otros los hacen indistinguibles. Lo anterior permite cuantificar el porcentaje de usuarios con cambios de comportamiento drásticos. En particular, un 78 % de los 932 usuarios que no fueron reconocidos por ningún algoritmo no compartían el mínimo de dos RoIs entre los dos periodos.

En tercer y último lugar, la Figura 5.6 muestra que para el 50 % restante de usuarios en que los algoritmos no coinciden, se dividen en diez subconjuntos con distinta frecuencia. Lo anterior evidencia la existencia de usuarios que fueron reconocidos por algunos algoritmos y no por otros, en particular todos los algoritmos reconocen cierto número de usuarios que ninguno de los otros dos algoritmos logra reconocer. Ciertamente se observa en mayor medida para el algoritmo TPM, que reconoce 600 usuarios que con los otros algoritmos se emparejaron incorrectamente. De lo anterior se concluye que diferentes usuarios pueden tener diferentes formas de ser distinguibles, y también, que la mayoría de los usuarios que pudieron ser reconocidos, fueron reconocidos con el algoritmo TPM.

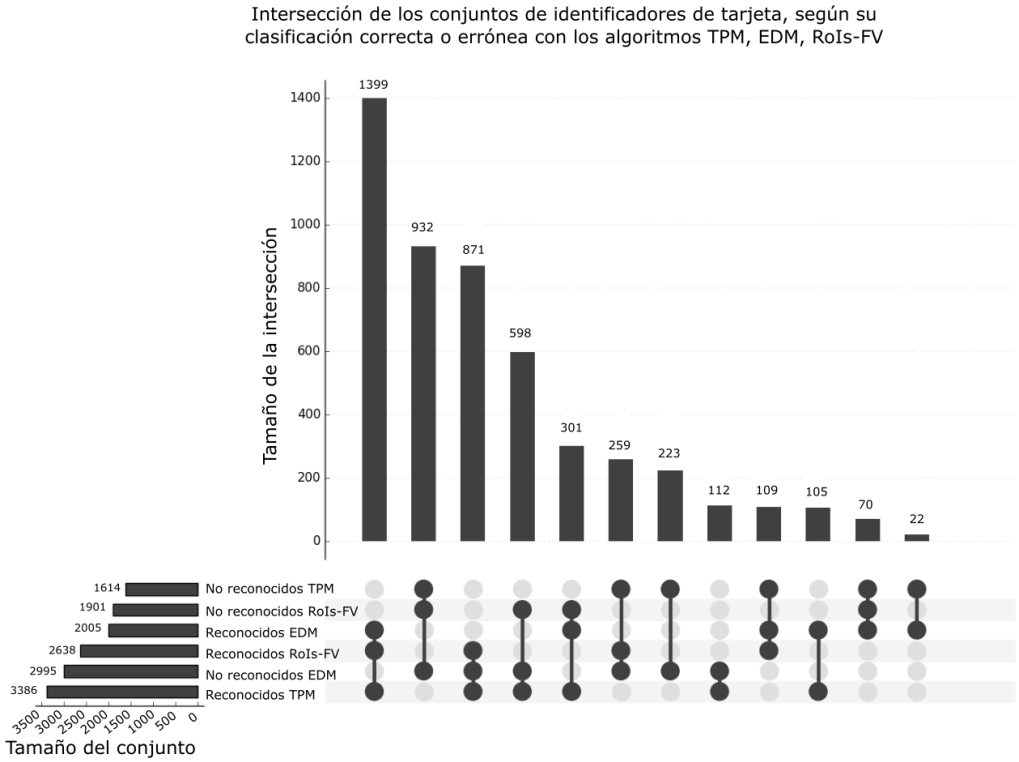


Figura 5.6: Intersección de los conjuntos de tarjetas identificadas y no identificadas con los algoritmos TPM, EDM y RoIs-FV.

Como el algoritmo TPM tuvo un rendimiento superior al algoritmo RoIs-FV y EDM, se decidió unir sus resultados con algunas variables del algoritmo RoIs-FV que permitieran profundizar el análisis de los usuarios correcta e incorrectamente reconocidos. La Figura 5.7a muestra la frecuencia de usuarios reconocidos y no reconocidos por el algoritmo TPM según el número de RoIs compartidas por los usuarios entre los dos cortes temporales. Se puede observar que la mayoría de los usuarios que no comparten 2 RoIs fue incorrectamente

identificado. Sin embargo también se observa que más de 500 usuarios que solo comparten una Región de Interés fueron reconocidos correctamente por el algoritmo TPM. Lo anterior permite concluir por una parte que exigir un mínimo de RoIs compartidas mejoraría la razón entre usuarios identificados y no identificados del algoritmo TPM. Por otra parte, el método de exigir 2 RoIs compartidos está siendo muy exigente en ciertos casos, lo cual se puede deber al límite de caminata de 500 metros. Una posible solución a este problema sería aumentar el radio de caminata con el cual se extraen los RoIs o aumentar la distancia que define si dos RoIs son compartidos o no.

La Figura 5.7b muestra el número de usuarios reconocidos y no reconocidos por el algoritmo TPM, según el número de vecinos de 2 RoIs. De esta figura se desprende que los usuarios que no comparten 2 RoIs son reconocidos y no reconocidos con casi la misma frecuencia. Además, los usuarios que tienen al menos un vecino de 2 RoIs tienen una frecuencia decreciente con un máximo notorio en los usuarios que tienen solo un potencial *match*. Finalmente se observa que la razón entre reconocidos y no reconocidos disminuye al aumentar el número de vecinos, por lo que se concluye que la capacidad de reconocer a los usuarios disminuye al aumentar la popularidad de los lugares de interés.

La Figura 5.7c muestra la frecuencia de los usuarios reconocidos y no reconocidos según la preferencia de modo de transporte. Se definió a los usuarios de buses como aquellos usuarios que realizan al menos un 80 % de sus viajes en bus. Del mismo modo se definió a los usuarios de metro como aquellos usuarios que realizan al menos un 80 % de viajes en metro. La Figura 5.7c muestra claramente que la razón entre usuarios reconocidos y no reconocidos es mucho menor en los usuarios de metro. Este fenómeno se justifica por la popularidad de las estaciones de metro y porque la cantidad de estaciones de metro es mucho menor que la cantidad de paradas de buses; ambas características hacen que sea más difícil distinguir a los usuarios de metro.

Las imágenes de la Figura 5.8 muestran la similitud de los usuarios en los dos periodos observados en dos visualizaciones. La Figura 5.8a muestra la silueta de la distribución de los tres algoritmos según el indicador de similitud (TPM) o distancia (RoIs-FV y EDM) de la movilidad de cada usuario en los cortes temporales Abril 2013 y Septiembre 2013. La Figura 5.8b muestra el diagrama de caja de las distribuciones de los indicadores de la similitud o distancia de la movilidad de los usuarios en ambos cortes temporales.

En la Figura 5.8a se puede observar claramente la diferencia entre las distribuciones de los tres algoritmos. La principal diferencia es que el algoritmo TPM se distribuye a lo largo de todo el espectro de similitud, en cambio los algoritmos EDM y RoIs-FV se encuentran concentrados principalmente en el rango [0,0-0,5]. La forma de las distribuciones también es diferente entre el indicador de similitud del algoritmo TPM y los algoritmos EDM y RoIs-FV. El máximo del indicador de similitud TPM está en similitud 0, en cambio los máximos de los indicadores de distancia de los otros algoritmos se encuentran cercanos a la distancia mínima.

Por su parte, la Figura 5.8b muestra que para el algoritmo TPM la mediana está en 0,4, esto indica que más de la mitad de los usuarios poseen una movilidad significativamente diferente en los dos periodos según este algoritmo. En cambio, las medianas de los indicadores de distancia de los algoritmos EDM y RoIs-FV son menores a 0,2, lo cual señala que la mayoría de los usuarios se mueve bastante similar según estos algoritmos.

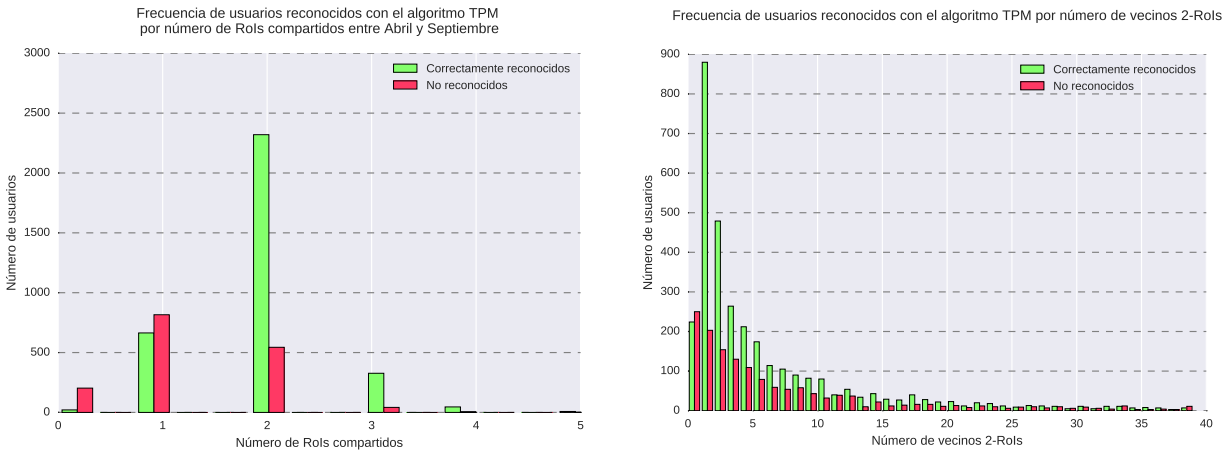


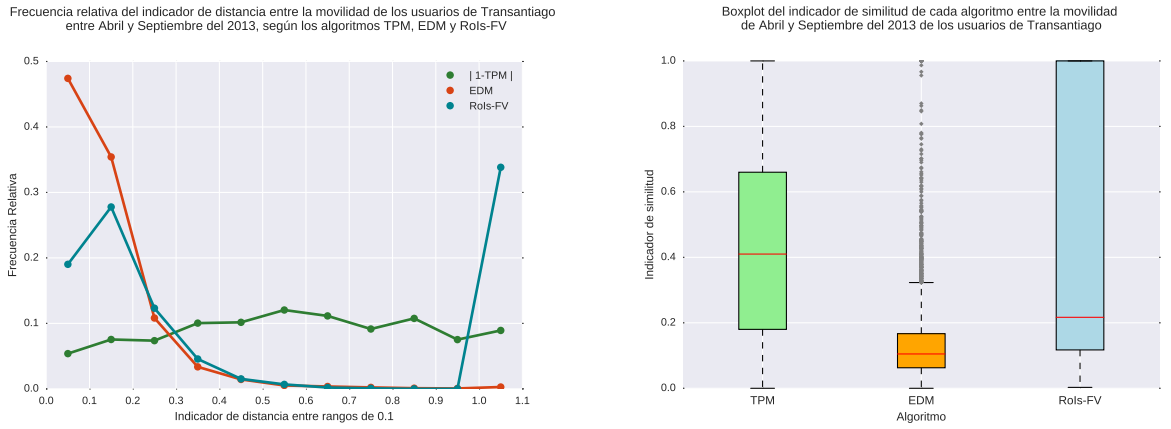
Figura 5.7: Resultados del algoritmo TPM según diferentes características de los usuarios.

Las observaciones anteriores refuerzan la idea de que la variabilidad del movimiento de los usuarios depende en gran medida del tipo de medición que se realice. También, considerando las propiedades de los distintos algoritmos, se concluye que la variabilidad medida es mayor al utilizar distancias espaciales basadas en el nombre de las paradas que al utilizar distancias geográficas. Por otra parte, se observó que las distancias geográficas y los métodos basados en características de la movilidad utilizados (comúnmente en algoritmos de *clustering*) tienden a homogeneizar a los usuarios con comportamientos similares.

La Tabla 5.3 resume el rendimiento de cada algoritmo según fueron analizados en sus respectivas secciones. El algoritmo TPM supera significativamente a los otros métodos. Por

Tabla 5.3: Tabla comparativa del rendimiento de los tres algoritmos

Algoritmo	**Tasa de identificación**	**Tasa de abstención**	**Tasa de error**	**Identificación/Error**
TPM	66,72	0,00	33,28	2,00
EDM	40,01	0,00	59,99	0,24
RoIs-FV	52,76	9,22	38,02	1,39



(a) (b)

Figura 5.8: Resultados generales de los tres algoritmos al reconocer usuarios de Transantiago.

Tabla 5.4: Tabla comparativa del rendimiento de los tres algoritmos utilizando un umbral de mínima similitud.

Algoritmo	**Tasa de identificación**	**Tasa de abstención**	**Tasa de error**	**Identificación/Error**
TPM	24,88	71,88	3,24	7,68
EDM	32,88	48,58	18,54	1,77
RoIs-FV (Canberra distance)	7,16	92,54	0,3	23,87

el contrario, el algoritmo EDM tiene un rendimiento notoriamente peor que los otros dos, ya que tiene la menor tasa de identificación, la mayor tasa de error y toma más tiempo en ejecutarse. El algoritmo RoIs-FV tiene un rendimiento regular, sin embargo su tasa de error es mayor que la del algoritmo TPM, a pesar de que el algoritmo RoIs-FV es capaz de abstenerse.

La Tabla 5.4 presenta los resultados de los algoritmos aplicando un umbral de mínima similitud. Se aplicó la condición presentada en la Sección 5.3.1 para seleccionar el óptimo en algoritmo RoIs-FV, es decir, se busca el umbral donde se maximice la razón entre la tasa de identificación y la tasa de error. En general, todos los algoritmos aumentaron la razón entre reconocidos y reconocidos incorrectamente con el costo de disminuir el número de identificados. De todos los resultados destaca la mejora en el rendimiento del algoritmo EDM, el cual disminuye su tasa de error en un 50 %. Sin embargo, se concluye que si bien el algoritmo EDM presenta un rendimiento regular al utilizar umbral de mínima similitud y al considerar el tiempo de ejecución, los algoritmos TPM y RoIs-FV presentan mejor rendimiento, ya sea al privilegiar la tasa de identificación o la razón entre los reconocidos y no reconocidos.

Finalmente, seleccionar el mejor rendimiento dependerá de cuán importante sea el número de reconocidos correctamente versus los reconocidos incorrectamente. Como una primera aproximación, es interesante observar cuánto se puede reducir la tasa de error, especialmente si es que se pudiese mantener aumentando el tamaño de la muestra de usuarios.

Capítulo 6

Variabilidad del comportamiento de usuarios de transporte público de Gatineau

En este capítulo se presentan los resultados obtenidos al aplicar los algoritmos TPM, EDM y RoIs-FV para medir la variabilidad del comportamiento de los usuarios del transporte público de Gatineau. Los resultados presentados en este capítulo corresponden al experimento explicado en la Sección 4.3.2, evaluados según los indicadores de variabilidad definidos en la Sección 4.4.1. Se utilizó la base de datos del sistema de transporte público de Gatineau, la cual almacena las transacciones de más de 100.000 usuarios, registradas durante el año 2012 y 2013.

Este capítulo se divide en cinco secciones. La primera sección describe brevemente los datos de variabilidad generados por usuario. Las secciones siguientes presentan resultados agregados de todos los usuarios del sistema. Específicamente, las tres secciones intermedias describen los resultados asociados a cada algoritmo evaluado bajo diferentes experimentos. La última sección corresponde a un análisis general de los resultados obtenidos por los tres algoritmos.

6.1. Formato de los resultados por usuario

El objetivo de esta sección es describir los datos generados luego de medir la variabilidad de cada usuario. Dos razones motivan esta sección: en primer lugar, ayudar a la comprensión de los resultados de las siguientes secciones. En segundo lugar, mostrar que los resultados individuales tienen potencial para ser aplicados en otras líneas de investigación.

Según lo descrito en la Sección 4.3.2, al finalizar el proceso de modelación de la etapa 2, cada usuario tiene asociado un objeto en el que se almacenan los datos relativos a la variación de su movilidad durante el periodo en que registró transacciones. En particular destacan tres atributos:

1. La matriz de variaciones
2. El arreglo de variaciones cercanas
3. El arreglo de variaciones contiguas

Luego, sobre estos atributos se calculan los indicadores descritos en la Sección 4.4.1, cuyos resultados serán presentados en las secciones siguientes.

En la Sección 4.3.2 se describe la matriz de variaciones como la matriz con las similitudes entre las subtablas de transacciones asociadas a cada ventana de tiempo. Es decir, esta matriz almacena los cambios de la movilidad de un usuario al comparar cada ventana de tiempo con el resto. La Figura 6.1 muestra tres mapas de calor de las similitudes almacenadas en la matriz de variaciones de tres usuarios del sistema de transporte público de Gatineau: usuarios A, B y C. Los eje X e Y representan las ventanas del periodo de actividad de cada usuario, en este caso, ventanas de tiempo de una semana. Los valores de las celdas [i, j] corresponden a la variabilidad entre las ventanas i y j, calculadas, en este ejemplo, con el algoritmo TPM.

De la Figura 6.1 se distingue inmediatamente el carácter triangular superior de la matriz de variaciones. Si bien todas las ventanas son comparadas con todas, solo se comparan en un sentido, donde cada ventana es comparada con las ventanas futuras. Además, en los mapas de calor de los usuarios B y C es posible observar algunas columnas grises, las cuales corresponden a ventanas de tiempo con menos de 8 transacciones por semana, es decir, con menos del mínimo de transacciones definido para extraer un perfil de movilidad comparable. Si bien las ventanas sin transacciones son omitidas, los índices en los ejes X e Y señalan el identificador de ventana correspondiente.

La Figura 6.1a ilustra la matriz de variaciones del usuario A, cuyo periodo de actividad es de 7 semanas. En el mapa de calor se perciben cambios en mayor y menor medida entre la movilidad de las diferentes ventanas, lo cual indica que la variación de la movilidad no es homogénea. Sin embargo, la mayoría de las ventanas presenta alta similitud, lo cual sugiere que, a pesar del breve periodo de actividad, el usuario A presenta un comportamiento altamente consistente.

La Figura 6.1b grafica la matriz de variaciones del usuario B, cuyo periodo de actividad es de 24 semanas, de las cuales 23 semanas registran viajes y 22 semanas presentan más de 8 transacciones. El mapa de calor muestra variaciones en todo el espectro de similitud, con la particularidad de que parecen haber columnas de colores similares, i.e. ventanas cuya movilidad es igualmente similar a todas las otras. Este fenómeno evidencia que el usuario B posee un patrón de movilidad presente en la mayoría de las ventanas. Luego, las ventanas de verde más oscuro corresponden a ventanas donde solo se presenta el núcleo de la movilidad, por ende, similar a todas las ventanas. Por otro lado, las columnas de verde más claro, evidencian semanas donde el núcleo de movilidad aparece acompañado de un alto porcentaje de viajes nuevos, por tanto, corresponden a ventanas de baja similitud en comparación al resto del periodo. Por lo anterior, el usuario B ejemplifica la existencia de usuarios cuyo comportamiento puede estar formado por componentes altamente regulares y variables, más que estar encasillado en una sola categoría.

La Figura 6.1c muestra la matriz de variaciones del usuario C, cuyo periodo de actividad es de 69 semanas, sin embargo, solo registra viajes en 15 semanas, de las cuales 12 presen-

tan más de 8 transacciones. El mapa de calor parece estar formado por tres regiones: dos regiones triangulares con similitudes de grado medio-alto, y una región rectangular de bajas similitudes, la cual está demarcada por la falta de transacciones suficientes en la semana 8. La falta de similitud entre las semanas posteriores a la semana 8 con las semanas previas, sugiere que el usuario C cambió su comportamiento durante las semanas en que no se percibieron transacciones. Es decir, es posible asociar la ausencia del usuario C en el sistema de transporte con un cambio drástico en su movilidad en la ciudad.

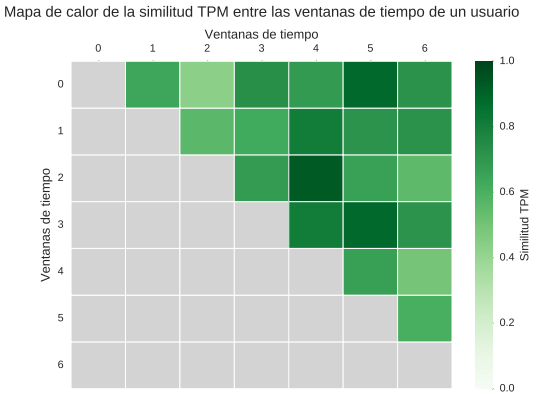
A partir de la matriz de variaciones se puede extraer el arreglo de variaciones cercanas y el arreglo de variaciones contiguas. Ambos arreglos representan las variaciones entre ventanas secuenciales. La diferencia, es que el arreglo de variaciones contiguas almacena solo las variaciones entre ventanas con distancia igual a 1, es decir omite variaciones entre ventanas separadas por periodos de inactividad o periodos con ventanas de menos de 8 transacciones por semana.

La Figura 6.2 muestra tres series temporales correspondientes a los arreglos de variaciones cercanas de los usuarios A, B y C de la Figura 6.1. El eje X corresponde a las ventanas del periodo de actividad del usuario, y el eje Y corresponde al indicador de variación de la movilidad, el cual en este ejemplo es el indicador de similitud del algoritmo TPM. Los puntos verdes señalan la primera ventana con transacciones suficientes y se ubican en la línea que denota el promedio de las similitudes del usuario. Los puntos azules marcan la similitud de cada ventana con la anterior. Los puntos amarillos representan ventanas con menos de 8 transacciones por semana, y los puntos rojos señalan ausencia de transacciones.

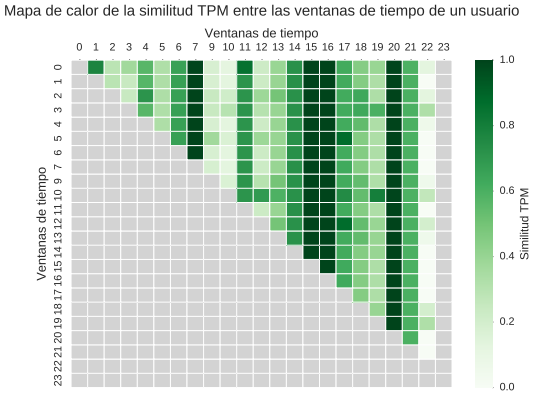
En la Figura 6.2a es posible observar que el usuario A registra una variabilidad promedio mayor que 0,6, y con un rango de variabilidad acotado entre 0,55 y 0,8. Para el usuario A los arreglos de variaciones cercanas y variaciones contiguas son equivalentes.

La Figura 6.2b muestra que si bien el usuario B posee una similitud promedio cercana a 0,6, no tan distante a la del usuario A, el rango de variabilidad ocupa todo el espectro de similitud. En particular, en las semanas 7, 15, 16 y 20 la movilidad es exactamente igual a la semana anterior, y en la semana 22 es completamente diferente. Es importante notar que considerando la diferencia entre el promedio y rango de la variabilidad de este usuario, resultaría difícil describir su comportamiento a través de datos agregados. El arreglo de variaciones contiguas del usuario B correspondería al conjunto de similitudes graficado en la Figura 6.2b, salvo por la similitud de la ventana 9, que corresponde a la similitud entre la ventana 7 y 9, es decir separadas por más de una ventana.

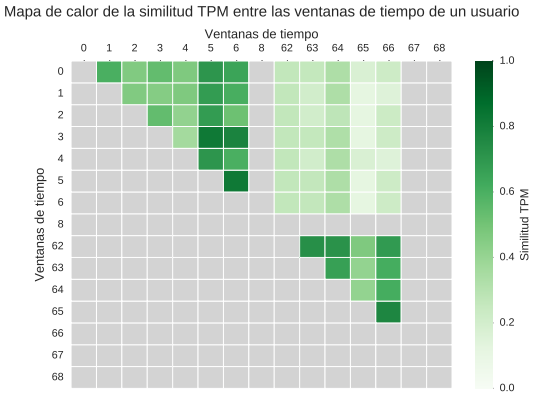
La Figura 6.2c indica que la similitud promedio del usuario C es cercana a 0,6 y su rango de variabilidad es entre 0,2 y 0,8. En este gráfico el mínimo de similitud se observa en la semana 62, el cual corresponde a la similitud entre la ventana 6 y la ventana 62, es decir la movilidad entre periodos con más de un año de separación. La singularidad del usuario C, es que muestra un comportamiento medianamente regular con breves periodos de alta frecuencia de viajes. Lo anterior no habría sido advertido al utilizar un indicador de frecuencia agregado como la relación entre el número de transacciones y el largo del periodo de actividad. Para este usuario, el arreglo de variaciones contiguas corresponde al conjunto de similitudes graficado en la Figura 6.2c, salvo por la similitud registrada en la semana 62.



(a) Usuario A

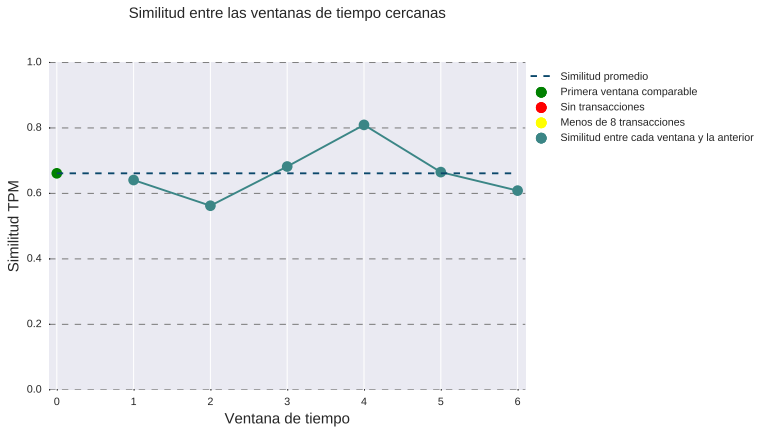


(b) Usuario B

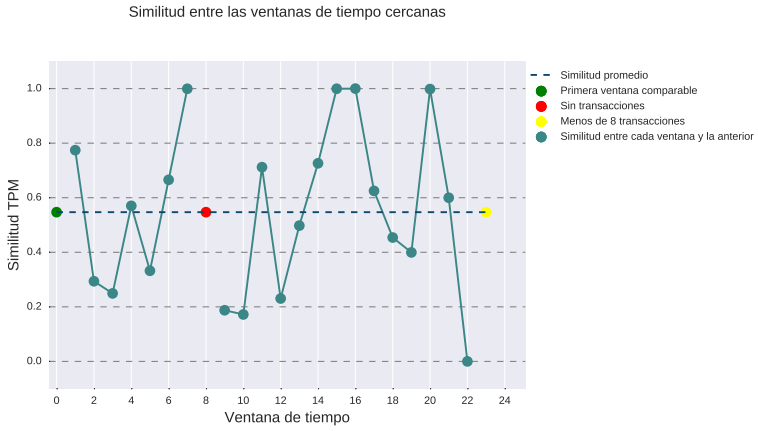


(c) Usuario C

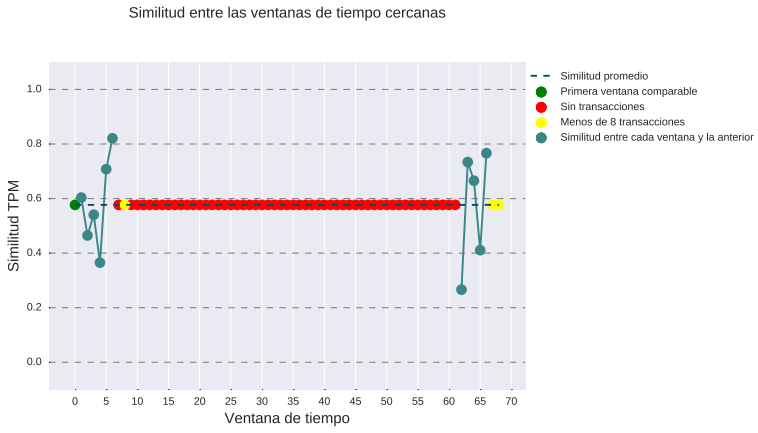
Figura 6.1: Mapa de calor de las variaciones entre las ventanas de tiempo de tres usuarios de ejemplo.



(a) Usuario A



(b) Usuario B



(c) Usuario C

Figura 6.2: Series temporales de la variabilidad almacenada en los arreglos de variaciones cercanas de tres usuarios de ejemplo.

Los ejemplos presentados corresponden a tres usuarios de los más de 100.000 de la base de datos de Gatineau. Como resultaría extenso hacer un análisis individual para todos, en las siguientes secciones se presenta un análisis agregado de las distintas variaciones de los usuarios. No obstante, el análisis individual automatizado a través de técnicas de minería de datos permitiría encontrar patrones de la variabilidad de los usuario. Por ejemplo, en la matriz de variaciones se podrían identificar y describir usuarios con variaciones periódicas. También, a través del análisis de las series temporales, se podría estudiar el comportamiento de los usuarios después de periodos de inactividad. La potencialidad de los datos generados en esta tesis coincide con la idea general de que los datos de tarjetas inteligentes continúan presentando desafíos y oportunidades para desarrollar herramientas que mejoren la comprensión de los usuarios de transporte público.

6.2. Resultados asociados al algoritmo TPM

Esta sección presenta los resultados agregados de la medición de la variabilidad de los usuarios utilizando el indicador de similitud TPM. El algoritmo TPM fue evaluado variando tres factores:

1. La proximidad de las ventanas
2. Número de semanas de las ventanas de tiempo
3. Agregación espacial de las posiciones de origen y destino que componen la matriz de transición de probabilidad

6.2.1. Análisis de la influencia de la proximidad de las ventanas

En este experimento se evaluó la diferencia de los indicadores de variabilidad (ver Sección 4.4.1), luego de considerar la matriz de variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas, donde cada grupo de variaciones es más riguroso que el anterior respecto de la cercanía entre las ventanas que se comparan.

Los gráficos de la Figura 6.3 muestran la silueta de la frecuencia relativa de cada indicador de variabilidad sobre la similitud TPM entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido a partir de la matriz de variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas. En este experimento se consideraron ventanas de una semana con un mínimo de 8 transacciones por ventana.

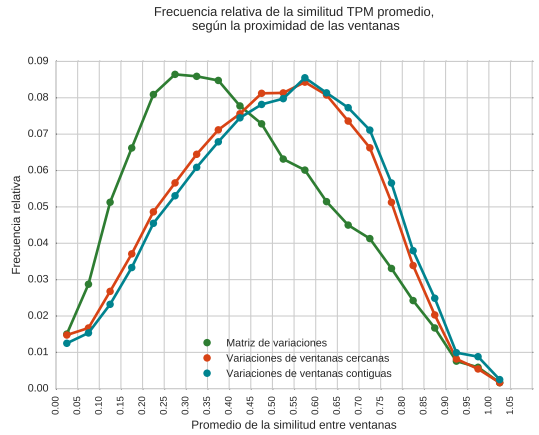
En la Figura 6.3a se observa la distribución del indicador promedio. La distribución del promedio de la matriz de variaciones abarca todo el espectro de similitud, no obstante, se concentra principalmente en la mitad inferior del espectro. Lo anterior indica que la mayoría de los usuarios posee una alta variabilidad al comparar ventanas que no necesariamente se encuentran en orden secuencial. La distribución de la matriz de variaciones se separa notoriamente de las distribuciones de las ventanas cercanas y ventanas contiguas. Estas últimas tienen una distribución similar entre ellas, concentrada en la mitad del espectro, con máximo

en el rango [0,55-0,60]. La distribución de las ventanas contiguas presenta un leve desplazamiento hacia el extremo derecho del espectro (mayor similitud) que la distribución de ventanas cercanas.

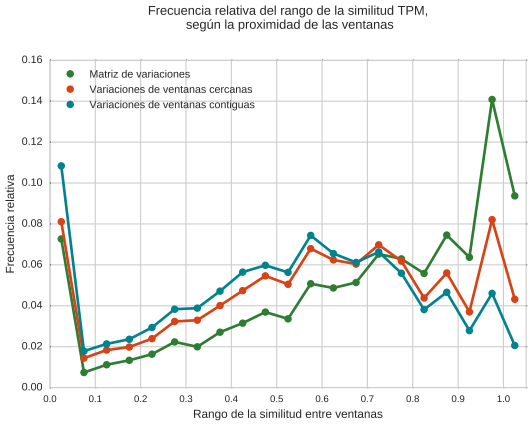
Los gráficos de la Figura 6.3 de indicadores distintos al promedio muestran distribuciones más bien similares entre las distintas variaciones comparadas. En general, las tres distribuciones muestran dos máximos: un máximo en 0 y otro dependiente del indicador. Sin embargo al mirar con detención, se observa que en valores cercanos al extremo izquierdo las variaciones de ventanas contiguas presentan una mayor concentración. Por el contrario, en valores cercanos al extremo derecho se ve que la matriz de variaciones presenta mayor frecuencia relativa. De esta forma, se concluye que la matriz de variaciones presenta mayor variabilidad que las variaciones contiguas. Esta tendencia se observa en mayor medida en los indicadores rango (Figura 6.3b) y en el coeficiente de variación (Figura 6.3f).

En la Figura 6.3b se observa que para los tres tipos de variaciones existe una alta concentración de usuarios con rango mayor a 0,5. Esto indica una alta variación en la mayoría de los usuarios. En el caso de la matriz de variaciones, se distingue claramente un máximo cercano a 1. Este último, indica que existe alrededor de un 14 % de usuarios con al menos dos ventanas de tiempo con ningún viaje en común (similitud 0), y otro par de ventanas cuya movilidad es completamente similar (similitud 1). No obstante, también se observa que las tres variaciones tienen un máximo local en torno a 0. En particular para las variaciones de ventanas contiguas hay alrededor de un 10 % de usuarios que conserva las variaciones de su movilidad constante ventana a ventana.

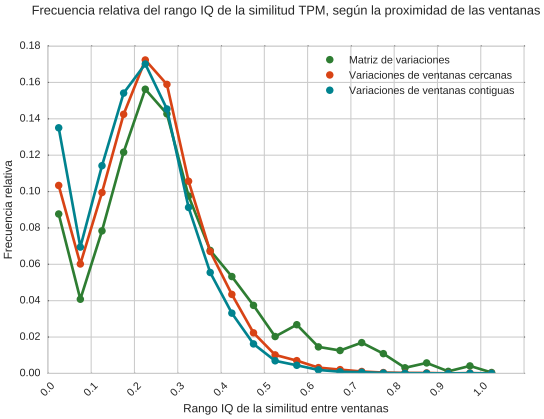
El rango es el único indicador que señala una alta variación de los usuarios, lo cual tiene sentido porque es el indicador más sensible, ya que considera los valores extremos de cada usuario. En general los otros indicadores de variación señalan comportamientos relativamente estables, especialmente en las variaciones de ventanas cercanas y contiguas. Por último es importante notar que los indicadores distintos al promedio ilustran la presencia de usuarios cuya similitud ventana a ventana es extremadamente constante. Sin embargo, el promedio no exhibe ningún máximo local cercano a 1 que indique movilidad constante. De lo anterior se concluye que aquellos usuarios con variabilidad constante no necesariamente se relacionan a usuarios con movilidad regular, sino más bien, indica que hay usuarios que mantienen constante el porcentaje de viajes nuevos ventana a ventana.



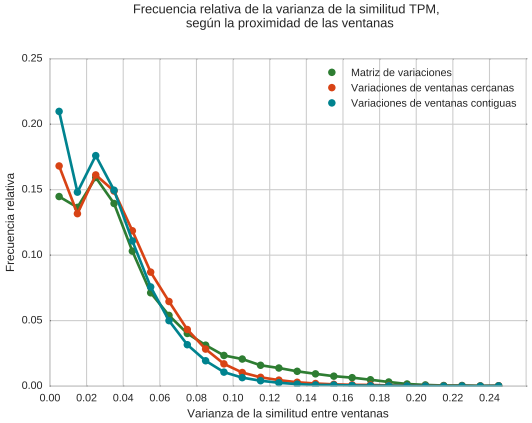
(a) Promedio



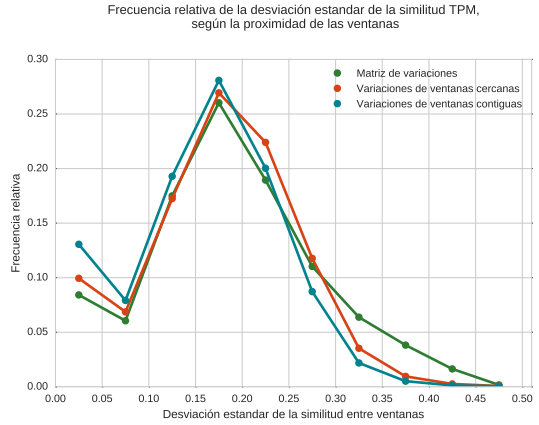
(b) Rango



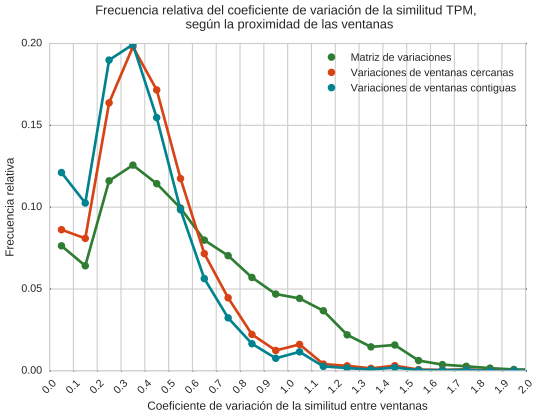
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.3: Silueta de la frecuencia relativa de cada indicador sobre la similitud TPM entre las ventanas de un usuario, según la proximidad de las ventanas comparadas.

A modo de resumen, la Tabla 6.1 muestra el promedio de las distribuciones de cada indicador. Como el algoritmo TPM entrega una medida de similitud, entonces el promedio es una medida de similitud. En la tabla se observa que su valor promedio aumenta al incrementar la proximidad de las ventanas. En cambio, los otros indicadores son medidas de variación, y en

la tabla se observa que sus valores disminuyen al incrementar la proximidad de las ventanas. Luego, la Tabla 6.1 confirma la tendencia observada previamente: a mayor cercanía de las ventanas mayor es la constancia de la variación del movimiento.

Para confirmar la tendencia observada se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon (Wilcoxon, 1945), con un nivel de significación de 0,01. Se eligió este test por ser una prueba no paramétrica que permite relacionar muestras relacionadas (que es el caso, ya que en cada experimento se comparan resultados de los mismos usuarios). El test se realizó sobre la distribución del indicador promedio de las siguientes combinaciones:

1. la matriz de variaciones con las variaciones de ventanas cercanas
2. la matriz de variaciones con las variaciones de ventanas contiguas
3. las variaciones de ventanas cercanas con las variaciones de ventanas contiguas

En todos los casos anteriores el p-valor resultó 0.0, por tanto se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. De este modo, se concluye que a mayor proximidad entre las ventanas comparadas menor es la variabilidad observada.

Tabla 6.1: Valores promedio de los indicadores de variabilidad de la similitud TPM según la proximidad de las ventanas.

Indicador	Todas con todas	Ventanas más cercanas	Ventanas contiguas
Promedio	0,42	0,50	0,51
Rango	0,67	0,58	0,52
Rango IQ	0,28	0,23	0,21
Var	0,04	0,04	0,03
STD	0,19	0,17	0,16
CV	0,58	0,41	0,37

6.2.2. Análisis de la influencia del tamaño de las ventanas de tiempo

En este experimento se evaluó la diferencia de los indicadores de variabilidad según el tamaño de la ventana de tiempo en que se divide el periodo de actividad de cada usuario.

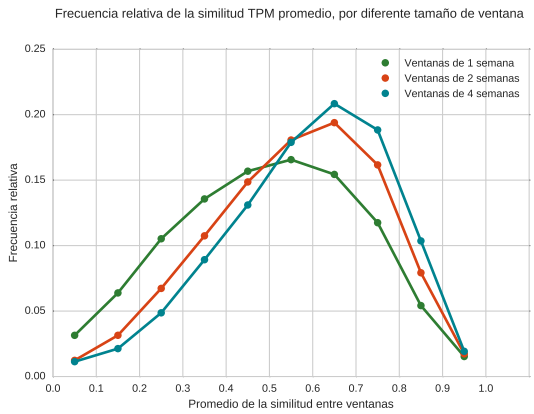
Los gráficos de la Figura 6.4 muestran la silueta de la frecuencia relativa de cada indicador de estabilidad sobre la similitud TPM entre las ventanas cercanas de cada usuario. En cada gráfico se compara el indicador obtenido utilizando ventanas de una semana, dos semanas y cuatro semanas, con un mínimo de 8 transacciones por semana.

En todos los gráficos de la Figura 6.4 se observa claramente un orden entre las tres distribuciones. En el caso de los indicadores de variabilidad (aquellos distintos al promedio), la distribución de las ventanas de cuatro semanas es la más cercana al extremo izquierdo del espectro de cada indicador, es decir, la menos variable. Le siguen en orden las ventanas de dos semanas y una semana. En el caso del indicador promedio (Figura 6.4a), se obser-

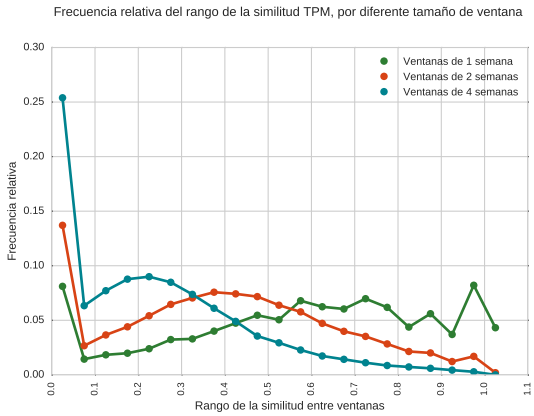
va lo inverso: las ventanas de cuatro semanas se encuentran orientadas al extremo derecho, indicando mayor similitud de la movilidad de los usuarios.

Los gráficos de la Figura 6.4 de indicadores distintos al promedio muestran distribuciones con dos máximos: uno cercano al 0, y otro que depende del tamaño de la ventana y del indicador. Al ser un resultado transversal, se concluye que para los tres tamaños de ventana analizados se encuentra un grupo importante de usuarios que muestra absoluta constancia en la variación de la movilidad. En particular, al utilizar ventanas de cuatro semanas se observa que alrededor de un 25 % de los usuarios exhibe máxima estabilidad.

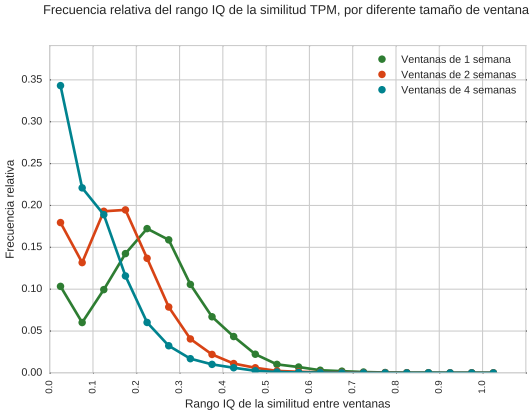
En el caso del rango, la Figura 6.4b muestra que las distribuciones de las ventanas de una semana y de dos semanas exhiben una importante concentración de usuarios con al menos un par de ventanas con alta variabilidad (rango mayor a 0,5). Sin embargo, de la Figura 6.4c se desprende que, independiente del tamaño de la ventana, la gran mayoría de los usuarios presentan un rango interquartil entre 0,0 y 0,5. Es decir, se observa una baja variabilidad entre la mayoría de las ventanas de cada usuario, incluso al utilizar ventanas de una semana.



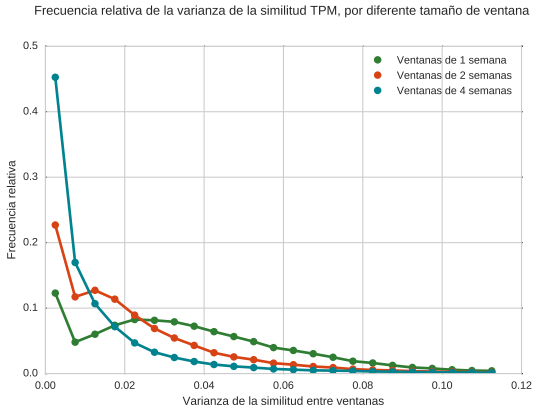
(a) Promedio



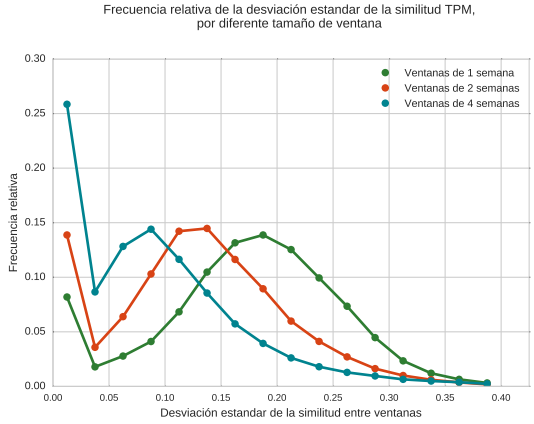
(b) Rango



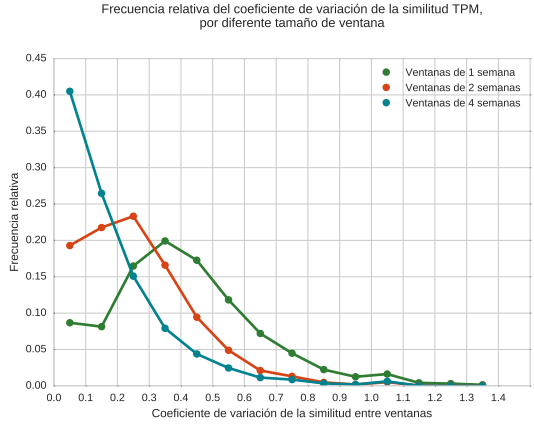
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.4: Los gráficos de esta figura muestran la silueta de la frecuencia relativa de cada indicador sobre la similitud TPM entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido al comparar las variaciones entre ventanas cercanas al variar el tamaño de la ventana entre una semana, dos semanas y cuatro semanas.

La Tabla 6.2 muestra el promedio de las distribuciones de cada indicador según el tamaño de las ventanas de tiempo. Se observa que el valor promedio del indicador promedio es el único en aumentar al incrementar el tamaño de la ventana. En el resto de los indicadores,

se observa una clara disminución de la variación al aumentar el tamaño de la ventana. De esta forma, la Tabla 6.1 confirma la tendencia observada previamente: a mayor tamaño de las ventanas mayor es la estabilidad de la variación del movimiento.

Para confirmar la tendencia observada se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon, con un nivel de significación de 0,01. Se compararon las distribuciones del indicador promedio de las siguientes combinaciones de tamaño de ventana:

1. ventanas de una semana con ventanas de dos semanas
2. ventanas de una semana con ventanas de cuatro semanas
3. ventanas de dos semanas con ventanas de cuatro semanas

En todos los casos anteriores el p-valor resultó 0.0, por lo que se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. De este modo se confirma que la tendencia es significativa, por tanto se concluye que a mayor tamaño de las ventanas comparadas menor es la variabilidad observada.

Tabla 6.2: Valores promedio de los indicadores de variabilidad de la similitud TPM según el tamaño de las ventanas.

Indicador	**Una semana**	**Dos semanas**	**Cuatro semanas**
Promedio	0,50	0,56	0,59
Rango	0,58	0,40	0,24
Rango IQ	0,23	0,15	0,10
Var	0,04	0,02	0,01
STD	0,17	0,13	0,09
CV	0,41	0,26	0,17

6.2.3. Análisis de la influencia del nivel de agregación espacial

En este experimento se evaluó la diferencia de los indicadores de variabilidad según el nivel de agregación espacial de los orígenes y destinos que conforman la matriz de probabilidad de transición del algoritmo TPM. Los niveles de agregación que se analizaron fueron: paradas de buses, zona 400, *clusters* de 500 metros, y zona 66. El nivel más desagregado (paradas de buses) corresponde a usar los identificadores de las paradas de buses del sistema de transporte de Gatineau. Los niveles zona 400 y zona 66, corresponden a zonificaciones disponibles que agrupan las paradas en 400 y 66 zonas respectivamente. El nivel *cluster* de 500 metros corresponde a las zonas resultantes luego de utilizar *clustering* jerárquico con un límite de 500 metros de radio por *cluster*. Este último nivel de agregación permite agrupar las transacciones de los usuarios en zonas personalizadas. Luego, las matrices de probabilidad de transición de cada usuario se construyen según los viajes entre las zonas del nivel de agregación que se haya seleccionado.

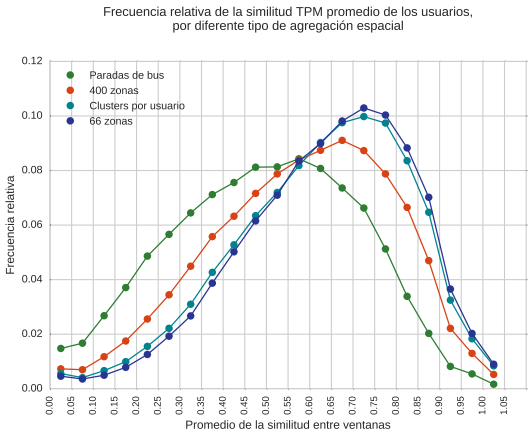
Los gráficos de la Figura 6.5 muestran la silueta de la frecuencia relativa de cada indicador de estabilidad sobre las variaciones entre ventanas cercanas de cada usuario, medidas con el algoritmo TPM. Se utilizaron ventanas de una semana, con un mínimo de 8 transacciones por

semana. En cada gráfico se compara el indicador obtenido utilizando los niveles de agregación disponibles: paradas de buses, zona 400, *clusters* de 500 metros, y zona 66.

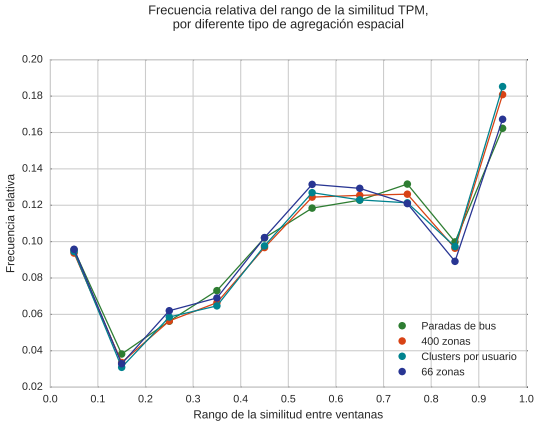
El gráfico de la Figura 6.5a muestra las distribuciones del indicador promedio según los diferentes niveles de agregación espacial. En esta imagen es posible distinguir que las distribuciones se encuentran separadas según su cercanía con el extremo derecho del espectro de similitud. Utilizando el nivel de paradas de bus, los usuarios exhiben una similitud de la movilidad mucho menor que utilizando el resto de los niveles. Luego le sigue el nivel de agregación de 400 zonas, cuya distribución alcanza un máximo en el rango [0,65-0,7], lo cual se considera alta similitud. Los niveles de agregación de *clusters* y zona 66 tienen distribuciones levemente distinguibles, compartiendo la moda en el rango [0,7-0,75]. La distribución de zona 66 muestra mayor frecuencia relativa en los valores de similitud mayores a 0,75. En cualquiera de los niveles de agregación mayores a las paradas de buses, se observa que la mayoría de los usuarios posee una alta similitud entre las ventanas de movilidad.

Los resultados del coeficiente de variación, graficados en la Figura 6.5f permite distinguir mejor las distribuciones entre los cuatro niveles de agregación, donde nuevamente el nivel de mayor agregación presenta mayor frecuencia relativa en los valores de baja variación. También se observa que con cualquier nivel de agregación espacial, la mayoría de los usuarios presenta coeficientes de variación menores que 0,4, lo cual indica baja variabilidad del indicador de similitud.

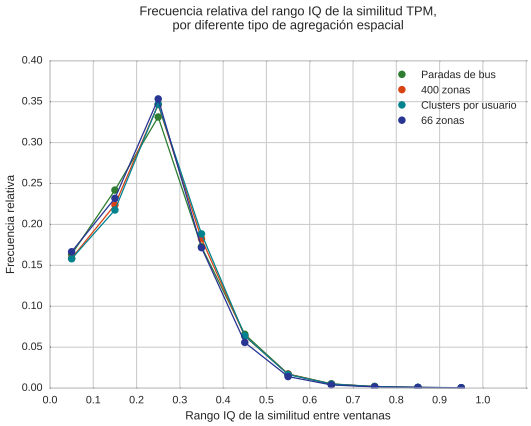
Los gráficos de la Figura 6.5 correspondientes a los indicadores: rango, rango IQ, varianza y desviación estándar, exhiben una alta similitud en la distribución de todos los niveles de agregación. De estos indicadores, solo el rango muestra diferencias notorias. En la Figura 6.5b, se puede observar que la distribución del rango, en todos los niveles de agregación, tiene un máximo local en los valores cercanos a 1, lo cual indica alta variabilidad por parte de un número importante de usuarios. No obstante, la frecuencia relativa del máximo local cercano a 1 de la zona 66 es bastante menor que en las otras zonificaciones. De lo anterior se concluye que en general los distintos niveles de agregación no provocan cambios relevantes en la dispersión de la similitud de la movilidad de los usuarios, salvo en casos de variación extrema, donde el máximo nivel de agregación disponible parece suavizar el nivel de variación observado.



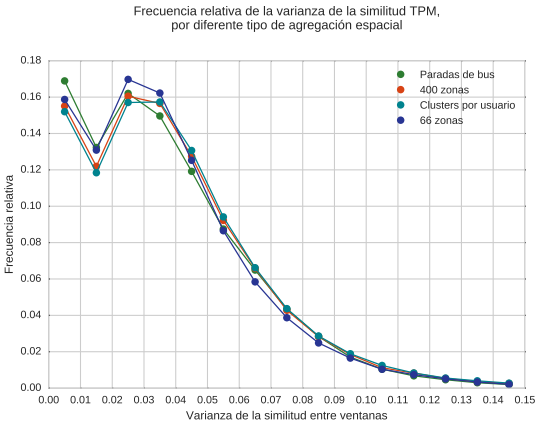
(a) Promedio



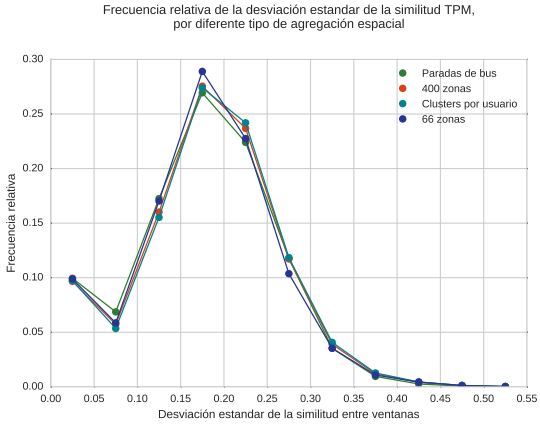
(b) Rango



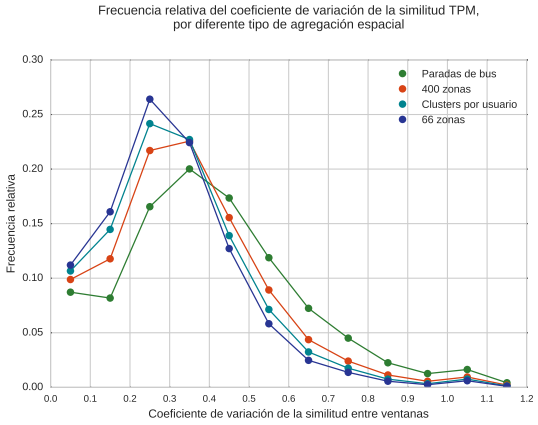
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.5: Los gráficos de esta figura muestran la silueta de la frecuencia relativa de cada indicador sobre la similitud TPM entre las ventanas de cada usuario. En cada gráfico se compara el indicador calculado sobre la matriz de variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas. Se consideraron ventanas de una semana con un mínimo de 8 transacciones por ventana.

La Tabla 6.3 muestra el promedio de las distribuciones de cada indicador según el nivel de agregación espacial. El rango, rango IQ, la varianza y la desviación estándar en general se mantienen constantes. Esto indica que las diferentes agregaciones espaciales no afectan de manera importante la dispersión de la similitud registrada. Por otro lado, el indicador promedio y coeficiente de variación si varían, lo cual señala que el nivel de agregación espacial si afecta a la similitud de la movilidad entre las ventanas. Resumiendo, el nivel de agregación espacial determina el nivel de similitud que se observa entre las ventanas, pero no altera la variación de la similitud a lo largo del tiempo.

De la Tabla 6.3 también se desprende que no existe una tendencia clara de la variabilidad de la movilidad en relación al nivel de agregación espacial. Por un lado, se observa que en los niveles de agregación: parada, zona 400 y zona 66, el indicador promedio aumenta en el mismo orden que el nivel de agregación. En consecuencia, pareciera que mayor agregación espacial aumenta la estabilidad de la movilidad observada. Sin embargo, al agrupar las transacciones de los usuarios en *clusters* de 500 metros, si bien se realiza una partición menos agregada que el nivel zona 400, los resultados de la tabla muestran una estabilidad similar al nivel zona 66. Por lo anterior se concluye que utilizar una agrupación espacial individual para cada usuario permite obtener una alta estabilidad de la movilidad, sin necesidad de homogeneizar los datos espaciales con una zonificación más agregada.

Es importante notar que el efecto observado en la similitud de la movilidad al utilizar *clustering* de las transacciones no sería notorio si los usuarios siempre viajaran entre las mismas paradas orígen-destino. Por consiguiente, los resultados evidencian que incluso en viajes homólogos existe variabilidad en las paradas que se utilizan para abordar o descender.

Tabla 6.3: Valores promedio de los indicadores de variabilidad de la similitud TPM según el nivel de agregación espacial.

Indicador	**Parada**	**Zona 400**	**Clusters 500m**	**Zona 66**
Promedio	0,50	0,58	0,63	0,64
Rango	0,58	0,59	0,59	0,58
Rango IQ	0,23	0,23	0,23	0,23
Var	0,04	0,04	0,04	0,04
STD	0,17	0,18	0,18	0,18
CV	0,41	0,35	0,32	0,30

Para verificar que los resultados de las distintas zonas de agregación son significativamente diferentes se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon, con un nivel de significación de 0,01. Se compararon las distribuciones del indicador promedio de las siguientes combinaciones de zonas de agregación:

1. parada con zona 400
2. parada con *clusters* de 500m
3. parada con zona 66
4. zona 400 con *clusters* de 500m
5. zona 400 con zona 66
6. *clusters* de 500m con zona 66

En todos los casos anteriores el p-valor resultó 0.0, por lo que se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. De este modo se confirma que las diferencias observadas son significativas.

6.2.4. Resultados de la variabilidad de los usuarios agregados por semana del periodo 2012-2013

Además de analizar las variaciones de la movilidad individual, resulta interesante observar desde una perspectiva general cómo se comporta la variación de todos los usuarios del sistema de transporte a través del tiempo. La Figura 6.6 presenta el diagrama de caja de la similitud TPM entre la movilidad de cada usuario respecto la semana anterior, para cada semana del periodo 2012-2013. Cada diagrama de caja representa la división de las similitudes de los usuarios en quartiles, donde la línea roja denota la mediana y la caja verde marca el rango del segundo y tercer quartil. Los puntos verdes representan valores atípicos.

En la Figura 6.6 es posible observar que la mediana se encuentra en la mayoría de los casos entre 0,5 y 0,6, salvo semanas especiales. En particular, las semanas correspondientes a las fiestas de fin de año presentan una notoria caída en la similitud de la movilidad de los usuarios. Además la semana 95 presenta una caída de la similitud tal que el 50 % de los usuarios muestra una similitud TPM menor a 0,2 en relación a la movilidad de la semana 94. Esto se debe a la implementación del corredor Rapibus en el sistema de transporte de Gatineau.

Dejando de lado las semanas especiales, es posible observar que el algoritmo TPM distribuye las variaciones de los usuarios a lo largo de todo el espectro de similitud. De esta forma, es posible dividir los usuarios en tres grupos: un 50 % de los usuarios con variaciones de valores intermedios, un 25 % de usuarios con alta variabilidad y un 25 % de usuarios con alta similitud.

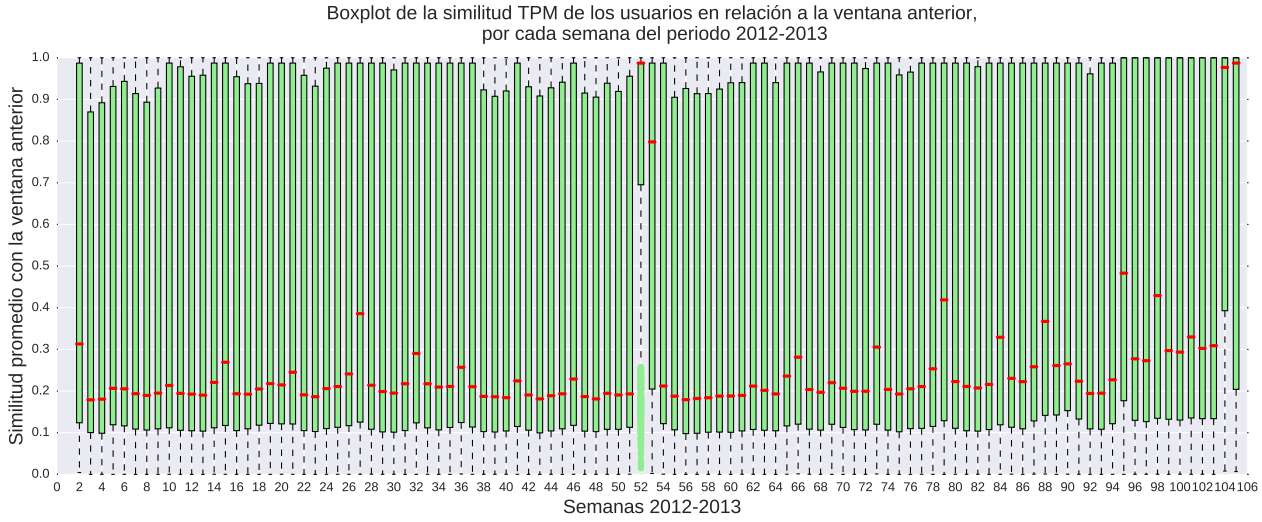


Figura 6.6: Diagrama de caja de la similitud TPM de los usuarios en relación a la ventana anterior, por cada semana del periodo 2012-2013.

6.3. Resultados asociados al algoritmo EDM

Esta sección presenta los resultados agregados de la medición de la variabilidad de los usuarios utilizando la distancia EDM. Debido al tiempo de ejecución de este algoritmo, solo se evaluó la influencia en las variaciones resultantes al cambiar el tamaño de la ventana de tiempo. Se utilizaron ventanas de tiempo de una, dos y cuatro semanas.

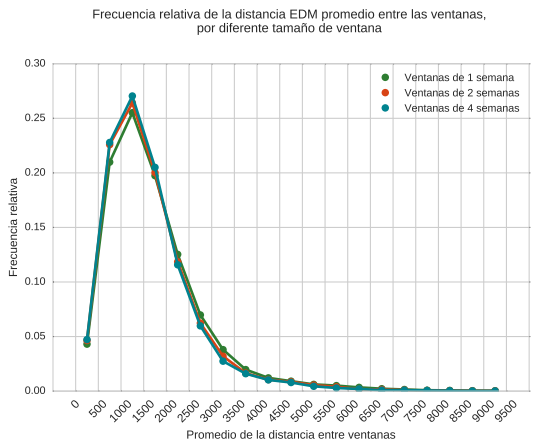
Los gráficos de la Figura 6.7 muestran la silueta de la frecuencia relativa de cada indicador de variabilidad sobre la distancia EDM entre las ventanas cercanas de cada usuario. En cada gráfico se compara el indicador obtenido utilizando ventanas de una semana, dos semanas y cuatro semanas, con un mínimo de 8 transacciones por semana.

En la Figura 6.7a se observa que la curva del promedio es casi indistinguible para los tres tamaños de ventanas analizados. Las tres curvas tiene forma de distribución normal truncada, con una moda cercana a 1.500. A pesar de que la cola de la distribución supere el valor 8,000, la mayoría de los usuarios tienen una distancia EDM promedio entre sus ventanas menor a 2.000. Lo anterior se interpreta como que la acumulación de los costos de las operaciones mínimas para transformar las trayectorias de una ventana a otra, es menor a 2 kilómetros. Es decir, incluso considerando las variaciones en la movilidad y las visitas a nuevos lugares, las variaciones en el centroide de la trayectoria son de baja magnitud.

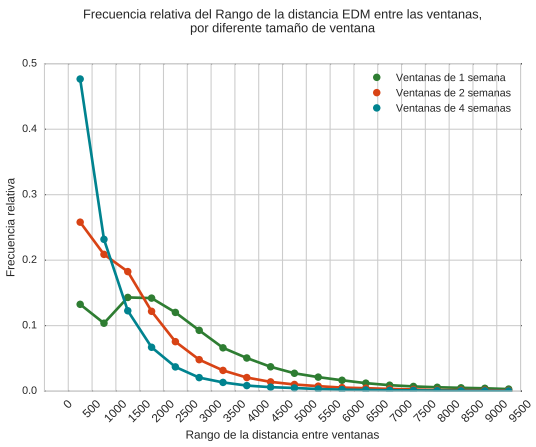
En los gráficos de la Figura 6.7 correspondientes a los indicadores: rango, rango IQ, desviación estándar y coeficiente de variación, es posible distinguir las curvas de las tres distribuciones comparadas, donde a mayor tamaño de la ventana mayor es la concentración en valores cercanos a 0, es decir de baja variabilidad. Resulta particularmente notable que cerca del 50 % de los usuarios muestre un rango cercano a 0 cuando se consideran ventanas de cuatro semanas. Esto indica que para la mayoría de los usuarios, la variación geográfica de los lugares visitados mes a mes es casi nula.

Todos los gráficos de la Figura 6.4 indican que utilizando la distancia EDM, la mayoría de los usuarios presenta baja variabilidad de la movilidad. Al ser un resultado transversal, y comparando con los resultados del algoritmo TPM donde si se evidencia variabilidad (para los mismos usuarios), se concluye que la distancia EDM homogeneiza el comportamiento de los usuarios al considerar solo la trayectoria geográfica y no el modo de transporte.

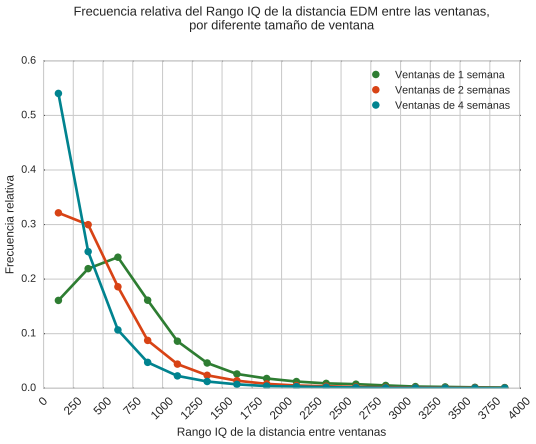
Finalmente, al observar que el tamaño de las ventanas no influencia notoriamente al promedio de la distancia EDM entre las ventanas, pero que si influencia la dispersión de las distancias, se concluye que utilizar ventanas de tiempo de mayor tamaño permite disminuir la influencia de aquellas escasas ventanas que registran mayor variación de la movilidad.



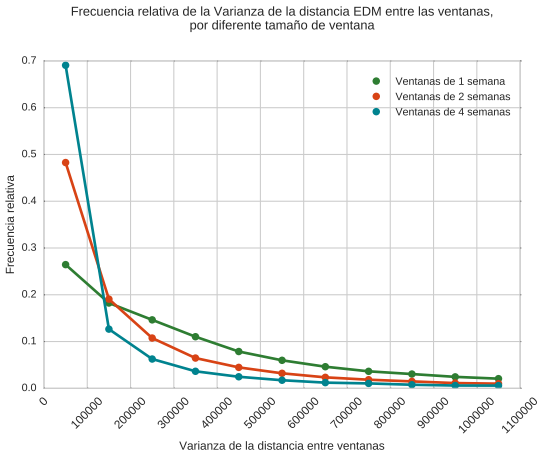
(a) Promedio



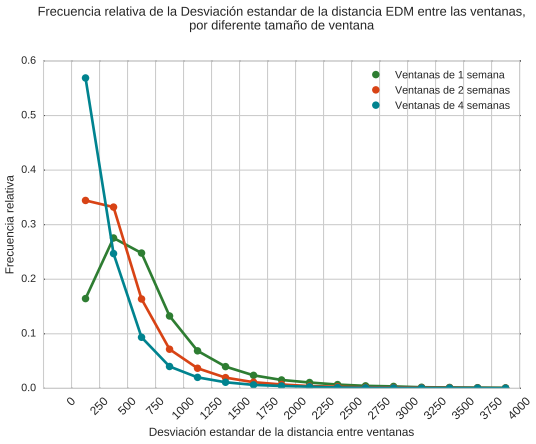
(b) Rango



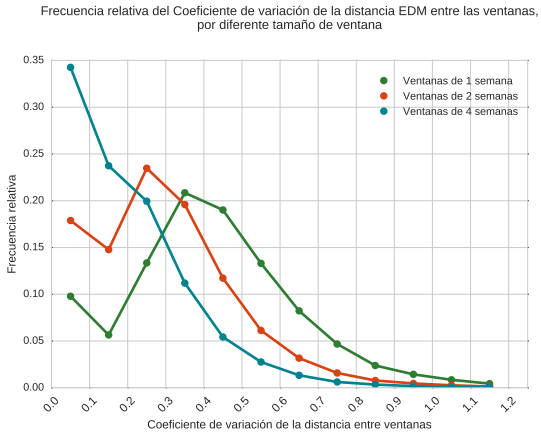
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.7: Los gráficos de esta figura muestran la silueta de la frecuencia relativa de cada indicador sobre la similitud EDM entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido al comparar las variaciones entre ventanas cercanas al variar el tamaño de la ventana entre una semana, dos semanas y cuatro semanas.

La Tabla 6.4 muestra el promedio de las distribuciones de cada indicador según el tamaño de las ventanas de tiempo. Se observa que el valor promedio de todos los indicadores disminuye al incrementar el tamaño de la ventana. Sin embargo, para el indicador promedio la diferencia entre los resultados con distintos tamaños de ventana es mucho menos evidente que en los otros indicadores. De esta forma, la Tabla 6.1 confirma la tendencia observada previamente: a mayor tamaño de las ventanas mayor es la estabilidad de la variación del movimiento.

Para verificar la tendencia observada se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon, con un nivel de significación de 0,01. Se compararon las distribuciones del indicador promedio de las siguientes combinaciones de tamaño de ventana:

1. ventanas de una semana con ventanas de dos semanas
2. ventanas de una semana con ventanas de cuatro semanas
3. ventanas de dos semanas con ventanas de cuatro semanas

En todos los casos anteriores el p-valor resultó 0.0, por lo que se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. De este modo se verifica que la diferencia entre los conjuntos de indicadores es significativa, por tanto se concluye que a mayor tamaño de las ventanas comparadas menor es la variabilidad observada.

Tabla 6.4: Valores promedio de los indicadores de variabilidad de la distancia EDM según el tamaño de las ventanas.

Indicador	**Una semana**	**Dos semanas**	**Cuatro semanas**
Promedio	1.733,70	1.647,70	1.615,05
Rango	2.497,91	1.462,99	862,31
Rango IQ	746,97	495,65	328,93
Var	876.998,38	459.364,73	283.441,19
STD	684,38	460,10	312,36
CV	0,42	0,29	0,19

6.3.1. Resultados de la variabilidad de los usuarios agregados por semana del periodo 2012-2013

La Figura 6.8 presenta el diagrama de caja de la distancia EDM entre la movilidad de cada usuario respecto la semana anterior, para cada semana del periodo 2012-2013. Cada diagrama de caja representa la distribución de las distancias entre ventanas de los usuarios. Las líneas rojas denotan la mediana y las cajas naranjas marcan el rango entre el segundo y tercer quartil. Los puntos naranjos representan valores atípicos.

En la Figura 6.8 es posible observar que la mediana se encuentra en la mayoría de los casos entre 0,2 y 0,3, salvo semanas especiales. En particular, las semanas correspondientes a las fiestas de fin de año presentan un alza en el indicador de distancia, evidenciando que los usuarios cambian sus trayectorias en esta época del año. Dejando de lado las semanas especiales, es posible observar que el algoritmo EDM distribuye las variaciones de la mayoría de los usuarios en el rango 0,0-0,5. También se observa la presencia de valores atípicos en el

extremo superior del espectro de distancia, lo cual indica que la frecuencia de esos valores es muy baja comparada con el resto del espectro. Por lo anterior, a pesar de que este método encasilla a la gran mayoría de los usuarios como muy regulares, es posible identificar a aquellos usuarios que destacan por la alta variabilidad geográfica de sus trayectorias.

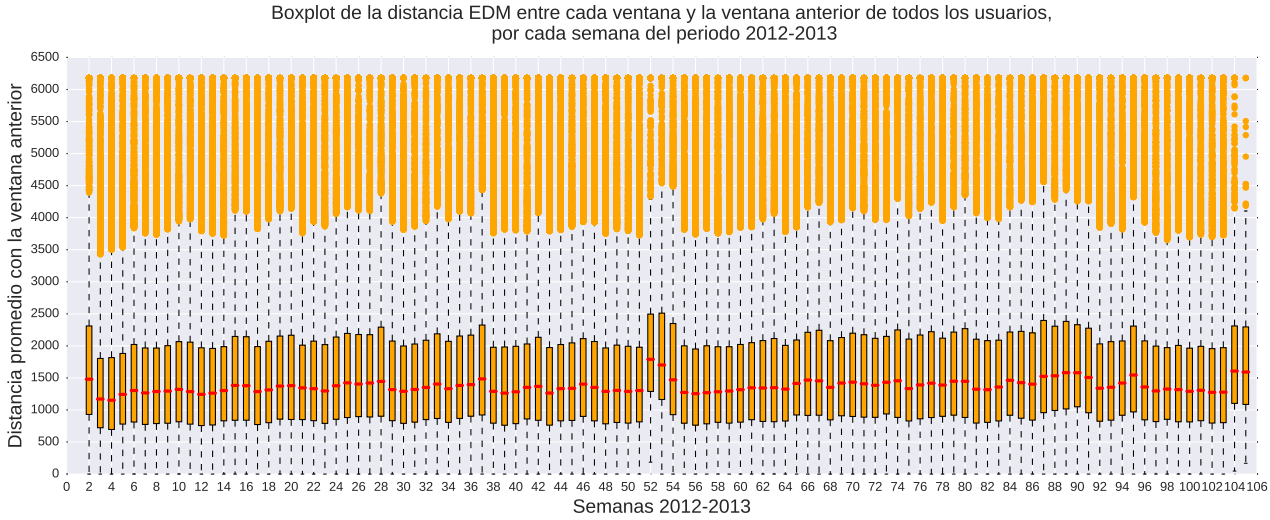


Figura 6.8: Diagrama de caja de la distancia EDM de los usuarios en relación a la ventana anterior, por cada ventana del periodo 2012-2013.

6.4. Resultados asociados al algoritmo RoIs-FV

Esta sección presenta los resultados agregados de la medición de la variabilidad de los usuarios utilizando el algoritmo de comparación de la movilidad RoIs-FV. El algoritmo RoIs-FV fue evaluado variando dos factores:

1. La proximidad de las ventanas
2. Número de semanas de las ventanas de tiempo

Es importante recordar que este algoritmo si bien entrega una distancia de la movilidad, solo lo hace cuando los registros comparados cumplen con un mínimo de similitud. En este trabajo se estableció que antes de calcular la distancia, se exigieran al menos dos RoIs en común. En caso de no cumplirse la condición el resultado es la etiqueta cualitativa "No comparables". Por otra parte, para medir las variaciones de la movilidad de un usuario a lo largo del tiempo resulta necesario cuantificar los casos en que cambian sus RoIs, de lo contrario la variabilidad total del usuario omitiría los casos de variaciones drásticas induciendo un sesgo de estabilidad. Debido a lo anterior, se optó por asociar la etiqueta "No comparables.ªl máximo de disimilitud.

Como se indica en la descripción del algoritmo RoIs-FV (Sección 3.3), la distancia resultante depende de la dimensión del vector de características y la distancia utilizada. Por lo tanto establecer un máximo dependerá de los resultados obtenidos. Con el objetivo de minimizar las distorsiones producidas al cuantificar las movilidades "No comparables", se realizó

un proceso de detección de valores atípicos de las distancias RoIs-FV. En particular se utilizó el método Z-score modificado, el cual indicó que valores que corresponden a *outliers* son los mayores a 9,28 en el caso de los resultados al variar la proximidad entre ventanas, y 9,38 al variar el tamaño de las ventanas. Con esta información se procedió a enmascarar los valores atípicos con la nueva distancia máxima de cada experimento (9,28 y 9,38), y a los valores "No comparables"se les asoció el valor máximo más 0,2 (9,48 y 9,58), para equilibrar entre el efecto que produce en la variabilidad observada y también eventualmente, distinguir su aparición.

Por último, en esta sección se decidió por no normalizar los resultados para mantener la relación de los valores con la distancia evaluada. En este caso, se utilizó la distancia manhattan sobre vectores con 28 características (ver Sección 4.3). Por lo anterior se desprende que una distancia RoIs-FV equivale a que la diferencia acumulada entre los valores de todas las características entre dos vectores es 1.

6.4.1. Análisis de la influencia de la proximidad entre las ventanas

En este experimento se evaluó la diferencia de los indicadores de variabilidad (ver Sección 4.4.1), luego de considerar la matriz de variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas obtenidas con el algoritmo RoIs-FV, donde cada grupo de variaciones es más riguroso que el anterior respecto de la cercanía entre las ventanas que se comparan.

Los gráficos de la Figura 6.9 muestran la silueta de la frecuencia relativa de cada indicador de variabilidad sobre la distancia RoIs-FV entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido a partir de la matriz de variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas. En este experimento se consideraron ventanas de una semana con un mínimo de 8 transacciones por ventana.

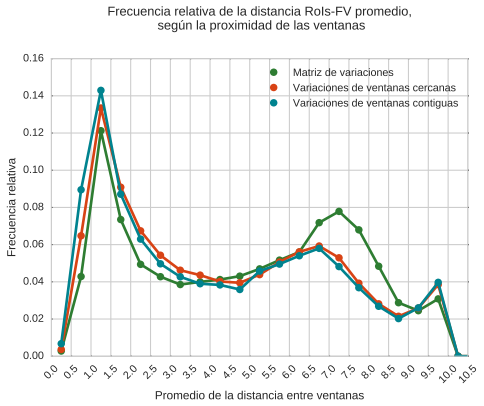
En la Figura 6.9a se observa la distribución del indicador promedio. La distribución del promedio de las tres variaciones evaluadas abarcan todo el espectro de similitud, todas con tres máximos locales. Por un lado se distingue una alta concentración de usuarios con variabilidad promedio en el intervalo 0,5-2,5, lo cual indica baja variabilidad del comportamiento. Por otro lado, se distinguen también dos concentraciones en el extremo derecho del espectro, lo cual indica alta variabilidad. En particular el máximo local de valores mayor a 9,5 indica la presencia de usuarios que cambian sus RoIs durante su periodo de actividad.

También en la Figura 6.9a se aprecia que si bien la forma de la curva de los tres tipos de variaciones es similar, la frecuencia del indicador promedio de la matriz de variaciones es mucho mayor en valores de alta variabilidad. En la misma línea, la frecuencia de las variaciones cercanas y variaciones contiguas es mayor en valores de baja variabilidad. Lo anterior indica que a mayor proximidad de las ventanas comparadas menor es la variabilidad percibida.

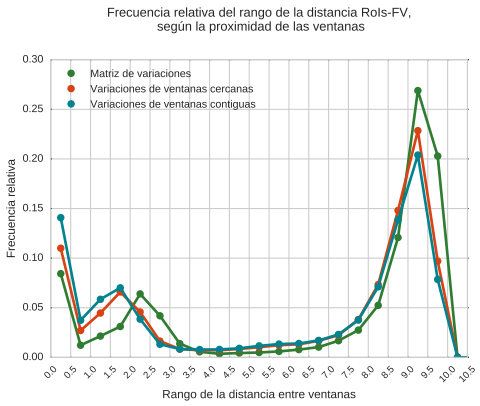
Los gráficos de la Figura 6.9 de indicadores promedio, rango, rango IQ y desviación estándar muestran distribuciones en las que destaca la diferencia entre dos grupos de compor-

tamientos: usuarios de baja variabilidad y usuarios de alta variabilidad. Resulta interesante observar la diferencia entre la distribución del rango, el rango IQ y el promedio. En la distribución del rango (Figura 6.9b) la mayor concentración de usuarios se encuentra en valores de alta variabilidad, a diferencia del promedio y del rango IQ donde la moda de las distribuciones se encuentra en valores de baja variabilidad. Lo anterior indica que gran parte de los usuarios que presenta baja variabilidad en algún momento de su periodo de actividad presenta también, alta variabilidad. Esto destaca la importancia de entender la movilidad humana como un concepto formado por componentes regulares y variables, más que tratar de clasificar a los usuarios en una de estas categorías.

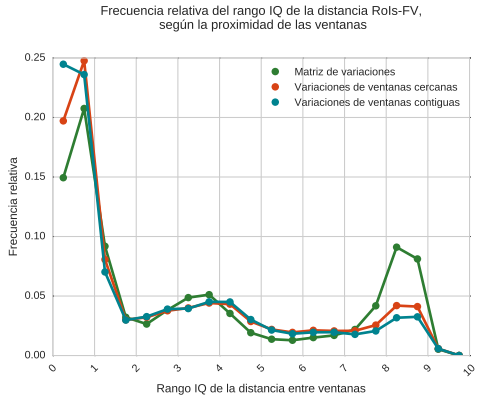
El coeficiente de variación es el único indicador que señala una mayor concentración de la matriz de variaciones en valores de baja dispersión de la variabilidad de los usuarios. Por el contrario, los indicadores rango, rango IQ y desviación estándar señalan mayor concentración de las variaciones de ventanas contiguas y cercanas en valores de baja dispersión. Esta aparente discrepancia se produce porque los valores del promedio de la matriz de variaciones es mayor, luego valores de mayor dispersión son normalizados en el coeficiente de Variación. Por lo anterior se concluye que si bien la dispersión relativa de la matriz de variaciones es menor, considerando que los tres tipos de variaciones analizadas pertenecen al mismo dominio de distancia, las variaciones de ventanas contiguas y cercanas exhiben menor variabilidad que la matriz de variaciones.



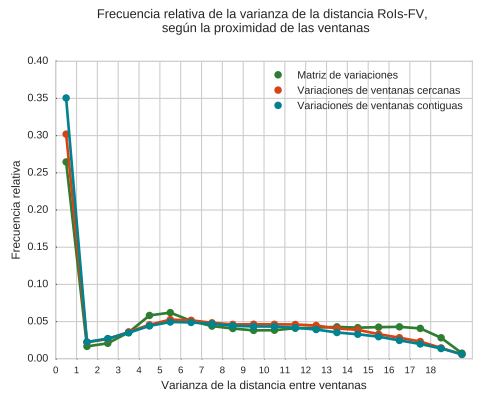
(a) Promedio



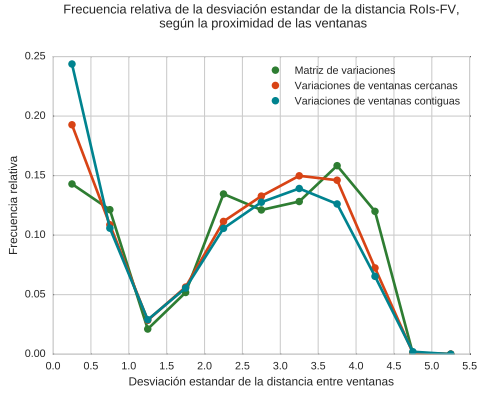
(b) Rango



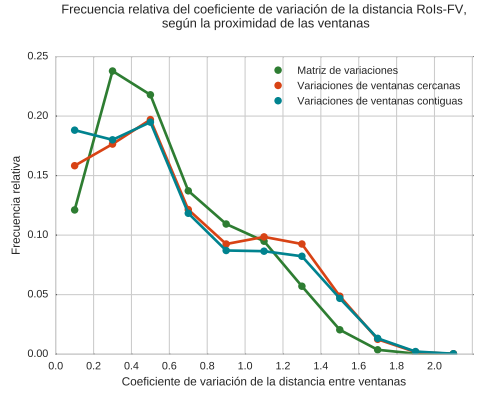
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.9: Los gráficos de esta figura muestran la silueta de la frecuencia relativa de cada indicador sobre la similitud RoIs-FV entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido al comparar las variaciones, las variaciones entre ventanas cercanas y las variaciones entre ventanas contiguas. Se consideraron ventanas de una semana con un mínimo de 8 transacciones por ventana.

La Tabla 6.5 muestra el promedio de las distribuciones de cada indicador para las variabilidades obtenidas con el algoritmo RoIs-FV, según la proximidad de las ventanas de tiempo. Se observa que el valor de todos los indicadores disminuye al incrementar la cercanía entre las ventanas comparadas. Se observa que la tendencia para el indicador coeficiente de variación

es de menor magnitud que en los otros indicadores, lo cual coincide con lo discutido previamente sobre la normalización de la dispersión de la variabilidad según el promedio observado. También destaca que la diferencia entre los resultados de las variaciones entre ventanas más cercanas y ventanas contiguas, varían según el indicador. Por ejemplo, para el promedio y el coeficiente de variación la diferencia entre estos dos tipos de variaciones es de mucha menor magnitud que la observada en el indicador rango y varianza. Lo anterior indica la presencia de valores atípicos en las variaciones de ventanas más cercanas que no se encuentran en las variaciones de ventanas contiguas. En general, la Tabla 6.1 confirma la tendencia observada previamente: a mayor cercanía de las ventanas mayor es la estabilidad de la variación del movimiento.

Para verificar la tendencia observada se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon, con un nivel de significación de 0,01. Se compararon las distribuciones del indicador promedio de las siguientes combinaciones:

1. la matriz de variaciones con las variaciones de ventanas cercanas
2. la matriz de variaciones con las variaciones de ventanas contiguas
3. las variaciones de ventanas cercanas con las variaciones de ventanas contiguas

Para la primera y tercera combinación, el p-valor resultó muy cercano a 0.0. Para la segunda combinación, el p-valor resultó $6{,}4e^{-7}$. Por lo anterior, se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. De este modo se confirma que la tendencia es significativa, por tanto se concluye que a mayor proximidad entre las ventanas comparadas menor es la variabilidad observada.

Tabla 6.5: Valores promedio de los indicadores de variabilidad de la distancia RoIs-FV según la proximidad de las ventanas.

Indicador	**Matriz variaciones**	**Ventanas cercanas**	**Ventanas contiguas**
Promedio	4,84	4,34	4,21
Rango	6,89	6,19	5,73
Rango IQ	3,44	2,76	2,52
Var	7,59	6,80	6,24
STD	2,38	2,20	2,05
CV	0,60	0,65	0,62

6.4.2. Análisis de la influencia del tamaño de las ventanas de tiempo

En este experimento se evaluó la diferencia de los indicadores de variabilidad obtenidos con el algoritmo RoIs-FV, según el tamaño de la ventana de tiempo en que se divide el periodo de actividad de cada usuario.

Los gráficos de la Figura 6.10 muestran la silueta de la frecuencia relativa de cada indicador de variabilidad sobre la distancia RoIs-FV entre las ventanas cercanas de cada usuario.

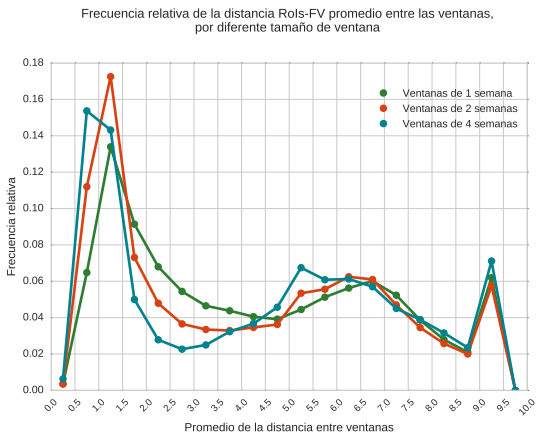
En cada gráfico se compara el indicador obtenido utilizando ventanas de una semana, dos semanas y cuatro semanas, con un mínimo de 8 transacciones por semana.

En los gráficos de la Figura 6.10 de indicadores distintos al promedio, se observa claramente un orden entre las tres distribuciones según el tamaño de la ventana. La distribución de las variaciones entre ventanas de una semana presentan mayor concentración al extremo derecho del espectro de cada indicador, es decir, valores de alta variabilidad. Por el contrario, las ventanas de cuatro semanas, muestran mayor concentración al extremo izquierdo del espectro, es decir, indican que gran parte de los usuarios presentan baja variabilidad al comparar su movilidad mes a mes. En la misma línea, se observa que no solo cambia la magnitud de la frecuencia, sino también el valor de la moda: Los indicadores rango, rango IQ y coeficiente de variación presentan modas más cercanas a 0 (valores de baja variabilidad) a mayor el tamaño de la ventana.

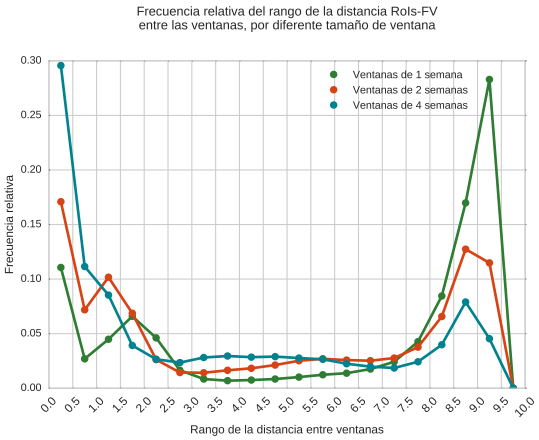
En el caso del promedio, la Figura 6.10a muestra que la moda de la distribución de las variabilidades con ventanas de una semana es menor a ventanas de 2 y 4 semanas, al igual que lo observado para los indicadores rango, rango IQ y coeficiente de variación. Es decir, confirma la tendencia de menor variabilidad observada a mayor tamaño de ventana. Sin embargo, para valores de máxima variabilidad (en el intervalo 0,9-0,95), se observa una frecuencia relativa mayor al utilizar ventanas de una semana. Lo anterior se explica por la relación entre el número de pares de ventanas que presentan cambios de RoIs versus el número de pares de ventanas comparados. Hay que tener en cuenta que al dividir el periodo de actividad en ventanas de cuatro semanas se comparan menos pares de ventanas que al dividir el periodo de actividad en ventanas de una semana. Por tanto, cuando un usuario cambia de RoIs (que en la mayoría de los casos corresponde a un fenómeno aislado), tiene un menor efecto en el promedio de la variabilidad medida semana a semana que en el promedio de la variabilidad mes a mes.

En el caso del rango, la Figura 6.10b muestra que las distribuciones de los tres tamaños de ventanas exhiben una importante concentración de usuarios con al menos un par de ventanas con alta variabilidad (rango mayor a 7), en particular las ventanas de una semana donde hay cerca de un 30 % de usuarios que presenta máxima variabilidad (cambio de RoIs). Sin embargo, de la Figura 6.4c se desprende que, independiente del tamaño de la ventana, la gran mayoría de los usuarios presentan un rango interquartil menor a 2. Por tanto, las concentraciones de usuarios cuya movilidad presenta un gran rango de variabilidad, se explica por la presencia de un número acotado de pares de ventanas con alta variabilidad. Es decir, se observa una baja variabilidad entre la mayoría de las ventanas de cada usuario, incluso al utilizar ventanas de una semana.

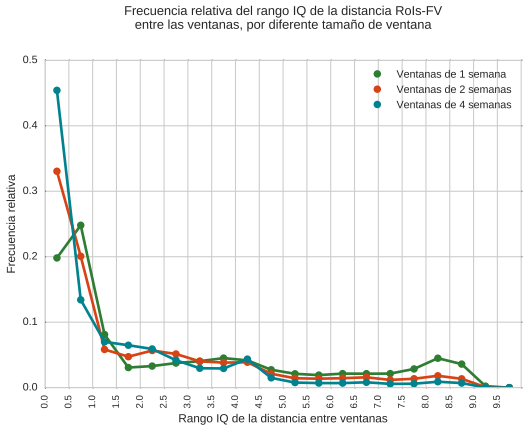
Los gráficos de la Figura 6.10 de los indicadores promedio, rango y desviación estándar, permiten diferenciar dos grupos de comportamiento: usuarios con alta regularidad y usuarios con alta dispersión en la variabilidad de su movilidad. De acuerdo a lo observado anteriormente, la alta dispersión corresponde a la presencia de un número más bien pequeño de pares de ventanas con alta variabilidad, en particular los casos en que los usuarios cambian de RoIs. Es por esto que los dos grupos de usuarios se deben entender como: usuarios que mantienen sus RoIs estables, y usuarios que presentan cambios de RoIs en algún momento de su periodo de actividad.



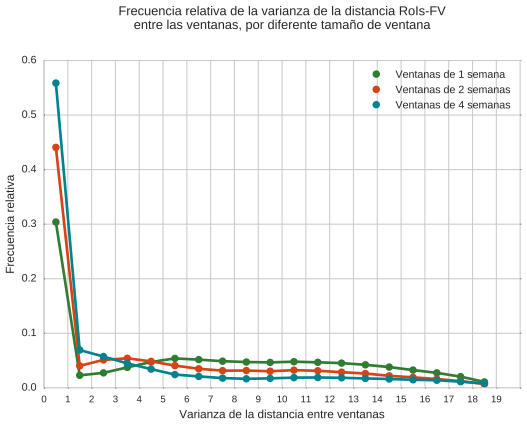
(a) Promedio



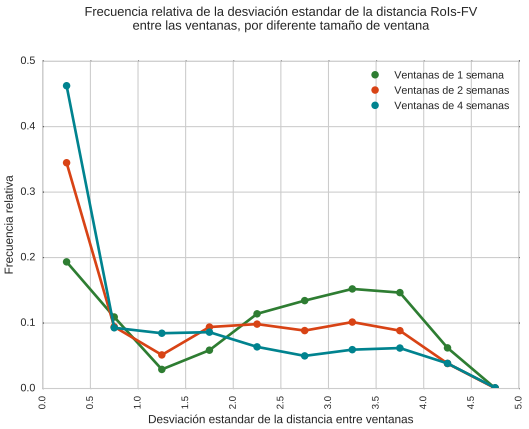
(b) Rango



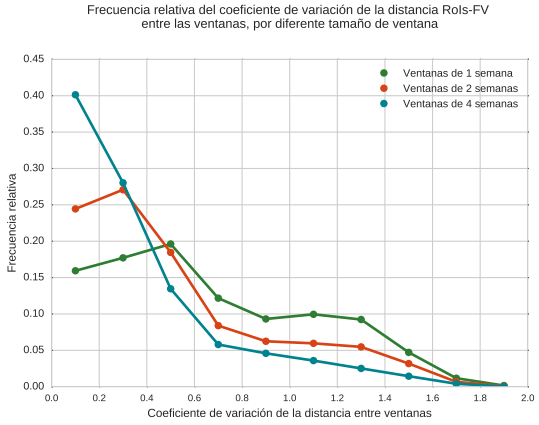
(c) Rango IQ



(d) Varianza



(e) Desviación estándar



(f) Coeficiente de variación

Figura 6.10: Los gráficos de esta figura muestran la silueta de la frecuencia relativa de cada indicador sobre la similitud RoIs-FV entre las ventanas de cada usuario. En cada gráfico se compara el indicador obtenido al comparar las variaciones entre ventanas cercanas al variar el tamaño de la ventana entre una semana, dos semanas y cuatro semanas.

La Tabla 6.6 muestra el promedio de las distribuciones de cada indicador de variabilidad utilizando la distancia RoIs-FV, según el tamaño de las ventanas de tiempo. Se observa

que el valor promedio de todos los indicadores distintos al promedio disminuye de manera consistente al incrementar el tamaño de la ventana. En el caso del promedio, el menor valor es alcanzado con ventanas de dos semanas, mientras que ventanas de una y cuatro semanas tienen valores similares. Lo anterior se explica por lo comentado anteriormente: con ventanas de una y cuatro semanas, se percibe un mayor número de cambios de RoIs, por tanto la distancia promedio entre las ventanas es mayor. Por lo anterior se concluye de la Tabla 6.6 que a pesar de que el promedio de la variación de la movilidad ventana a ventana es mayor al utilizar ventanas de dos semanas, la dispersión de las variaciones de la movilidad disminuye al aumentar el tamaño de una a cuatro semanas.

Como la tendencia del promedio no es clara, para evaluar si es significativa la diferencia entre las distribuciones de este indicador, se llevó a cabo el test estadístico prueba de los rangos con signo de Wilcoxon (con un nivel de significación de 0,01). Se compararon las distribuciones del indicador promedio de las siguientes combinaciones de tamaño de ventana:

1. ventanas de una semana con ventanas de dos semanas
2. ventanas de una semana con ventanas de cuatro semanas
3. ventanas de dos semanas con ventanas de cuatro semanas

En todos los casos anteriores el p-valor resultó 0.0 o muy cercano a 0, por lo que se rechaza la hipótesis de que los conjuntos de indicadores provengan de la misma distribución. Luego, se requiere más información (hacer experimentos con más usuarios o utilizar otros largos de ventanas) para poder concluir sobre el efecto del tamaño de la ventana en la estabilidad de la movilidad utilizando el algoritmo RoIs-FV.

Tabla 6.6: Valores promedio de los indicadores de variabilidad de la distancia RoIs-FV según el tamaño de las ventanas.

Indicador	**Una semana**	**Dos semanas**	**Cuatro semanas**
Promedio	4,31	4,13	4,32
Rango	6,12	4,53	3,19
Rango IQ	2,73	1,98	1,48
Var	6,66	4,61	3,36
STD	2,18	1,65	1,26
CV	0,65	0,50	0,35

6.4.3. Resultados de la variabilidad de los usuarios agregados por semana del periodo 2012-2013

La Figura 6.11 presenta el diagrama de caja de la distancia RoIs-FV entre la movilidad de cada usuario respecto la semana anterior, para cada semana del periodo 2012-2013. Cada diagrama de caja representa la distribución de las distancias entre ventanas de los usuarios. Las líneas rojas denotan la mediana y las cajas naranjas marcan el rango entre el segundo y tercer quartil. Los puntos celestes representan valores atípicos.

En la Figura 6.11 es posible observar que la mediana se encuentra en la mayoría de los casos entre 0,17 y 0,3, salvo semanas especiales. En particular, las semanas correspondientes a las fiestas de fin de año presentan un alza en el indicador de distancia, evidenciando que los usuarios cambian su comportamiento en el sistema de transporte en esta época del año.

Dejando de lado las semanas especiales, es posible observar que el algoritmo RoIs-FV distribuye las variaciones de los usuarios en todo el espectro de distancia. A diferencia de la distribución observada con los algoritmos TPM y EDM, el algoritmo RoIs-FV presenta una diferencia significativa en el rango de los valores menores a la mediana versus el rango de los valores mayores a la mediana. Esto evidencia que utilizando la distancia RoIs-FV, los usuarios se dividen en dos grupos: una alta concentración de usuarios extremadamente regulares, y el resto de usuarios distribuidos en un amplio rango de variabilidad.

Por último, se observa que durante el año 2013 el límite superior de los diagramas de caja se desplaza hacia la distancia máxima, lo cual indica que la dispersión de la movilidad de los usuarios en el año 2013 es mayor a la dispersión registrada en el año 2012.

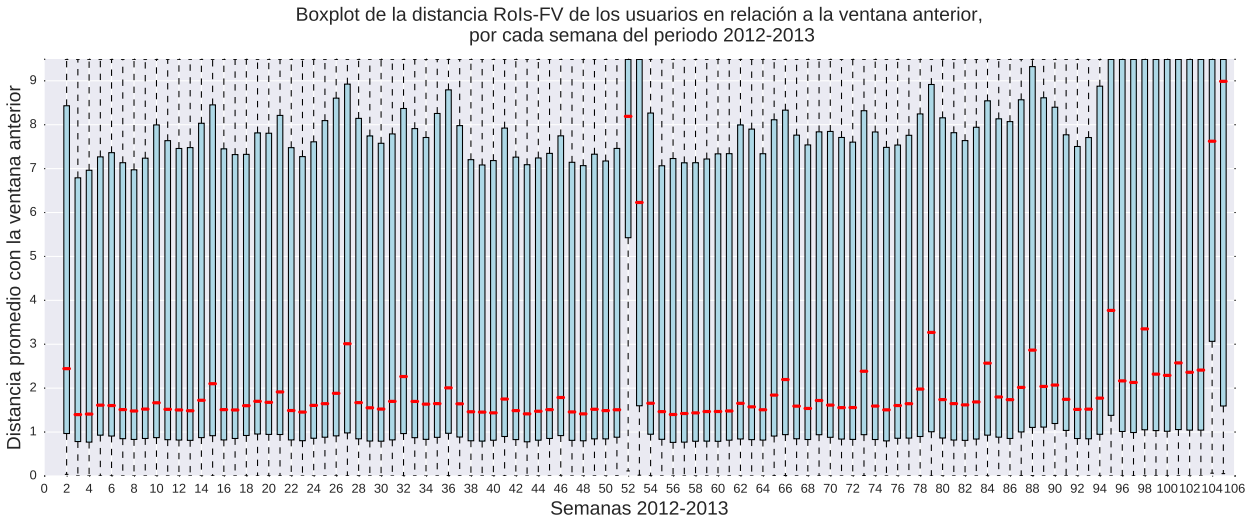


Figura 6.11: Diagrama de caja de la distancia RoIs-FV de los usuarios respecto a la semana anterior, por cada semana del periodo 2012-2013.

6.5. Resultados generales

En esta sección se realiza una comparación de los indicadores de variabilidad generados por los algoritmos TPM, EDM y RoIs-FV, y se discuten los resultados generales. Para comparar los resultados de los tres algoritmos, se procedió a normalizar los valores obtenidos con los algoritmos EDM y RoIs-FV. Sobre los resultados de ambos algoritmos se aplicó el algoritmo de detección de valores atípicos *Z Score modificado*. Luego, se enmascararon los valores atípicos utilizando el máximo sugerido por el algoritmo de detección de *outliers*. En el caso del algoritmo EDM, se definió 7000 el máximo de las variaciones observadas entre las ventanas de los usuarios de Gatineau. En el caso del algoritmo RoIs-FV, se definió el valor 9,58 como el máximo de las variaciones observadas entre las ventanas de los usuarios de

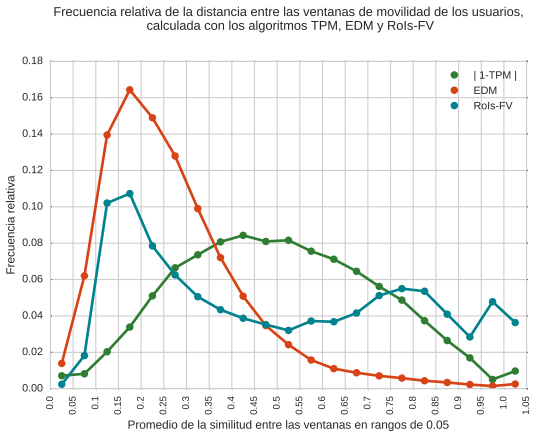
Gatineau. Finalmente, utilizando una normalización min-max, se transformaron los valores al rango [0,1].

La Figura 6.12 muestra dos gráficos que resumen la diferencia entre las distribuciones de la variabilidad de los usuarios utilizando los algoritmos TPM, EDM y RoIs-FV. Los resultados presentados corresponden a las variaciones entre ventanas cercanas, con ventanas de una semana y un mínimo de 8 transacciones por semana.

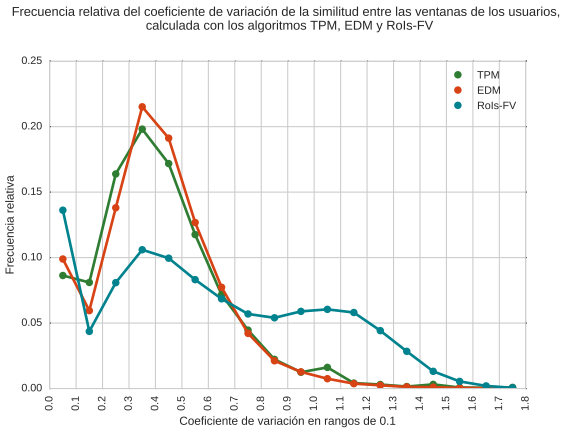
La Figura 6.12a presenta la distribución del promedio de la distancia entre las ventanas de los usuarios, según los tres algoritmos comparados. Para una mejor comprensión, la distribución del algoritmo TPM aparece invertida, de esta forma las tres distribuciones se interpretan como distancias: valores cercanos a 0 indican baja variabilidad y valores cercanos a 1 alta variabilidad. Es posible observar diferencias claras en las tres distribuciones, a pesar de corresponder al mismo grupo de usuarios. Lo anterior confirma la idea de que la variación de la movilidad percibida depende en gran medida de la métrica utilizada.

En la Figura 6.12a la distribución más distinguible es la del algoritmo RoIs-FV, ya que exhibe tres máximos locales que corresponden a concentraciones de usuarios con distintos grados de variabilidad. Esto resulta particularmente útil en estudios que requieren extraer muestras de usuarios regulares. Por su parte, los algoritmos TPM y EDM presentan distribuciones con solo una moda, sin embargo con concentraciones muy diferentes. El algoritmo EDM concentra a la mayoría de los usuarios en el extremo izquierdo del espectro (baja variabilidad), mientras que el algoritmo TPM indica que la variabilidad de los usuarios se distribuye por todo el espectro. Además, el algoritmo TPM presenta un leve aumento al acercarse a valores de alta variabilidad, indicando la presencia de usuarios que cambian su movilidad drásticamente. La diferencia entre los algoritmos TPM y EDM es interesante, ya que ambos algoritmos corresponden a medidas netamente espaciales, salvo que el algoritmo TPM utiliza la variación entre los nombres de los pares origen-destino, mientras que el algoritmo EDM utiliza la distancia geográfica. Se concluye que la distancia geográfica entre las trayectorias homogeneiza la variación de los usuarios al no advertir cambios en los medios de viaje.

La Figura 6.12b muestra la distribución del coeficiente de variación de la distancia entre las ventanas de los usuarios, según los tres algoritmos comparados. Los tres algoritmos presentan un máximo local en 0, es decir evidencian la presencia de usuarios cuya variación de la movilidad es completamente constante. Los algoritmos TPM y EDM muestran distribuciones similares, donde la gran mayoría de los usuarios presentan coeficientes de variación menor que 0,7, lo cual indica que en general los usuarios muestran baja dispersión en la variabilidad de su movilidad. Por su parte, la distribución del algoritmo RoIs-FV muestra dos máximos locales menores a 0,7, esto señala que la mayoría de los usuarios poseen una baja dispersión en la variabilidad del comportamiento en transporte público. Sin embargo, la distribución del algoritmo RoIs-FV también presenta una concentración no menor de usuarios con coeficiente de variación mayor a 1, lo cual indica alta dispersión en la variabilidad de la movilidad.



(a) Similitud promedio



(b) Coeficiente de variación

Figura 6.12: Los gráficos (a) y (b) de esta figura muestran la silueta de la frecuencia relativa del indicador promedio y coeficiente de variación sobre la similitud entre las ventanas de cada usuario obtenidas con los tres algoritmos.

La Figura 6.13 presenta tres mapas de calor que representan la frecuencia de los usuarios según su similitud entre ventanas y el porcentaje de ventanas con las transacciones suficientes para ser comparadas. Se realizó esta intersección de variables para ilustrar si es que existe algún rango de variabilidad de la movilidad asociado a la frecuencia de los usuarios al utilizar transporte público. Los tres mapas de calor presentados corresponden al promedio de las variaciones entre ventanas cercanas de cada usuario, obtenidas con los algoritmos TPM, EDM y RoIs-FV, utilizando ventanas de una semana con un mínimo de 8 transacciones por semana. En cada mapa de calor el eje X corresponde al porcentaje de ventanas con más de 8 transacciones; y el eje Y corresponde a la similitud (TPM) o distancia (EDM y RoIs-FV) obtenida por cada algoritmo. A continuación se describe cada mapa de calor:

La Figura 6.13a muestra el mapa de calor correspondiente a graficar la similitud TPM. En primer lugar, se observa que la frecuencia de usuarios en cada celda se encuentra menos centralizada que en los mapas de calor de los algoritmos EDM y RoIs-FV. No obstante, la Figura muestra una concentración evidente de usuarios con porcentajes de ventanas comparables en el rango [50 %-90 %], donde se aprecia un aumento de la frecuencia de usuarios según los valores de similitud en el intervalo [0,3-0,8]. También se observa la ausencia de usuarios extremadamente regulares para cualquier porcentaje de ventanas comparables. En cambio para valores extremadamente irregulares (tramo [0,0-0,1]) se observa una mayor concentración en ventanas con menor porcentaje de ventanas comparables. Es decir se observa una relación entre baja frecuencia de viajes y poca regularidad de la movilidad, pero solo para usuarios extremadamente irregulares.

La Figura 6.13b muestra el mapa de calor correspondiente a graficar la distancia EDM. Lo primero que llama la atención es la ausencia de usuarios con similitudes mayores a 0,5. Del mismo modo, se percibe una falta de usuarios con regularidad extrema (en el tramo [0,0-0,1]). Por último, la alta concentración de usuarios entre el rango de distancia [0,1-0,4] exhibe un aumento de la frecuencia de los usuarios según el porcentaje de ventanas comparables. Sin embargo esta tendencia es propia de la distribución del porcentaje de ventanas, por lo que no resulta concluyente.

La Figura 6.13c muestra el mapa de calor correspondiente a graficar la distancia RoIs-FV. Nuevamente se observa una ausencia de usuarios extremadamente regulares. Por otra parte, la distribución de los usuarios en el rango de distancia [0,1-0,4] exhibe un comportamiento similar al observado en la distancia EDM, es decir en aquel rango de distancia se observa que a mayor porcentaje de ventanas comparables mayor la frecuencia de los usuarios. Sin embargo, a diferencia del mapa de calor del algoritmo EDM, se observan otras dos zonas de concentración de usuarios de menor magnitud. La primera de estas zonas corresponde a usuarios con más de 50 % de ventanas comparables que muestran distancias en el rango [0,6-0,9]. La segunda de estas zonas es similar a una concentración observada en el algoritmo TPM, que corresponde a usuarios de alta variabilidad (valores de distancia cercanos a 1) en usuarios con bajo porcentaje de ventanas comparables (menor a 40 %). Finalmente en este gráfico se observa que para usuarios con un porcentaje de ventanas comparables menor al 10 % los resultados de la distancia se vuelven binarios: o extremadamente regular o extremadamente variable. Lo anterior confirma la idea de que los usuarios con distancias promedios con valores intermedios son producto de la ponderación de una mayoría de ventanas regulares con una pequeño conjunto de ventanas extremadamente irregular (principalmente casos con cambio de RoIs). Para confirmar esta observación se procede a analizar la variaciones obtenidas, no como promedios de cada usuario, sino que desagregados entre los pares de ventanas.

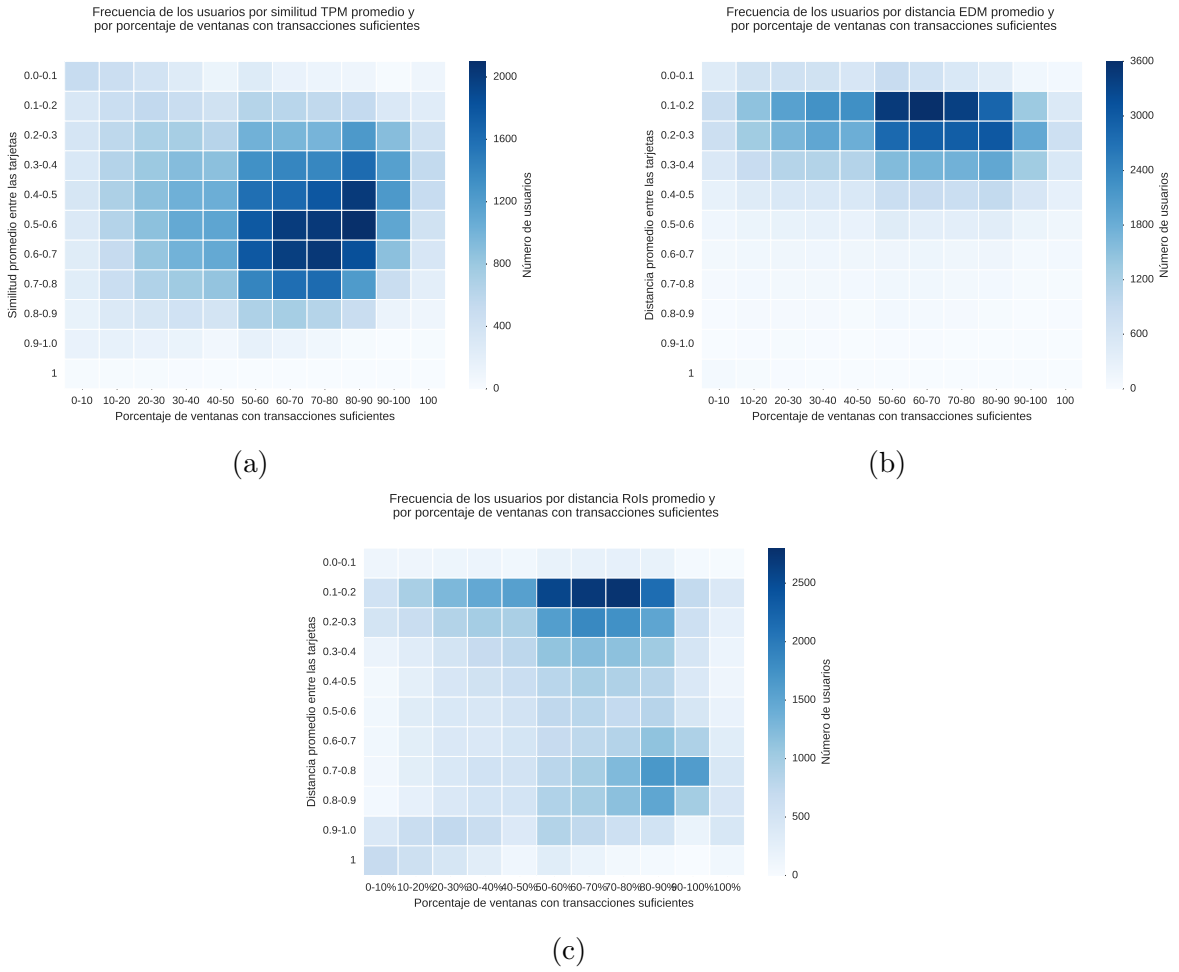


Figura 6.13: Los gráficos de esta figura muestran la frecuencia de los usuarios según similitud entre ventanas y porcentaje de ventanas con las transacciones suficientes para ser comparadas. Los gráficos (a),(b) y (c) utilizan la similitud TPM, EDM y RoIs-FV respectivamente.

La Figura 6.14 muestra tres mapas de calor que representan la frecuencia de los pares de ventanas comparados según la variación de la movilidad y el número de transacciones promedio. Los tres mapas de calor presentados corresponden a las distintas variaciones entre cada par de ventanas cercanas de cada usuarios, obtenidas con los algoritmos TPM, EDM y RoIs-FV, utilizando ventanas de una semana con un mínimo de 8 transacciones por semana. Es decir, cada celda almacena el número de pares de ventanas cuya diferencia de movilidad pertenece al rango de cada fila, y donde el promedio entre las transacciones entre cada par de ventana pertenece al indicado en cada columna. En cada mapa de calor el eje X corresponde al número de transacciones promedio por par de ventana; y el eje Y corresponde a la similitud (TPM) o distancia (EDM y RoIs-FV) obtenidas por cada algoritmo. A continuación se describe cada mapa de calor:

La Figura 6.14a ilustra el mapa de calor correspondiente al algoritmo TPM. Es posible observar que en general la frecuencia de los pares de ventanas disminuye al aumentar la distancia entre la movilidad de los pares de ventana y al aumentar el número de transacciones. La moda de la frecuencia se observa en el rango de similitud [0,7-0,8] para pares de ventanas

con 9 transacciones en promedio. Si bien se observa que las similitudes se distribuyen en todo el espectro de similitud, se observa la falta de pares de ventana con más de 10 transacciones que presenten alta similitud. La menor variabilidad observada entre ventanas de 8 a 10 transacciones se debe, probablemente, a que dichas ventanas corresponden a semanas con viajes exclusivos entre hogar y trabajo u hogar y estudio, es decir comportamientos extremadamente regulares. En el mismo sentido, la ausencia de ventanas con alta similitud y mayor número de viajes, se explica porque semanas con más de 10 transacciones en general involucran la presencia de viajes recreativos o extraordinarios, por tanto comportamientos más variables. La Figura 6.14b muestra el mapa de calor correspondiente al algoritmo EDM. Del mismo modo que en el mapa del algoritmo TPM, la frecuencia de los pares de ventanas disminuye al aumentar la distancia entre la movilidad de los pares de ventana y al aumentar el número de transacciones. En particular, se percibe la existencia de pares de ventanas con extrema similitud entre la movilidad (distancia menor a 0,1). También se observa que la moda se encuentra en valores de distancia en el rango [0,1-0,2] para pares de ventanas con 9 transacciones. La Figura 6.14c grafica el mapa de calor correspondiente al algoritmo RoIs-FV. En esta figura se observa la distribución decreciente de los usuarios según número de transacciones. La moda se observa en valores extremadamente regulares (distancia menor a 0,1) para pares de ventanas con 9 transacciones en promedio. Lo importante de este gráfico es que muestra que las distancias RoIs-FV promedio de los usuarios se encuentran efectivamente formadas por una combinación de pares de ventanas extremadamente regulares y en menor medida por pares de ventanas donde se perciben cambios de RoIs.

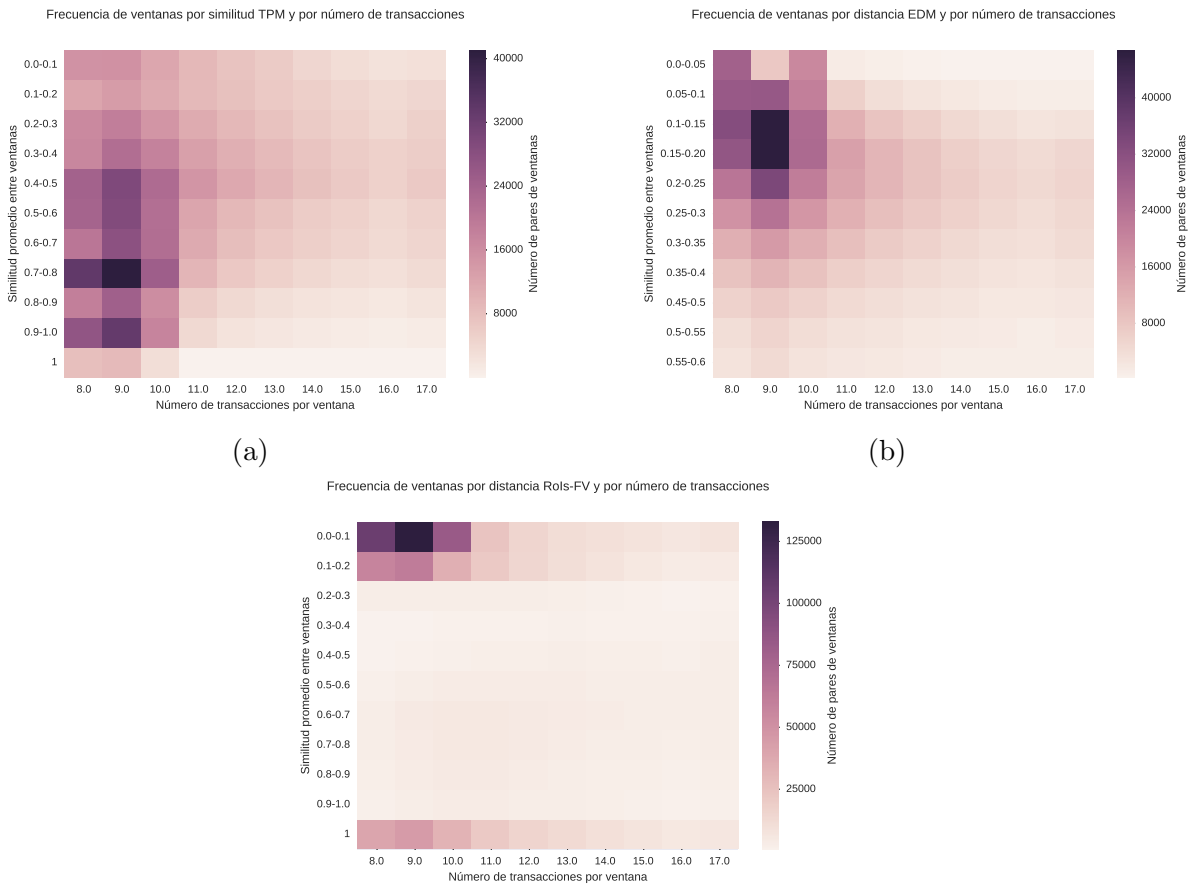


Figura 6.14: Los gráficos de esta figura muestran la frecuencia de las ventanas según similitud y número de transacciones. Los gráficos (a),(b) y (c) utilizan la similitud TPM, EDM y RoIs-FV respectivamente.

Resulta interesante observar los cambios de RoIs de los usuarios, ya que permiten separar los cambios drásticos de los cambios cotidianos en la movilidad de los usuarios. La Figura 6.15 muestra la frecuencia del número de usuarios según el número de pares de ventanas del periodo de actividad en los que el usuario no comparte el mínimo de dos RoIs. Considerando que la mayoría de los usuarios tiene 2 RoIs, la Figura 6.15 en la mayoría de los casos corresponde al número de veces que el usuario cambia uno o dos de sus zonas de actividad más importantes. En la figura se observa que la mayor frecuencia corresponde a usuarios que no cambian de RoIs durante su periodo de actividad, lo cual se condice con la idea general de que las RoIs corresponden a las zonas relacionadas al hogar y trabajo, las cuales se esperan sean estables en la mayoría de los usuarios. Por otro lado, el porcentaje de usuarios con cambios de RoIs resulta significativo y mayor de lo esperado. Diversos motivos pueden explicar estos resultados:

- Cambios efectivos en la movilidad de los usuarios.
- El radio de 500 metros que define a cada RoI, puede estar subestimando la distancia caminada por los usuarios, provocando que usuarios que utilizan paraderos de buses levemente más lejanos que el radio que determina su zona de actividad principal, se

malinterpreten como cambios de RoIs.

- La presencia de usuarios con viajes multimodales puede aumentar los cambios de RoIs percibidos, ya que las RoIs podrían indicar lugares de transferencias en vez de zonas donde se desempeñan actividades. Luego, un cambio de ruta podría interpretarse como un cambio de RoIs.
- El porcentaje mínimo de transacciones que almacenan las RoIs (70 %) puede no ser el indicado para todos los usuarios. Por ejemplo, usuarios que realizan muchas transacciones podrían tener RoIs que no representan lugares de importancia, luego estos lugares podrían no volver a ser visitados.

Por lo anterior, como trabajo futuro se estima conveniente realizar una validación de las RoIs con datos provenientes de encuestas a usuarios de transporte público.

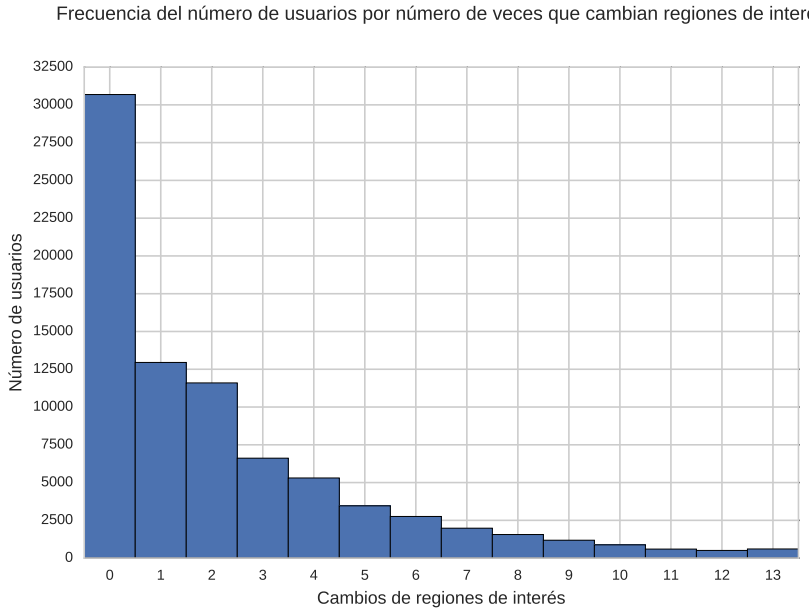


Figura 6.15: Histograma del número de usuarios por número de ocasiones en la que cambia de RoIs.

La Figura 6.16 presenta cómo varía el promedio de la variación semanal de la movilidad de todos los usuarios del sistema de transporte público de Gatineau, según el algoritmo TPM, EDM y RoIs-FV. A pesar de que los valores resultantes de los tres algoritmos están separados, la forma de las curvas es similar, lo cual indica que los tres algoritmos capturan en mayor o menor medida, las mismas variaciones. Es posible observar que la Figura 6.16 muestra tres máximos locales notorios en las semanas 52, 95 y 104. Los máximos de las semanas 52 y 104 corresponden a cambios en la movilidad asociados a las fiestas de fin de año. El tercer máximo, en la semana 95, corresponde a los cambios producidos por la puesta en marcha de un nuevo sistema de corredores (Rapibus), el 19 de octubre del 2013. El impacto del nuevo corredor en el comportamiento de los usuarios es claramente percibido por el algoritmo TPM, ya que los usuarios cambian las paradas de origen y/o destino de sus viajes. Por el contrario, el algoritmo EDM no muestra una variación importante, producto de que la variación se produce en el modo de viaje más que en el viaje en sí mismo.

Se puede observar en la Figura 6.16 que las tres curvas tienen una forma similar y que están separadas en casi todo el periodo de observación. La mayor diferencia entre el comportamiento de las curvas es observado una semana luego de la implementación de Rapibus. Por un lado, los algoritmos TPM y EDM muestran una variación promedio restaurada, lo que significa que los usuarios se adaptaron a una nueva forma de viaje. Y por otro lado, el algoritmo RoIs-FV parece mantenerse en un nivel de variación mayor en comparación al observado previo a la implementación del corredor, para luego alzarse de nuevo a finales del año 2013. Lo anterior muestra que existe un grado de variabilidad en el comportamiento de los usuarios que permanece semanas después de la implementación del corredor.

El análisis de la variación del uso del sistema desde la perspectiva de los usuarios puede ser útil para medir la capacidad de adaptación de los usuarios frente a cambios en el sistema de transporte público. Este tipo de análisis tiene el potencial de realizarse sobre diferentes zonas, periodos de tiempo (días, semanas, meses, etc.), y sobre determinados grupos de usuarios.

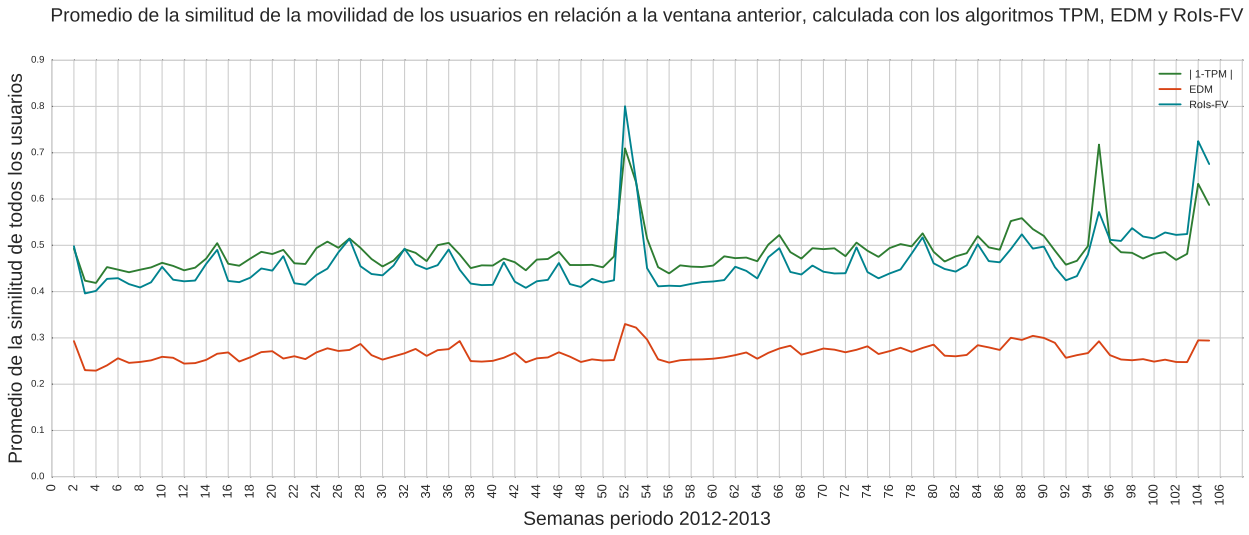


Figura 6.16: Promedio de la similitud de la movilidad de los usuarios en relación a la semana anterior durante cada semana del periodo 2012-201. Se compara la similitud obtenida con los algoritmos TPM, EDM y RoIs-FV.

En conclusión, se observa que las características del algoritmo EDM no lo hacen idóneo para medir la variabilidad de los usuarios de transporte. Debido a la baja sensibilidad a las variaciones, el algoritmo EDM resulta útil en la búsqueda por similitud de trayectorias de diferentes usuarios. Los algoritmos TPM y RoIs-FV generan mediciones de la variabilidad de la movilidad que promueven un mayor entendimiento del comportamiento en transporte público. En particular, el algoritmo TPM permite medir incluso leves variaciones en la forma de viaje, por lo que resulta útil en la medición de la adaptabilidad de los usuarios a cambios en el sistema de transporte. Por su parte, el algoritmo RoIs-FV permite medir variaciones en distintas dimensiones de la movilidad y provee una división entre usuarios regulares y poco regulares que resulta especialmente útil en estudios del comportamiento de los usuarios.

Capítulo 7

Resumen y conclusiones

Este capítulo está organizado en tres secciones: La Sección 7.1 presenta un resumen de los hallazgos y conclusiones del trabajo realizado. En la Sección 7.2 se presentan las conclusiones generales y las líneas de trabajo futuro. En la Sección 7.3 se identifican las limitaciones de esta tesis. Finalmente, en la Sección 7.4 se proponen recomendaciones para futuras aplicaciones del trabajo desarrollado.

7.1. Resumen y hallazgos

En el presente trabajo se han implementado y evaluado tres algoritmos para caracterizar y comparar registros de movilidad, con el objetivo de medir cuan reconocible es la movilidad de los usuario de transporte público. Uno de estos algoritmos ha sido diseñado y calibrado en el desarrollo de esta tesis, utilizando como base la literatura preexistente sobre caracterización de usuarios de transporte público. Los tres algoritmos implementados extraen y comparan distintos aspectos del comportamiento de los usuarios.

Se han empleado dos enfoques para determinar si es posible reconocer usuarios mediante su movilidad: medir la variabilidad del comportamiento de los usuarios y medir la capacidad de distinguir este comportamiento. El primer enfoque fue realizado sobre los registros de transacciones de los más de 100.000 usuarios del sistema de transporte público de Gatineau, Canadá. El segundo enfoque fue desarrollado sobre una muestra aleatoria de 5.000 usuarios del sistema de transporte público de Santiago de Chile.

Los algoritmos utilizados en esta investigación son los siguientes: un algoritmo propuesto por De Mulder et al., el cual describe la movilidad mediante una matriz de probabilidad de transición entre pares origen-destino; un algoritmo propuesto por Yuan y Raubal, el cual utiliza la trayectoria geográfica de cada usuario. Y finalmente el algoritmo RoIs-FV, desarrollado en esta tesis, el cual extrae las principales zonas de actividad de cada usuario y un vector de características (espaciales, temporales, entre otras) que describe su comportamiento en el sistema de transporte público.

Los tres algoritmos mencionados se utilizaron para medir la similitud entre la movilidad almacenada en distintas ventanas temporales de registros de movilidad de los usuarios. Luego, se extrajeron distintos indicadores de variabilidad sobre las similitudes medidas a lo largo del periodo de actividad de cada usuario. Se realizó un análisis de sensibilidad de la distribución de los indicadores de variabilidad de los usuarios, ajustando los siguientes parámetros: la cercanía entre las ventanas temporales, el tamaño de la ventana y el nivel de agregación espacial de los registros de movilidad.

Los tres algoritmos implementados también fueron evaluados y comparados de acuerdo a su capacidad de reconocer usuarios mediante su movilidad. Para esto se midió la capacidad de emparejar tarjetas provenientes de dos conjuntos de transacciones de cortes temporales independientes. Se analizó la influencia de diferentes factores según cada algoritmo, con especial énfasis en la calibración del algoritmo RoIs-FV.

Los principales resultados del trabajo realizado se resumen a continuación.

7.1.1. Variabilidad de los usuarios del sistema de transporte público de Gatineau

Los resultados de este experimento indican que los tres algoritmos evaluados obtienen grados de variabilidad diferentes para los usuarios del transporte público de Gatineau. Por un lado, los tres algoritmos presentan grandes diferencias en la distribución de la variabilidad individual promedio. Por otro lado, al analizar la variabilidad del comportamiento de todos los usuarios durante el periodo 2012-2013, se observan similitudes en la tendencia de las variaciones semana a semana de los tres algoritmos, sin embargo, la magnitud de las variaciones de cada algoritmo es notoriamente diferente. Estos hallazgos refuerzan la idea discutida por Jones y Clarke, quienes advirtieron que los resultados de cualquier métrica que mida la variabilidad se enfocará en ciertos aspectos de la movilidad, lo que determinará el tipo de regularidad observada. Por lo anterior, resulta necesario destacar el efecto de cada método sobre la variabilidad percibida:

El algoritmo TPM ha mostrado una gran sensibilidad a los cambios en el comportamiento de viajes en transporte público. El promedio de la similitud TPM de la movilidad de los usuarios tiene una distribución que abarca todo el espectro de similitud y mediante la cual se infiere que la mayoría de los usuarios conserva gran parte de sus viajes estables en el tiempo. La alta sensibilidad del algoritmo hace que se perciba una falta de usuarios extremadamente regulares. También provoca que usuarios con alto porcentaje de viajes nuevos semana a semana, sean evaluados como altamente irregulares, a pesar de mantener una componente de viajes altamente regular en el tiempo.

El algoritmo EDM es el método que presenta menor sensibilidad a variaciones en el comportamiento de los usuarios. La causa más probable de esta observación, es que aun considerando los cambios en el modo de viaje, la mayor parte del tiempo los usuarios viajan entre las mismas zonas geográficas. El promedio de la distancia EDM entre las ventanas de movilidad de los usuarios se concentra en los valores asociados a baja distancia, interpretados como baja variabilidad.

El algoritmo RoIs-FV realiza una distinción entre usuarios extremadamente regulares y el resto de los usuarios. Esta diferencia resulta particularmente útil en estudios del comportamiento en transporte público. Por ejemplo, se podría evaluar si esta diferencia se mantiene para distintos tipos de usuarios, como estudiantes, adultos o adultos mayores. Por otro lado, la extracción de las zonas de mayor actividad de los usuarios permite detectar cuándo un usuario cambia alguna de las zonas más importantes, generalmente asociadas a hogar y trabajo.

La diferencia de los resultados entre el algoritmo TPM y EDM resulta particularmente interesante, considerando que ambos comparan aspectos de la movilidad netamente espaciales. Según sus respectivos resultados, se postula que las métricas de distancia geográfica son menos sensibles a los cambios en la movilidad, que las distancias basadas en nombres de lugares visitados. Por consiguiente, la distancia geográfica resulta menos adecuada para medir la variación del comportamiento en transporte público.

En relación a los factores que determinan el grado de variación percibido, se observan las siguientes tendencias transversales:

- A mayor cercanía entre las ventanas de tiempo comparadas mayor similitud entre la movilidad almacenada en las ventanas
- Considerando ventanas de tamaño una, dos y cuatro semanas: A mayor tamaño de ventana, mayor la similitud entre la movilidad almacenada en las ventanas

También se observó que para el caso del algoritmo TPM, la agregación de los datos espaciales aumenta la similitud de la movilidad promedio de los usuarios. De esta forma, es posible adaptar el algoritmo TPM para asociar viajes homólogos, y así considerar variaciones provocadas solo por viajes nuevos y semánticamente diferentes.

En términos generales, los tres algoritmos coinciden en que la mayoría de los usuarios presenta comportamientos recurrentes. Lo anterior resulta esperable considerando la evidencia sobre la regularidad de la movilidad humana, expuesta en la Sección 2.1.2. Aun así, este trabajo provee evidencia de que, a pesar de la baja frecuencia de registros, los datos de transporte público permiten extraer perfiles de movilidad que capturan las relación entre comportamientos regulares y variables de la movilidad individual.

7.1.2. Reconocimiento de usuarios mediante la observación de la movilidad en transporte público

Al comparar los conjuntos de tarjetas reconocidas y no reconocidas por los tres algoritmos, se encontró que las dos intersecciones más grandes corresponden a los conjuntos donde los tres algoritmos coincidían, es decir: al conjunto de usuarios reconocidos y no reconocidos por los tres algoritmos. Lo anterior indica la existencia de un porcentaje no menor de usuarios con comportamientos distinguibles, capaces de describir la unicidad del comportamiento humano. En este mismo sentido, se demuestra que los tres algoritmos, en mayor o menor medida, son capaces de extraer, representar y distinguir la movilidad de los usuarios. Por otro lado, también permite cuantificar los usuarios cuya movilidad cambia de tal manera en el intervalo

entre Abril del 2013 y Septiembre del 2013, que no es posible reconocerlos por ninguno de los tres métodos; indicando así, la posibilidad de que aquellas tarjetas hayan cambiado de usuario portador.

En comparación a los métodos presentados en la literatura (ver Sección 2.3), los tres algoritmos analizados en este trabajo presentan rendimientos regulares considerando las diferencias entre las características de las bases de datos utilizadas. Por ejemplo, Naini et al. utilizó registros de llamadas telefónicas, y obtuvo una tasa de identificación del 78 % al emparejar la movilidad de 1.000 usuarios, esta tasa disminuyó al 21 % al emparejar 50.000 usuarios. Luego, el rendimiento de los tres algoritmos evaluados en este trabajo es comparable con el trabajo de Naini et al.. Lo mismo ocurre al comparar con el trabajo de De Mulder et al. y Gambs et al.. No se puede realizar una comparación cuantitativa más exhaustiva, ya que las variaciones del número de usuarios, el número de registros por usuario y el número de posiciones que componen los registros, afectan la tasa de identificación. Sin embargo, de acuerdo con los resultados obtenidos es posible argumentar que a pesar de las diferencias entre registros de telefonía, registros de dispositivos GPS y registros de tarjetas inteligentes, el porcentaje de usuarios que puede ser reconocido por su movilidad es similar. En particular, en caso de que las bases de datos sean realmente anónimas y de datos temporalmente independientes, resulta imposible distinguir los usuarios reconocidos correctamente de aquellos emparejados incorrectamente.

A diferencia de los trabajos encontrados en la literatura, en esta investigación se realizó una búsqueda de umbrales de similitud mínimos, con el fin de reducir la tasa de error, a través de la abstención del emparejamiento de tarjetas, en casos donde los posibles *match* son de baja similitud. De esta forma, se logró aumentar la relación entre la tasa de identificación y la tasa de error. Finalmente, la decisión de seleccionar el mejor algoritmo de reconocimiento dependería de la importancia que se le asigne al porcentaje de usuarios reconocidos versus el porcentaje de usuarios incorrectamente emparejados.

En particular, los resultados asociados a cada algoritmo indican lo siguiente:

El algoritmo TPM permite reconocer a un 66,72 % de los usuarios, y utilizando un umbral de mínima similitud de 0,65, se obtiene una razón entre identificados y no identificados de 7.68. El indicador de similitud resultante de este algoritmo es altamente sensible a cambios en la movilidad, lo cual permite diferenciar y reconocer a la mayoría de los usuarios. Por otro lado, su sensibilidad a las variaciones de la movilidad de un mismo usuario, provoca que en algunos usuarios se observe un bajo grado de similitud consigo mismo. Por lo anterior, es incierto su rendimiento como validador de que una tarjeta este siendo utilizada por el mismo portador.

El algoritmo EDM presenta una tasa de reconocimiento del 40,01 %, y al utilizar un umbral de máxima distancia, la tasa de error disminuye de un 58,99 % a un 28,96 %, mientras que la tasa de identificación solo disminuye a 32,88 %. El indicador de distancia del algoritmo EDM genera una homogenización de los usuarios. Este fenómeno se percibe a través de la alta tasa de error en el reconocimiento, a pesar de la alta similitud entre un usuario consigo mismo. Esta característica lo hace potencialmente eficaz para búsqueda de usuarios con movilidades afines.

RoIs-FV tiene un rendimiento regular, superior al algoritmo EDM, e inferior al algoritmo TPM. El algoritmo RoIs-FV muestra una tasa de reconocimiento del 51.94 % con una tasa de abstención del 9,48 %., Por otra parte, al exigir un umbral de 0,3 como máxima distancia, se obtiene una identificación del 5,04 % junto con una tasa de abstención del 94,58 %. De los resultados de este algoritmo se observa que el mecanismo de detección de Regiones de Interés puede mejorar; ya que existe un grupo de usuarios que no comparten RoIs, pero si son reconocidos por los otros algoritmos. Por último, a pesar de no tener un sobresaliente resultado reconociendo usuarios, este algoritmo presenta especial potencial en validar la constancia del comportamiento de los usuarios en el tiempo.

7.2. Conclusiones generales y líneas de trabajo futuro

Los algoritmos evaluados reportan diferentes resultados en la medición de la variabilidad de los usuarios y en la capacidad de distinguir usuarios según su movilidad. En relación a los objetivos de esta tesis, el algoritmo TPM y el algoritmo propuesto, RoIs-FV, presentan mejores resultados que el algoritmo EDM.

La variabilidad de los usuarios del sistema de transporte público de Gatineau no es homogénea. No obstante, la mayoría de los usuarios presenta una movilidad suficientemente regular en el tiempo, lo cual sugiere que los algoritmos permitirían extraer patrones temporalmente estables con el objetivo de verificar que una tarjeta está siendo utilizada por un mismo usuario.

Respecto a la capacidad de distinguir a los usuarios mediante la comparación de la movilidad en la red de transporte público, se concluye que la mayoría de los usuarios puede ser reconocido. Sin embargo, la tasa de reconocimiento conlleva una alta tasa de error. Esta tasa de error se explica principalmente por cambios en el comportamiento y por la presencia de usuarios con movilidad poco distinguible. En el caso de Santiago, donde las tarjetas no necesariamente están asociadas a un mismo usuario (por renovación u otros motivos), no es posible distinguir un usuario reconocido de un incorrectamente reconocido. Por lo anterior, se sugiere que el trabajo futuro debiese estar enfocado en detectar cambios en la movilidad que indiquen la posibilidad de que una tarjeta esté siendo utilizada por otro usuario, es decir, cambiar el paradigma utilizado en esta tesis, y asumir a priori que una tarjeta pertenece a un mismo usuario hasta que se demuestre lo contrario.

Más allá de los objetivos de este trabajo, los resultados mostraron un gran potencial relacionado a la minería de patrones de la variación del comportamiento en transporte público. Por ejemplo, se expuso como la medición de la variabilidad individual puede ser utilizada para el análisis de la adaptabilidad de los usuarios a cambios en el sistema de transporte. También se encontró factible encontrar patrones de variación periódicos o aquellos asociados cambios drásticos de la movilidad y su relación con periodos de ausencia de transacciones en el transporte público. De la misma manera, se observa potencial para trabajos multidisciplinarios, por ejemplo: relacionar el porcentaje de viajes regulares y variables de cada usuario con características sociodemográficas o relacionar estas últimas con la estabilidad de las zonas importantes (hogar y trabajo).

7.3. Limitaciones

A continuación se presentan las limitaciones del trabajo realizado:

- El análisis realizado sobre los datos del transporte de Gatineau y Santiago asume que existe una relación uno a uno entre usuario y tarjeta inteligente. En el caso de Gatineau, debido a que cada tarjeta posee una fotografía de la persona resulta sencillo asumir que el uso de la tarjeta es individual. Sin embargo, la falta de información personal en el caso de la tarjeta de Santiago, junto con el gran porcentaje observado de usuarios que cambian de tarjeta, genera incertidumbre en el experimento de reconocimiento de usuarios. En particular, podrían haber usuarios que no están siendo identificados simplemente porque son personas diferentes con registros de movilidad diferentes. Para eliminar esta incertidumbre se propone realizar un proceso de validación con datos provenientes de encuestas a usuarios de transporte público.
- Es importante tener en cuenta los límites propios de utilizar datos de transacciones de transporte público para evaluar el comportamiento de los usuarios. Por ejemplo, no es posible obtener la trayectoria completa de los usuarios en la ciudad, en particular de los usuarios multimodales. De la misma forma, hay que tener considerar que fenómenos como la evasión pueden alterar los resultados percibidos.
- El proceso de reconocimiento de usuarios se llevó a cabo sobre una muestra de 5.000 usuarios. Si bien la muestra es mayor a las generalmente observadas en la literatura, no es un escenario real. Por lo anterior, se propone analizar cómo afecta el tamaño de la muestra de usuarios al rendimiento de los algoritmos en el problema de reconocimiento de usuarios. También se sugiere construir un escenario en los que los conjuntos de usuarios a emparejar no estén compuestos necesariamente por los mismos individuos.

7.4. Recomendaciones

En esta sección se plantean recomendaciones para mejorar el trabajo realizado.

- Con el objetivo de evitar la redundancia de las variables descriptivas, se sugiere utilizar *Principal Component Analysis* sobre los registros del vector de características del algoritmo RoIs-FV.
- Se sugiere evaluar el rendimiento de los algoritmos utilizando un valor mayor de similitud del movimiento, según la presencia de patrones de movilidad recurrentes a lo largo del tiempo. De esta forma, se podría mitigar los casos en que un eventual alto porcentaje de viajes nuevos opaca la presencia de comportamientos altamente regulares.

En relación a la construcción de perfiles para reconocer usuarios mediante a su movilidad:

- No todos los paraderos son equivalentes en cuanto a popularidad de uso. Luego, se propone medir la importancia de un paradero en relación a la movilidad de una persona, junto con la popularidad de la parada en relación al uso del sistema. De esta forma, entre menos popular sea una parada frecuentemente visitada por un usuario particular,

será más descriptiva de la unicidad de la movilidad.

- Los resultados señalan que más de 2 Regiones de Interés por usuario, disminuye la relación entre usuarios reconocidos y no reconocidos. Esto sugiere que las Regiones de Interés no debiesen ser calculadas mediante un porcentaje estándar, sino que a través del porcentaje que permita extraer solo las 2 Regiones más representativas de un usuario.
- Para evitar redundancia, las Regiones de Interés fueron calculadas utilizando solo las paradas de origen. Sin embargo, se debiese evaluar la efectividad de utilizar las paradas de transferencia entre servicios y las paradas de bajada.
- Considerando que hay tarjetas cuya movilidad se empareja más de una vez debido a la popularidad de los lugares visitados; se plantea evaluar la metodología utilizada en Naini et al. Naini et al., donde el proceso de reconocimiento se lleva a cabo como un problema de optimización global, en vez de una búsqueda por similitud individual.

Bibliografía

Agard, B., Morency, C., y Trépanier, M. (2006). Mining public transport user behaviour from smart card data. In *12th IFAC symposium on information control problems in manufacturing-INCOM*, pages 17–19.

Bagchi, M. y White, P. (2004). What role for smart-card data from bus systems? *Municipal Engineer*, 157(1):39–46.

Beltrán, P., Cortes, C., Gschwender, A., Ibarra, R., Munizaga, M., Ortega, M., Palma, C., y Zuñiga, M. (2011). Obtención de información valiosa a partir de datos de transantiago. In *XV Congreso Chileno de Ingeniería de Transporte*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Chen, X., Lu, R., Ma, X., y Pang, J. (2014). Measuring user similarity with trajectory patterns: Principles and new metrics. In *Asia-Pacific Web Conference*, pages 437–448. Springer.

Chen, X., Pang, J., y Xue, R. (2013). Constructing and comparing user mobility profiles for location-based services. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 261–266. ACM.

De Montjoye, Y.-A., Quoidbach, J., Robic, F., y Pentland, A. S. (2013). Predicting personality using novel mobile phone-based metrics. In *Social computing, behavioral-cultural modeling and prediction*, pages 48–55. Springer.

De Mulder, Y., Danezis, G., Batina, L., y Preneel, B. (2008). Identification via location-profiling in gsm networks. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, pages 23–32. ACM.

Devillaine, F., Munizaga, M., y Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, (2276):48–55.

Gambs, S., Killijian, M.-O., y del Prado Cortez, M. N. (2014). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614.

Giannotti, F., Nanni, M., Pinelli, F., y Pedreschi, D. (2007). Trajectory pattern mining. In

Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 330–339. ACM.

González, M. C., Hidalgo, C. A., y Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Guyon, I. y Elisseeff, A. (2006). An introduction to feature extraction. *Feature extraction*, pages 1–25.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., y González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318.

Huff, J. O. y Hanson, S. (1986). Repetition and variability in urban travel. *Geographical Analysis*, 18(2):97–114.

Iglewicz, B. y Hoaglin, D. C. (1993). *How to detect and handle outliers*, volume 16. Asq Press.

Jones, P. y Clarke, M. (1988). The significance and measurement of variability in travel behaviour. *Transportation*, 15(1-2):65–87.

Kusakabe, T. y Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, 46:179–191.

Lee, M.-J. y Chung, C.-W. (2011). A user similarity calculation based on the location for social network services. In *Database Systems for Advanced Applications*, pages 38–52. Springer.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., y Ma, W.-Y. (2008). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM.

Liu, H. y Schneider, M. (2012). Similarity measurement of moving object trajectories. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 19–22. ACM.

Lv, M., Chen, L., y Chen, G. (2013). Mining user similarity based on routine activities. *Information Sciences*, 236:17–32.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., y Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12.

Morency, C., Trepanier, M., y Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203.

Munizaga, M. A. y Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18.

Naini, F. M., Unnikrishnan, J., Thiran, P., y Vetterli, M. (2016). Where you are is who you are: User identification by matching statistics. *IEEE Transactions on Information Forensics and Security*, 11(2):358–372.

Ortega-Tong, M. A. (2013). Classification of London's public transport users using smart card data. Master's thesis, Massachusetts Institute of Technology.

Pelletier, M.-P., Trépanier, M., y Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Pendyala, R., Parashar, A., y Muthyalagari, G. (2001). Measuring day-to day variability in travel characteristics using gps data. In *79th annual meeting of the Transportation Research Board*.

Richardson Corvalán, C. I. (2014). Construcción y caracterización de perfiles de clientes en base a su movilidad. Bachelor's thesis, Universidad de Chile.

Robinson, S., Narayanan, B., Toh, N., y Pereira, F. (2014). Methods for pre-processing smart-card data to improve data quality. *Transportation Research Part C: Emerging Technologies*, 49:43–58.

Schlich, R. y Axhausen, K. W. (2003). Habitual travel behaviour: evidence from a six-week travel diary. *Transportation*, 30(1):13–36.

Song, C., Qu, Z., Blumm, N., y Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Spurr, T., Chu, A., Chapleau, R., y Piché, D. (2015). A smart card transaction "travel diary" to assess the accuracy of the montréal household travel survey. *Transportation Research Procedia*, 11:350–364.

Thakur, G. S., Helmy, A., y Hsu, W.-J. (2010). Similarity analysis and modeling in mobile societies: the missing link. In *Proceedings of the 5th ACM workshop on Challenged networks*, pages 13–20. ACM.

Tirachini, A. (2015). Probability distribution of walking trips and effects of restricting free pedestrian movement on walking distance. *Transport policy*, 37:101–110.

Trépanier, M., Tranchant, N., y Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14.

Tukey, J. W. (1977). Exploratory data analysis.

Valenzuela, D. M. (2011). Técnicas de imputación para viajes con información incompleta a partir de datos transaccionales de transantiago. Bachelor thesis, Universidad de Chile.

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176):88–93.

Wagner, R. A. y Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Ying, J. J.-C., Lu, E. H.-C., Lee, W.-C., Weng, T.-C., y Tseng, V. S. (2010). Mining user similarity from semantic trajectories. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 19–26. ACM.

Yuan, Y. y Raubal, M. (2014). Measuring similarity of mobile phone user trajectories–a spatio-temporal edit distance method. *International Journal of Geographical Information Science*, 28(3):496–520.

Yue, Y., Lan, T., Yeh, A. G., y Li, Q.-Q. (2014). Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1(2):69–78.

Anexo A

Flujo de tarjetas entre diferentes cortes temporales

En este anexo se presentan las tablas que resumen la fluctuación de las tarjetas entre distintos cortes temporales.

Tabla A.1: Origen de las tarjetas de una semana de abril 2012 según las categorías Frecuentes y No frecuentes. Los porcentajes fueron calculados en relación al total de tarjetas observadas en la semana de abril 2012, que corresponde a 3.288.464 tarjetas.

	Semana abril 2012	
Origen de las tarjetas	Frecuentes	No frecuentes
Frecuentes 2011	13,87	6,39
No frecuentes 2011	5,87	11,22
Nuevas	28,34	34,31
Porcentaje total	48,08	51,92

Tabla A.2: Origen de las tarjetas de una semana de abril 2013 según las categorías Frecuentes y No frecuentes. Los porcentajes fueron calculados en relación al total de tarjetas observadas en la semana de abril 2013, que corresponde a 3.340.078 tarjetas.

	Semana abril 2013	
Origen de las tarjetas	Frecuentes	No frecuentes
Frecuentes 2011	1,21	1,51
No frecuentes 2011	1,41	3,27
Frecuentes 2012	14,24	6,11
No frecuentes 2012	5,72	11,16
Nuevas	25,69	29,68
Porcentaje total	48,27	51,73

Tabla A.3: Origen de las tarjetas de una semana de septiembre 2013 según las categorías Frecuentes y No frecuentes. Los porcentajes fueron calculados en relación al total de tarjetas observadas en la semana de septiembre 2013, que corresponde a 3.321.592 tarjetas.

	Semana septiembre 2013	
Origen de las tarjetas	Frecuentes	No Frecuentes
Frecuentes 2011	0,40	0,72
No frecuentes 2011	0,53	1,61
Frecuentes 2012	1,14	1,58
No frecuentes 2012	1,21	3,64
Frecuentes abril 2013	21,16	7,62
No frecuentes abril 2013	6,75	15,11
Nuevas	16,18	22,35
Porcentaje total	47,36	52,64

Tabla A.4: Origen de las tarjetas de una semana de abril 2015 según las categorías Frecuentes y No frecuentes. Los porcentajes fueron calculados en relación al total de tarjetas observadas en la semana de abril 2015, que corresponde a 3.306.914 tarjetas.

	Semana abril 2015	
Origen de las tarjetas	Frecuentes	No Frecuentes
Frecuentes 2011	0,27	0,00
No frecuentes 2011	0,39	0,92
Frecuentes 2012	0,49	0,67
No frecuentes 2012	0,66	1,58
Frecuentes abril 2013	0,83	0,97
No frecuentes abril 2013	1,18	2,93
Frecuentes septiembre 2013	10,02	5,24
No frecuentes septiembre 2013	4,42	8,60
Nuevas	28,08	32,76
Porcentaje total	46,33	53,67

Anexo B

Ejemplos de los algoritmos de caracterización y comparación de la movilidad

En este anexo se ejemplifican los tres algoritmos implementados en esta tesis utilizando los registros de dos cortes temporales de un usuario de Transantiago.

B.1. Registros de movilidad del usuario Guido

Guido se mueve parecido en los cortes temporales analizados. La Figura B.1 muestra las dos tablas de transacciones asociadas a Guido. En estas tablas es posible notar que las transacciones ocurren mayoritariamente en la estaciones de metro La Moneda y Simón Bolivar. Hay días en que el usuario utiliza estaciones alternativas al metro La Moneda, como las estaciones: Universidad de Chile y Santa Lucía. En ambas tablas se registra solo una actividad que no circunda las estaciones La Moneda y Simón Bolivar. En el primer corte temporal esta actividad se ve representada en un viaje al metro Tobalaba, y en el segundo corte temporal se observa un viaje con destino a la esquina de Av. Eliodoro Yañez con Av. Pedro de Valdivia. Además, este cambio repercute en el tipo de transporte utilizado: en el primer corte temporal, Guido utiliza exclusivamente el metro; en el segundo corte temporal utiliza metro y bus.

También es posible observar un evento anómalo en la tabla (A) de la Figura B.1: la transacción 7 no tiene bajada estimada. Hay solo 8 segundos de diferencia entre la transacción 7 y 8 y ambas son hechas en la estación de metro La Moneda. Por lo anterior, la transacción 7 se puede justificar como un error del sistema o bien, el usuario utilizó su tarjeta para pagar la tarifa a otra persona.

La Figura B.2 muestra la representación geográfica de las tablas asociadas al usuario Guido. La Figura B.2 (A) muestra las RoIs en ambos cortes temporales, representados a través de círculos rojos con radios de 500 metros. La Figura B.2 (B) muestra la trayectoria a

través de pines que marcan origen (verde), transbordo (azules) y destino (rojos), y el número corresponde a la cantidad de transacciones en estas paradas. En esta representación es más fácil ver las similitudes y diferencias espaciales del comportamiento de Guido entre ambos cortes temporales.

	tiempo_subida	id	tipo_transporte	tiempo_bajada	par_subida	par_bajada
0	2013-04-15 08:16:26	guido_id	METRO	2013-04-15 08:52:26	SIMON BOLIVAR	LA MONEDA
1	2013-04-15 18:53:57	guido_id	METRO	2013-04-15 19:13:57	LA MONEDA	SIMON BOLIVAR
2	2013-04-16 08:14:19	guido_id	METRO	2013-04-16 08:50:19	SIMON BOLIVAR	LA MONEDA
3	2013-04-16 18:27:29	guido_id	METRO	2013-04-16 18:47:29	LA MONEDA	SIMON BOLIVAR
4	2013-04-17 08:20:10	guido_id	METRO	2013-04-17 08:56:10	SIMON BOLIVAR	LA MONEDA
5	2013-04-17 18:30:42	guido_id	METRO	2013-04-17 18:50:42	LA MONEDA	SIMON BOLIVAR
6	2013-04-18 08:17:27	guido_id	METRO	2013-04-18 08:53:27	SIMON BOLIVAR	LA MONEDA
7	2013-04-18 16:32:53	guido_id	METRO	NaN	LA MONEDA	NaN
8	2013-04-18 16:33:01	guido_id	METRO	2013-04-18 16:46:01	LA MONEDA	TOBALABA
9	2013-04-18 18:37:00	guido_id	METRO	2013-04-18 18:43:00	TOBALABA	SIMON BOLIVAR
10	2013-04-19 08:21:45	guido_id	METRO	2013-04-19 08:59:45	SIMON BOLIVAR	SANTA LUCIA
11	2013-04-19 17:52:09	guido_id	METRO	2013-04-19 18:09:09	SANTA LUCIA	SIMON BOLIVAR

(A)

	tiempo_subida	id	tipo_transporte	tiempo_bajada	par_subida	par_bajada
0	2013-09-23 08:26:08	guido_id	METRO	2013-09-23 09:02:08	SIMON BOLIVAR	LA MONEDA
1	2013-09-23 19:08:45	guido_id	METRO	2013-09-23 19:28:45	LA MONEDA	SIMON BOLIVAR
2	2013-09-24 08:28:20	guido_id	METRO	2013-09-24 09:05:20	SIMON BOLIVAR	UNIVERSIDAD DE CHILE
3	2013-09-24 18:35:51	guido_id	METRO	2013-09-24 18:54:51	UNIVERSIDAD DE CHILE	SIMON BOLIVAR
4	2013-09-25 08:27:21	guido_id	METRO	2013-09-25 09:03:21	SIMON BOLIVAR	LA MONEDA
5	2013-09-25 18:32:49	guido_id	METRO	2013-09-25 18:52:49	LA MONEDA	SIMON BOLIVAR
6	2013-09-26 08:35:54	guido_id	METRO	2013-09-26 08:54:54	SIMON BOLIVAR	LA MONEDA
7	2013-09-26 15:35:58	guido_id	METRO	2013-09-26 15:45:58	LA MONEDA	PEDRO DE VALDIVIA
8	2013-09-26 15:51:20	guido_id	ZP	2013-09-26 15:57:25	T-14-127-NS-4	T-14-127-NS-15
9	2013-09-26 17:52:13	guido_id	BUS	2013-09-26 18:03:08	T-14-127-NS-15	T-14-127-NS-45
10	2013-09-26 18:13:27	guido_id	BUS	2013-09-26 18:30:30	L-18-6-PO-110	L-19-21-SN-15
11	2013-09-27 08:28:00	guido_id	METRO	2013-09-27 09:04:00	SIMON BOLIVAR	LA MONEDA
12	2013-09-27 17:13:49	guido_id	METRO	2013-09-27 17:33:49	LA MONEDA	SIMON BOLIVAR

(B)

Figura B.1: Las tablas (A) y (B) son las tablas de transacciones asociadas al usuario Guido en los cortes temporales de abril y septiembre respectivamente.

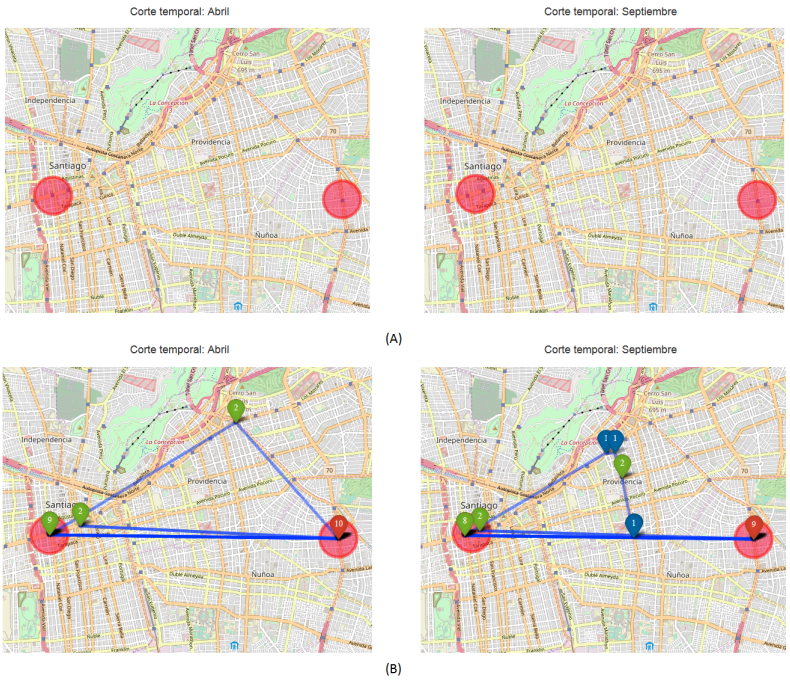


Figura B.2: El gráfico (a) muestra las RoIs del usuario Guido en los dos cortes temporales analizados. El gráfico (b) muestra la trayectoria del usuario Guido en los dos cortes temporales analizados.

B.2. Comparación de registros de Guido con el algoritmo TPM

Sean las tablas (A) y (B) de la Figura B.1 la entrada del algoritmo TPM. En primer lugar, se procede a extraer la secuencia de posiciones de cada corte temporal. La Figura B.3 muestra las secuencias de posiciones de la semana de abril y septiembre del usuario Guido.

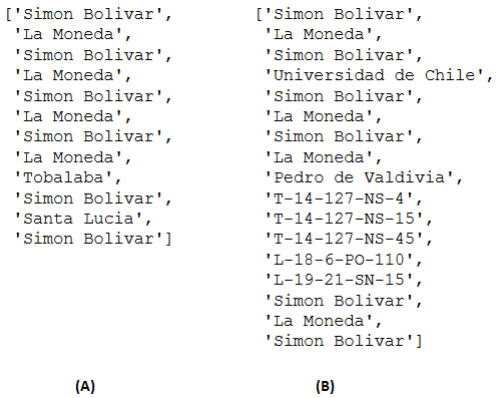


Figura B.3: Secuencia de posiciones usuario Guido, corte temporal abril y septiembre 2013.

Luego, es necesario extraer el conjunto mínimo de paradas de la secuencia de posiciones de abril, en este caso: [*'Simón Bolivar'*, *'La Moneda'*, *'Tobalaba'*, *'Santa Lucía'*]. Con este conjunto se construye la matriz de la Figura B.4, donde se almacena el conteo de viajes entre los pares Origen-Destino de la secuencia de posiciones. Por último, se divide cada celda por la suma total de viajes de la fila correspondiente, de esta forma se obtiene la matriz TPM, la cual se ilustra en la Figura B.5.

	SIMON BOLIVAR	LA MONEDA	TOBALABA	SANTA LUCIA
SIMON BOLIVAR	0	4	0	1
LA MONEDA	3	0	1	0
TOBALABA	1	0	0	0
SANTA LUCIA	1	0	0	0

Figura B.4: Matriz con los viajes Origen-Destino asociada al corte temporal de abril del usuario Guido.

	SIMON BOLIVAR	LA MONEDA	TOBALABA	SANTA LUCIA
SIMON BOLIVAR	0.00	0.8	0.00	0.2
LA MONEDA	0.75	0.0	0.25	0.0
TOBALABA	1.00	0.0	0.00	0.0
SANTA LUCIA	1.00	0.0	0.00	0.0

Figura B.5: Matriz TPM asociada al corte temporal de abril del usuario Guido.

Para comparar la movilidad entre el corte temporal de abril y septiembre del usuario Guido se procede a sumar el logaritmo de las probabilidades almacenadas en la matriz TPM de los viajes asociados a la secuencia de posiciones de septiembre. En el caso de que un viaje no tenga una probabilidad definida en la matriz TPM, se procede a utilizar el valor $10^{\frac{-800}{l-1}}$, con l el largo de la secuencia de posiciones de septiembre.

Utilizando la notación definida en la Sección 3.1, el cálculo del indicador de similitud de la movilidad de abril y septiembre del usuario Guido queda expuesto en el siguiente desarrollo:

$$\begin{aligned}
\text{sim}\, TPM_{Guido_abril,Guido_sept} &= \sum_{j=1}^{l-1} \log_{10} Pr(S^{j}_{Guido_sept} \rightarrow S^{j+1}_{Guido_sept}) \\
&= 4 \times \log_{10} Pr(\textit{Simon Bolivar} \rightarrow \textit{La Moneda}) + \\
&\quad 3 \times \log_{10} Pr(\textit{La Moneda} \rightarrow \textit{Simon Bolivar}) + \\
&\quad 9 \times \log_{10} Pr(\textit{Transiciones no definidas en la TPM}) \\
&= 4 \times \log_{10}(0{,}8) + 3 \times \log_{10}(0{,}75) + 9 \times \log_{10}(10^{\frac{-800}{16}}) \\
&= -450{,}76
\end{aligned}$$

Finalmente, para normalizar el valor $\text{sim}\, TPM_{Guido\_abril,Guido\_sept}$ a la escala [0-1], se utiliza una normalización min-max, donde el mínimo es -800 y el máximo es 0. De lo anterior se obtiene que el indicador de similitud TPM entre los cortes temporales de abril y septiembre del usuario Guido es 0,44.

B.3. Comparación de registros de Guido con el algoritmo EDM

Sean las tablas (A) y (B) de la Figura B.1 la entrada del algoritmo EDM. En primer lugar, se procede a extraer la trayectoria asociada a cada corte temporal. Para construir la trayectoria es necesario asociar a las tablas (A) y (B) la posición de cada parada visitada. Las tablas de la Figura B.6 muestran la información de las trayectorias de abril y septiembre del usuario Guido. La Figura B.2 muestra la representación geográfica de las trayectorias.

Transacción	Latitud	Longitud
1	-33.44591	-70.57204
2	-33.44509	-70.65463
3	-33.44591	-70.57204
4	-33.44509	-70.65463
5	-33.44591	-70.57204
6	-33.44509	-70.65463
7	-33.44591	-70.57204
8	-33.44509	-70.65463
9	-33.41823	-70.60145
10	-33.44591	-70.57204
11	-33.44285	-70.64581
12	-33.44591	-70.57204

(A)

Transacción	Latitud	Longitud
1	-33.44591	-70.57204
2	-33.44509	-70.65463
3	-33.44591	-70.57204
4	-33.44394	-70.65039
5	-33.44591	-70.57204
6	-33.44509	-70.65463
7	-33.44591	-70.57204
8	-33.44509	-70.65463
9	-33.42547	-70.61427
10	-33.42536	-70.61177
11	-33.43125	-70.60958
12	-33.44532	-70.60623
13	-33.44562	-70.60631
14	-33.44574	-70.57185
15	-33.44591	-70.57204
16	-33.44509	-70.65463
17	-33.44591	-70.57204

(B)

Figura B.6: Secuencia de posiciones usuario Guido, corte temporal abril y septiembre 2013.

Luego, se procede a construir una matriz M de tamaño $n+1 \times m+1$, donde n corresponde al largo de la trayectoria de abril, y m corresponde al largo de la trayectoria de septiembre. En la matriz M se almacenan los costos calculados mediante el algoritmo de distancia de edición, resultando la matriz de la Figura B.7. Los costos de la primera fila de la matriz M corresponden al desplazamiento del centroide de la trayectoria de abril al ir eliminando cada posición de esta trayectoria. Los costos de la primera columna de la matriz M corresponden al desplazamiento del centroide de la trayectoria de abril al ir agregando cada posición de la trayectoria de abril. El resto de los costos se calcula a través de la selección entre eliminar, insertar o reemplazar cada combinación de posiciones.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0.0	259.0	591.0	850.0	1152.0	1411.0	1743.0	2002.0	2334.0	2490.0	2643.0	2744.0	2767.0	2792.0	3052.0	3311.0	3643.0	3902.0
1	306.0	0.0	332.0	591.0	893.0	1152.0	1484.0	1743.0	2075.0	2231.0	2384.0	2485.0	2508.0	2533.0	2793.0	3052.0	3384.0	3643.0
2	699.0	393.0	0.0	259.0	561.0	820.0	1152.0	1411.0	1743.0	1899.0	2052.0	2153.0	2176.0	2201.0	2461.0	2720.0	3052.0	3311.0
3	1005.0	699.0	306.0	0.0	302.0	561.0	893.0	1152.0	1484.0	1640.0	1793.0	1894.0	1917.0	1942.0	2202.0	2461.0	2793.0	3052.0
4	1398.0	1092.0	699.0	393.0	34.0	293.0	561.0	820.0	1152.0	1308.0	1461.0	1562.0	1585.0	1610.0	1870.0	2129.0	2461.0	2720.0
5	1704.0	1398.0	1005.0	699.0	340.0	34.0	366.0	561.0	893.0	1049.0	1202.0	1303.0	1326.0	1351.0	1611.0	1870.0	2202.0	2461.0
6	2097.0	1791.0	1398.0	1092.0	733.0	427.0	34.0	293.0	561.0	717.0	870.0	971.0	994.0	1019.0	1279.0	1538.0	1870.0	2129.0
7	2403.0	2097.0	1704.0	1398.0	1039.0	733.0	340.0	34.0	366.0	522.0	675.0	776.0	799.0	824.0	1021.0	1279.0	1611.0	1870.0
8	2796.0	2490.0	2097.0	1791.0	1432.0	1126.0	733.0	427.0	34.0	190.0	343.0	444.0	467.0	492.0	752.0	1011.0	1279.0	1538.0
9	3052.0	2746.0	2353.0	2047.0	1688.0	1382.0	989.0	683.0	290.0	153.0	293.0	394.0	417.0	442.0	702.0	961.0	1293.0	1552.0
10	3358.0	3052.0	2659.0	2353.0	1994.0	1688.0	1295.0	989.0	596.0	459.0	514.0	613.0	636.0	661.0	444.0	702.0	1034.0	1293.0
11	3676.0	3370.0	2977.0	2671.0	2312.0	2006.0	1613.0	1307.0	914.0	777.0	768.0	814.0	837.0	862.0	762.0	1016.0	773.0	1032.0
12	3982.0	3676.0	3283.0	2977.0	2618.0	2312.0	1919.0	1613.0	1220.0	1083.0	1074.0	1088.0	1079.0	1102.0	864.0	762.0	1079.0	773.0

Figura B.7: Secuencia de posiciones usuario Guido, corte temporal abril y septiembre 2013.

De la Figura B.7 se desprende que la distancia de edición entre las trayectorias de abril y septiembre de Guido es 773. Para normalizar esta distancia se requiere de un máximo, el cual dependerá de las distancias del grupo de usuarios que se esté analizando. Para tener una idea de la magnitud de la distancia entre las trayectorias de Guido, en el caso de los usuarios de Transantiago la máxima distancia observada entre dos usuarios fue 24.942. Utilizando

Tabla B.1: Vector de características del usuario Guido en los cortes temporales abril y septiembre 2013.

Tipo de Característica	Característica	Guido, abril 2013	Guido, septiembre 2013
Temporal	Hora de inicio promedio primer viaje (semana/fin de semana)	29.881 / N/A	30.548 / N/A
	Hora de inicio promedio último viaje (semana/fin de semana)	66.495 / N/A	65.801 / N/A
	Número de días con viajes	5	5
	Moda del número de viajes por día	2	2
	Frecuencia de la moda del número de viajes por día	4	4
	Promedio de número de viajes por día (semana/fin de semana)	2.4 / 0	2.4 / 0
Espacial	Distancia viajada	77.3	78.4
	Mínima distancia viajada promedio	7.7	7.7
	Máxima distancia viajada diaria promedio	4.1	4.8
	Radio de giro	3.7	4.8
	Entropía temporalmente no correlacionada	1.6	2.2
	Entropía aleatoria	2.0	2.6
	Porcentaje de primeras paradas diferentes (semana/fin de semana)	0.0 / N/A	0.0 / N/A
	Porcentaje de últimas paradas diferentes (semana/fin de semana)	50.0 / N/A	50.0 / N/A
Demográfica	Tipo de tarjeta	Adulto	Adulto
Actividad	Promedio de tiempo de actividad más corta por día (semana/fin de semana)	486 / N/A	499 / N/A
	Promedio de tiempo de actividad más larga por día (semana/fin de semana)	783 / N/A	793 / N/A
Modo de transporte	Número de etapas por viaje más frecuente	1	1
	Porcentaje de días con viajes exclusivos en bus	0.0	0.0
	Porcentaje de días con viajes exclusivos en metro	100.0	80.0
	Porcentaje de viajes en bus	0.0	23.0

este máximo y haciendo una normalización min-max, la distancia de Guido entre abril y septiembre sería 0.03.

B.4. Comparación de registros de Guido con el algoritmo RoIs-FV

Sean las tablas (A) y (B) de la Figura B.1 la entrada del algoritmo RoIs-FV. En primer lugar, se procede a extraer las RoIs asociadas a cada tabla. Para extraer las RoIs es necesario asociar a las tablas (A) y (B) la posición de cada parada visitada. Luego, se realiza *clustering* jerárquico sobre las posiciones de las trayectorias presentadas en la Figura B.6. De este proceso se seleccionan los *clusters* que reúnan al menos el 70 % de las transacciones. En el caso del usuario Guido, cada corte temporal tiene asociadas dos RoIs, las cuales se grafican en la Figura B.2 (A). En esta figura se observa que las RoIs en ambos periodos ilustran las mismas áreas de interés.

Como el usuario Guido comparte al menos 2 RoIs en ambos periodos, se procede a extraer los vectores de características. La Tabla B.1 almacena los valores de cada descriptor del vector de características en el corte temporal de abril y septiembre 2013. El dominio de cada valor esta descrito en la Tabla 3.3.

Para que las características posean el mismo peso en la función de distancia es preciso normalizar las características. Sin embargo muchas de las características utilizadas no tienen un máximo definido, por lo que es necesario calcularlo a partir de la muestra observada. La Figura B.2 muestra los mínimos y máximos por cada característica utilizados en la muestra de Transantiago. Utilizando esta información se procede a normalizar los valores de las características del usuario Guido, resultando en dos vectores con todas sus dimensiones en el rango [0,1].

Tabla B.2: Mínimos y máximos de las variables descriptivas en usuarios de Transantiago.

Tipo de Característica	**Característica**	**Mínimo**	**Máximo**
Temporal	Hora de inicio promedio primer viaje (semana/fin de semana)	10.215 / 33.000	55.001 / 83.882
	Hora de inicio promedio último viaje (semana/fin de semana)	44.036 / 21.040	85.866 / 86.386
	Número de días con viajes	1	7
	Moda del número de viajes por día	1	9
	Frecuencia de la moda del número de viajes por día	1	7
	Promedio de número de viajes por día (semana/fin de semana)	1,2 / 0,0	3,5 / 5,0
Espacial	Distancia viajada	7,44	285,75
	Mínima distancia viajada promedio	0,03	10,0
	Máxima distancia viajada diaria promedio	0.56	31,3
	Radio de giro	0,0	11.552
	Entropía temporalmente no correlacionada	0,91	4,58
	Entropía aleatoria	1,58	4,76
	Porcentaje de primeras paradas diferentes (semana/fin de semana)	0,0 / 0,0	100,0 / 100,0
	Porcentaje de últimas paradas diferentes (semana/fin de semana)	0,0 / 0,0	100,0 / 100,0
Demográfica	Tipo de tarjeta	Variable nominal	
Actividad	Promedio de tiempo de actividad más corta por día (semana/fin de semana)	0 / 0	1.033 / 1.438
	Promedio de tiempo de actividad más larga por día (semana/fin de semana)	511 / 36	1.064 / 1.599
Modo de transporte	Número de etapas por viaje más frecuente	1	3
	Porcentaje de días con viajes exclusivos en bus	0,0	100,0
	Porcentaje de días con viajes exclusivos en metro	0,0	100,0
	Porcentaje de viajes en bus	0,0	100,0

Finalmente, el indicador de distancia del algoritmo RoIs-FV dependerá de la función de distancia que se utilice sobre los dos vectores de características normalizados. Por ejemplo, si se utiliza la función de distancia Manhattan, el indicador de distancia RoIs-FV del usuario Guido entre la semana de abril y septiembre es 0,87. Es necesario advertir que el indicador RoIs-FV también debe ser normalizado para ser comparado con otros algoritmos, y al igual que la distancia EDM, el máximo depende de los indicadores de todos los usuarios. En el caso de Transantiago se definió el máximo como 3,80, por tanto, al normalizar min-max el indicador de Guido resulta en 0,24.