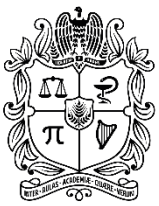


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Generación de matrices OD por motivo de viaje a través de minería de datos de información de sistemas automáticos de recaudo de tarifa en transporte público: Caso TransMilenio, Colombia

Sergio Alejandro Peña Pedreros

Universidad Nacional de Colombia
Facultad, Ingeniería Civil y Agrícola
Bogotá, Colombia
2020

Generación de matrices OD por motivo de viaje a través de minería de datos de información de sistemas automáticos de recaudo de tarifa en transporte público: Caso TransMilenio, Colombia

Sergio Alejandro Peña Pedreros

Trabajo de investigación presentado como requisito parcial para optar al título de:

Magister en Ingeniería - Transporte

Director (a):

Sonia Cecilia Mangones Matos, Ph.D.

Línea de Investigación:

Movilidad y desarrollo tecnológico

Grupo de Investigación:

Logística para el transporte sostenible y la seguridad - TRANSLOGYT

Universidad Nacional de Colombia

Facultad, Ingeniería Civil y Agrícola

Bogotá, D.C., Colombia

2020

A mi abuelo, que ha sido mi ejemplo de vida, me ha ensañado que con disciplina se pueden alcanzar grandes cosas.

A mi princesa, por todo su amor y motivación, porque juntos seguiremos cosechando grandes cosas.

Agradecimientos

A la empresa TransMilenio S.A., la Secretaría Distrital de Movilidad – SDM y a la Infraestructura de Datos Espaciales para el Distrito Capital – IDECA, por el acceso a los datos, los cual se ha convertido en insumo fundamental del presente trabajo.

A Cesar Moreno, por su valiosa explicación en SQL Server, a Ana Jiménez, Jose Borrego, Oscar Peña, Angela Cristancho, Ximena Velandia y Lina Palacio, por el tiempo dedicado para la revisión y generación de comentarios, que han aportado a mejorar este documento.

A la Universidad Nacional de Colombia y a la directora del trabajo de grado Sonia Mangones, por su tiempo, orientación y comentarios para lograr los objetivos del presente trabajo.

Resumen

El uso de los Sistemas Inteligentes de Transporte ha generado una nueva posibilidad para la obtención de grandes volúmenes de datos que permiten describir su funcionamiento. En el caso de los sistemas de transporte público, los Sistemas Automáticos de Recaudo de la tarifa, permiten obtener información del uso de las tarjetas inteligentes y datos del lugar de entrada, valor pagado y características del usuario. Estos datos soportan la obtención de información relevante para la planeación y operación de los sistemas de transporte público.

No obstante, hay información derivada al viaje que no se puede obtener de forma directa de los datos del uso de tarjetas inteligentes, como el motivo del viaje. A partir de la información de validaciones del sistema troncal del SITP (TransMilenio), se realiza el procesamiento y minería de datos para construir matrices de viajes a partir del encadenamiento del viaje y la aplicación de heurísticas. Posteriormente, se aplican modelos tipo logit multinomial, que permitan estimar el motivo del viaje, los cuales son comparados con los patrones de viajes obtenidos a partir de la encuesta origen destino en hogares de Bogotá, D.C.

Palabras clave: motivo de viaje, matrices de viaje, TransMilenio, método de encadenamiento de viaje, entradas, usos, validaciones, sistemas automáticos de recaudo.

Abstract

The use of Intelligent Transport Systems has created a new possibility for obtaining large volumes of data that can modify its operation. In the case of public transport systems, the information from the Automatic Fare Collection Systems plays an important role, since it can obtain information on the use of smart cards, data on the place of entry, value paid and the user's own characteristics.

However, there is information derived from the trip that cannot be obtained directly from the data from the use of smart cards, such as the reason for the trip, is that from the validation information of the SITP trunk system (TransMilenio), data processing and mining is carried out, travel matrices are built from chaining and the application of heuristics for their inference, in order to develop multinomial logit-type models, which obtain the reason for the trip, which are compared with the origin of destination household survey in Bogotá, D.C.

Keywords: OD matrices, TransMilenio, travel chain method, income (validations), automatic fare collection system (AFC), Smart cards.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

1. Introducción

La migración de personas, cambios socioeconómicos, condiciones de orden público e incluso cambios climáticos, generan modificaciones en el comportamiento de la demanda de viajes que ingresan al componente troncal del SITP, desde ahora, TransMilenio; razón por la cual, se requiere que la oferta de servicios de TransMilenio, se ajusten a estos cambios, con el fin de poder atender la demanda de viajes. Para esto, se utilizan las Matrices Origen-Destino (MOD) de viajes a nivel de estaciones, como elemento de la planeación, programación y operación de TransMilenio, considerando que reflejan los deseos de viaje de las personas entre diferentes puntos de origen y destino (Ortúzar and Willumsen, 2011).

En la actualidad, los métodos que se utilizan para obtener las MOD de TransMilenio consideran, por una parte, la recolección de datos en hogares o por interceptación de viajes; lo que implica importantes inversiones para su recolección, procesamiento y obtención (Ortúzar, 2015); y por otra, utilizan los datos de las tarjetas inteligentes para generar matrices semestrales de los viajes a nivel de Zona de Análisis de Transporte (ZAT) del modelo de transporte de propiedad de TransMilenio S.A., lo que refleja la importancia de los datos del recaudo (TransMilenio S.A., 2019). Sin embargo, las MOD de TransMilenio, no cuentan con información asociada a características como el motivo del viaje.

Por esto, el presente trabajo busca responder de qué forma se pueden obtener MOD a partir de datos del sistema automático de recaudo de tarifa, utilizando el método de encadenamiento de viajes a nivel de estación y/o portal de TransMilenio y haciendo una inferencia del motivo del viaje. En concordancia con lo anterior, el presente trabajo, tiene como objetivo, estimar MOD por motivo de viaje del componente troncal del SITP (TransMilenio) a partir de la minería de datos de las tarjetas inteligentes del Sistema de Recaudo Automático – AFC y la estimación del motivo del viaje a partir de modelos logit multinomiales.

Para lograr el objetivo general, se desarrollan cuatro objetivos específicos, que se abordan en los diferentes capítulos del documento; el primer objetivo específico que busca, analizar, explorar y preparar los datos temporales y espaciales de una semana de validaciones de TransMilenio y de las encuestas de viajes en hogares, se desarrolla en el capítulo 2, Procesamiento y minería de datos. El segundo objetivo específico que trata sobre, inferir matrices de viajes del componente troncal del SITP (TransMilenio), a partir de los datos de los sistemas AFC, aplicando el método de encadenamiento del viaje, se presenta con el capítulo 3, Estimación de matrices OD a partir de datos de validación de entradas al sistema.

El tercer objetivo específico el cual busca determinar las características temporales y espaciales que permiten inferir el motivo de viaje en el sistema troncal del SITP (TransMilenio), se desarrolla en el capítulo 4, Modelo de inferencia de motivo de viaje; el cuarto objetivo específico que se desarrolla para comparar las matrices origen destino – MOD inferidas, con MOD de otros estudios como la Encuesta Origen Destino en Hogares, se alcanza en el capítulo 5, Inferencia y comparación de matrices por motivo de viaje; finalmente, se incluye el capítulo 6, Conclusiones y recomendaciones, donde se resumen los principales resultados de la investigación.

Se espera que la metodología aquí utilizada, pueda ser de gran importancia, no solo para el sistema del componente troncal del SITP (TransMilenio), sino que pueda ser utilizada en los demás sistemas de transporte público que se encuentran en el marco del (CONPES - 3167, 2002), ya que dentro de la modernización de los Sistemas Estratégicos de Transporte Público (SETP) y Sistemas Integrados de Transporte Masivo (SITM), se incluye la utilización de sistemas centralizados de recaudo como los AFC, que son la fuente de datos para la obtención de los resultados buscados.

1.1 Justificación

Los métodos tradicionales y principales para la obtención de matrices origen destino (MOD), son las encuestas de viajes en hogares, las encuestas de interceptación, y las encuestas abordo de los vehículos, los cuales permiten recolectar información de los

viajeros y sus características del viaje, tales como, nivel de ingresos del hogar, tasa de motorización, propósito o motivo de viaje, modo de transporte, origen y destino del viaje, duración, entre otros.

No obstante, los costos y tiempos requeridos para la toma de información aún son elevados, a pesar de que se aplican muestreos estadísticos para definir valores de precisión y exactitud para estimar muestras que reduzcan el número de datos a recolectar (Willumsen, 2014); razón por la cual, los estudios no se realizan frecuentemente, sino de manera periódica, según las necesidades y disponibilidad de recursos.

La falta de información actualizada, limita la capacidad de respuesta de los sistemas de transporte, considerando que las grandes ciudades del mundo, sufren cambios constantes en el tamaño y preferencias de la población, como parte de la tendencia de aglomeración y migración a aquellas ciudades con mayor desarrollo económico, en busca de mejores condiciones de vida (Banco Mundial, 2009), lo que genera variaciones en la demanda de viajes y en general presiones sobre los sistemas de transporte urbano.

La limitación de datos e información para la planeación y operación de sistemas de transporte se puede contrarrestar, debido a que el incremento en el uso de tecnologías ha generado crecimientos exponenciales de datos e información, dando mayor relevancia a campos de interés que hacen uso de estos grandes volúmenes de datos (Big Data), para obtener información con valor (SAPORITI, 2016), convirtiéndola en un elemento que impulsa las industrias de finanzas, energía, transporte, educación, salud y comercio (Banco Interamericano de Desarrollo, 2018).

En el caso del transporte, los Sistemas Inteligentes de Transporte (ITS, por sus siglas en inglés) no solo ayudan en el recaudo de tarifas, control y vigilancia del transporte, sino que son fuente constante de datos. En el caso de los sistemas de transporte público colectivo la generación de datos pasivos automatizados, se pueden agrupar en dos grandes grupos según su fuente: una asociada a los Sistemas de Posicionamiento Global (GPS, por sus siglas en inglés) y otra al uso de tarjetas inteligentes (Alsger, 2017).

El uso de los datos provenientes de tarjetas inteligentes en los sistemas automático de recaudo de tarifas (AFC, por sus siglas en inglés) de los sistemas de transporte público

colectivo, son de gran interés debido a la cantidad de datos que se producen y al potencial para revelar el comportamiento dinámico de la demanda de viajes, considerando que se pueden obtener datos asociados a cada una de las tarjetas inteligentes, tales como: lugar, día, hora y tarifa, en el momento en que son usadas en la entrada y/o salida en un sistema de transporte público colectivo, lo que ha llevado a varios investigadores a realizar esfuerzos para procesar y analizar información de estos sistemas y obtener matrices origen destino de viajes (Li *et al.*, 2018).

En este sentido, los procesos de la minería de datos, son de utilidad para descubrir patrones y tendencias existentes, por medio de la aplicación de seis pasos básicos: la definición del problema, la preparación de los datos, la exploración de datos, la generación de modelos, la validación de modelos, y la implementación y actualización de los modelos, que por la cantidad de datos no pueden ser detectados por la exploración tradicional (Microsoft, 2019).

1.2 Estado del arte

1.2.1 Sistemas AFC con validación en el ingreso

En la actualidad la mayoría de los sistemas AFC, recolectan información asociada únicamente al ingreso en sistemas de transporte público colectivo, ya sea en estaciones o buses, razón de esto, varios investigadores han visto la necesidad de desarrollar métodos que permitan completar la información del viaje (Li *et al.*, 2018).

Autores como, (Alsger *et al.*, 2016), (Nunes, Galvao Dias and Falcao e Cunha, 2016), (Li *et al.*, 2011), (Wang, Attanucci and Wilson, 2011), (Nassir *et al.*, 2011), (Barry, Freimer and Slavin, 2009), (Zhao, Rahbee and Wilson, 2007) y (Chen and Fan, 2018), han utilizado el método de encadenamiento de viaje para obtener matrices origen destino. Este es un método desagregado debido a que permite inferir el destino del viaje de cada usuario en cada viaje, haciendo uso de algoritmos relativamente sencillos y utilizando información únicamente de los datos de las tarjetas.

El método de encadenamiento de viaje, tiene como objetivo, hacer una inferencia del destino del viaje mediante el encadenamiento del mismo, por medio de la revisión de los diferentes abordajes de un mismo usuario, en un mismo día, considerando tres supuestos planteados inicialmente por (Li *et al.*, 2011) que aún se mantienen, con algunas modificaciones. El primer supuesto considera que no existe un modo de transporte adicional entre dos viajes consecutivos, lo que significa que el destino del primer viaje es igual al origen del siguiente viaje, este supuesto se basa en que un porcentaje alto de pasajeros regresan a la estación anterior para comenzar su próximo viaje (Alsger, 2017). El segundo supuesto considera que cuando existen transferencias, estas son cortas, sobre este supuesto se han usados valores fijos y se han realizado análisis de sensibilidad de caminata, los cuales han utilizado variaciones de la distancia probable de caminata entre 400 y 2.000 metros. El tercer supuesto, plantea que un porcentaje alto de viajes regresan a la primera estación donde comenzó el primer viaje del día, cerrando la cadenada de viajes por día, sobre este supuesto las investigaciones varían en la definición de la hora de inicio y fin del día (Li *et al.*, 2018).

Otros investigadores como (Dou, H.; Liu, H.; Yang, 2007) y (Zhang, M.; Guo, Y.; Ma, 2014) han aplicado otra clase de modelos para la obtención de MOD con datos AFC de validación de ingreso, los cuales son de tipo probabilísticos. En estos métodos el tratamiento de datos se hace de forma agregada, considerando para cada estación o parada de origen el número de pasajeros que ingresa al sistema de transporte público colectivo, que son distribuidos según la estimación de probabilidades de descenso en las diferentes estaciones, dependiendo de la distancia y/o tiempo a la estación de origen.

1.2.2 Sistemas AFC con ingresos y salidas

A pesar de no ser la mayoría de los casos, existen sistemas AFC que registran información del ingreso y salida del viaje permitiendo obtener información total de la cadena de viaje; sobre esta información investigadores como (Jung and Sohn, 2017) y (Yu and Yang, 2006) han utilizado modelos de aprendizaje profundo para estimar la matriz de viajes. Dentro de las dificultades principales que se presentan al usar estos modelos, se encuentra el manejo de la cantidad de datos para poder entrenar las redes neuronales, la necesidad de información de ingreso y salida, y la complejidad teórica de los algoritmos empleados; dentro de las ventajas, se destaca su exhaustividad, su capacidad para estimar los viajes de cada usuarios y su opción para validar los resultados obtenidos (Li *et al.*, 2018).

Otros investigadores como (Alsger, 2017), han usado los sistemas AFC con ingreso y salida, para validar los supuesto del método de encadenamiento del viajes que registran solo el ingreso; a su vez, estimar la precisión del método y proponer mejoras en su aplicación.

1.2.3 Inferencia del motivo del viaje

Es importante considerar que la información que se puede obtener de las tarjetas inteligentes es limitada, considerando que no se puede obtener información de forma directa del motivo de viaje, transferencias en otros medios, duración del viaje, y otros atributos asociados al viajero. Adicionalmente, existen problemas de información faltante que pueden reducir el grado de precisión en la estimación de las MOD, como congestión en estaciones, medios alternativos de pago diferentes al torniquete y la evasión.

Para superar algunos de los vacíos en los datos, se definen algunos supuestos y métodos de revisión, que involucran no solamente el día para el cual se obtienen las MOD, sino que se consideran varios días de datos y partiendo de la hora del primer y segundo viaje, se infiere el motivo del viajes, ya sea de trabajo, estudio y otros, según la ventana de tiempo entre validaciones (Munizaga and Palma, 2012) o duración de la actividad.

Por otra parte, (Alsger, 2017) utiliza un método que involucra entradas, procesamiento, modelación y salidas el cual parte de los datos de las encuestas de viajes en hogares para obtener el motivo basados en la modelación de árboles de decisión. Estos modelos de inferencia del motivo de viaje utilizan métodos de validación cruzada, dividiendo los datos en dos grupos, uno para el entrenamiento del modelo de predicción y otro para la validación del modelo. Sobre el universo de datos se hace una preparación y exploración de información temporal y espacial, se estiman los modelos sobre los datos de entrenamiento y se validan los modelos con el grupo de datos restantes, así se obtienen las probabilidades que son aplicadas sobre las MOD obtenidas de los datos del AFC, para obtener submatrices por motivo de viaje.

1.2.4 Caso TransMilenio

El sistema TransMilenio comenzó operación el 18 de diciembre de 2000, con las primeras líneas de la fase uno, la Avenida Caracas y la Calle 80, como un sistema tronco-alimentado tipo autobús de tránsito rápido (BRT, por sus siglas en ingles), prestando el servicio de transporte público con buses articulados de plataforma alta que transitan sobre carriles exclusivos, con entrada de usuarios mediante el uso de tarjetas inteligentes usadas sobre torniquetes ubicados en estaciones y portales (Alcaldia Mayor de Bogotá, 2013).

En la actualidad, TransMilenio se encuentra en su fase tres y hace parte del Sistema Integrado de Transporte Público (SITP) como el componente troncal, funcionando todos los días de la semana sobre 12 líneas, 9 portales y 154 estaciones, que puede ofertar para un día laboral (lunes a viernes) 86 servicios, de los cuales 78 son de tipo expreso[1].

El sistema AFC recolectó para 2018 datos de las tarjetas inteligentes para más de 2,3 millones de entradas por día laboral, en las estaciones y portales de las tres fases del componente troncal, sin incluir el componente Dual (TransMilenio S.A., 2018); debido al potencial de los datos que son recolectados diariamente, TransMilenio S.A. consideró dentro del contrato de concesión 001 de 2011, en su cláusula 17, la construcción de matrices OD de viajes semestrales, para un día laboral normal, sábado y domingo desagregando por horario y zona de la ciudad para los servicios de TransMilenio Fase I, II, III y zonales. A su vez, la obtención de matrices de viajes de los usuarios de los servicios de Fase I, II, III y zonales para un día laboral normal, sábado y domingo desagregadas a nivel horario y zona de la ciudad; también, una matriz calibrada de viajes en punta mañana para día laboral normal, sábado y domingo de los servicios de Fase I, II, III y Zonales (TransMilenio S.A., 2019).

La metodología para la construcción de las MOD de TransMilenio utiliza las estimaciones de las entradas registradas por los usuarios a partir de la información de pago, la

[1] Servicios expresos, rutas de transporte público que no paran en todas las estaciones o portales por donde pasan.

información de GPS de los buses e información complementaria del sistema; los supuestos y precisión del algoritmo empleado para la obtención de las matrices, están basados en las definiciones hechas por (Munizaga and Palma, 2012), que resultan en matrices totales de viaje por sub sistema, sin información asociada al motivo del viaje o características socioeconómicas.

1.2.5 Hallazgos

Las MOD que son entregadas semestralmente para el SITP, por el operador del sistema AFC, no contienen información del motivo de viaje, limitando el uso de las matrices al control, en procesos de planeación y operación del sistema, y deja a un lado la importancia de que la demanda de transporte es una demanda derivada (Rus, Campos and Nombela, 2002), (Ortúzar, 2015). Esto impide que estas matrices puedan ser usadas para el seguimiento y propuesta de planes, programas y/o políticas, que se enfoquen en grupos específicos de demanda.

Por otra parte, a pesar de que los métodos han sido usados en diferentes sistemas de transporte público (Li *et al.*, 2018) no todas las investigaciones contrastan sus resultados en términos de validez estadística y representación de la realidad, a falta de información para comparar los datos. Adicionalmente, se resalta la importancia de la información de campo, considerando que los estudios que realizan la validación de la información, emplean los datos de encuestas y de seguimiento a usuario (Munizaga and Palma, 2012) y otros comparan la información estimada con sistemas AFC de ingreso y salida para validar su aceptación (Alsger, 2017).

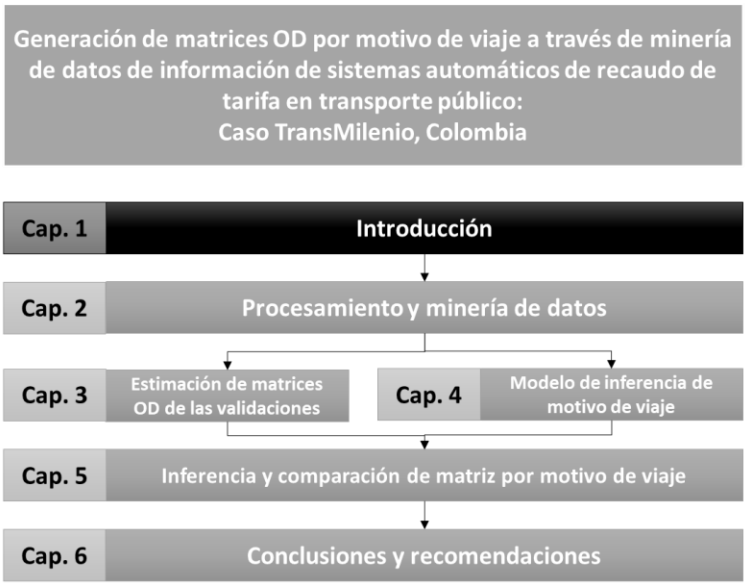
Finalmente, no se evidencia el uso de modelos de tipo logit multinomial para obtener matrices por motivo de viaje sobre el sistema TransMilenio.

1.3 Metodología

Como metodología para el desarrollo del presente trabajo, se considera el desarrollo de 6 capítulos que de forma secuencial permiten lograr el objetivo del trabajo de grado. La secuencia metodológica se presenta en la Figura 1-1, en donde se comienza con el presente numeral, que desarrolla la introducción y el estado del arte, con investigaciones

realizadas sobre el tema, la aplicación actual que realiza el sistema troncal del SITP (TransMilenio) y la importancia de completar estas matrices con información asociada al motivo del viaje.

Figura 1-1 Alcance del trabajo



Fuente: elaboración propia.

Seguido a la introducción, se da inicio al procesamiento y minería de datos, en donde se realiza la selección de información de interés, por medio de la revisión de datos asociados a TransMilenio, que incluyen la revisión del uso de tarjetas inteligentes, la programación y operación de rutas de TransMilenio, la encuesta origen destino en hogares (EODH) y datos del uso de suelo de Bogotá.

Posterior al tratamiento de los datos, se aplica el método de encadenamiento de viajes y sus heurísticas, que permiten inferir la información del destino del viaje y obtener las matrices de TransMilenio, para diferentes días de la semana, a nivel de estaciones y/o portales. A su vez, a partir del procesamiento y minería de datos, se obtienen las variables temporales y espaciales de mayor interés, que son de utilidad para la construcción de modelo logit multinomial, con las cuales se infiere el motivo de viaje.

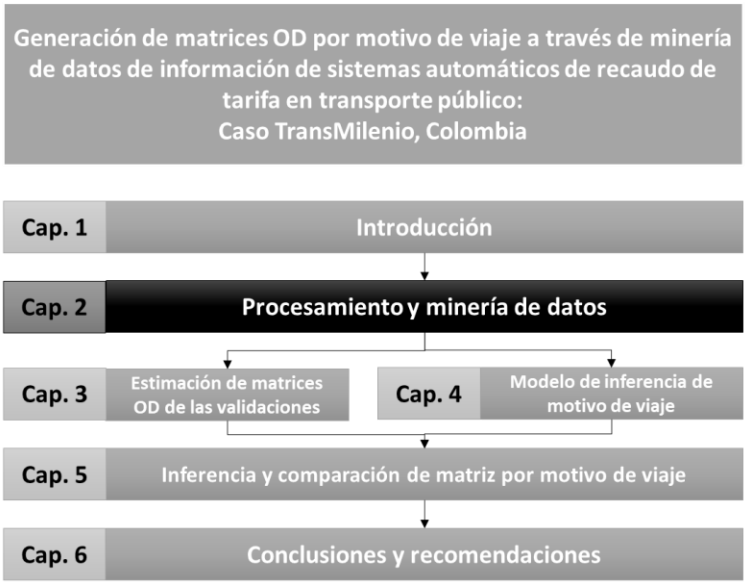
Sobre las matrices resultantes del encadenamiento de viajes, se aplica el modelo de inferencia de motivo de viaje, que permite obtener diferentes matrices a nivel de estaciones y/o portales, las cuales son comparadas con los resultados de la EODH. Finalmente, se presentan las conclusiones y resultados del trabajo, de mayor relevancia.

2. Procesamiento y minería de datos

2.1 Introducción

Este capítulo contiene la descripción del procesamiento y minería de datos, hacen parte estructural para la generación de matrices origen destino por motivo de viaje (ver Figura 2-1), ya que permite hacer el análisis, exploración y preparación de los datos temporales y espaciales de una semana de validaciones del AFC en el componente troncal del SITP (TransMilenio) y de las encuestas de viajes en hogares, junto con información de operación y programación de TransMilenio, y características del uso del suelo alrededor de estaciones, como insumo para la inferencia de matrices origen destino y la construcción de los modelos de motivo de viaje.

Figura 2-1 Metodología general – Capítulo 2



Fuente: elaboración propia.

El método de procesamiento y minería de datos, que se presenta en la Figura 2-2, aborda las fuentes de información que incluye, datos de TransMilenio, asociados a las validaciones de los Sistemas de Recaudo Automático (AFC) y la programación y operación de rutas de

TransMilenio, la información de la Secretaría Distrital de Movilidad (SDM), la encuesta origen destino en hogares (EODH) del año 2019 y las bases geográficas con información de uso de suelo de Infraestructura de Datos Espaciales para el Distrito Capital (IDECA), las cuales fueron procesadas en el software R para análisis estadísticos, SQL Server para la gestión de bases de datos, ArcGIS para los análisis de datos espaciales y Visum para el modelo de redes.

Figura 2-2 Método de procesamiento y minería de datos

Fuente: elaboración propia.

En el caso de los datos de TransMilenio, sobre las validaciones de los sistemas de recaudo automático, se realizó la selección de la semana de análisis a partir de la información de los 334 días de información desde el primero de enero al 30 de noviembre de 2018. Se agruparon los datos tanto para los días de la semana, como para los periodos de estacionalidad a lo largo del año. Se corroboró la independencia entre los grupos de datos, aplicando pruebas de hipótesis paramétricas de diferencias de medias, para corroborar la independencia entre los grupos de datos y así realizar la selección de una semana típica para el análisis, sobre los cuales se corrobora la normalidad de los datos por medio de un Q-Q Plot y la prueba Shapiro-Wilk.

Sobre la semana seleccionada, se obtuvieron las características asociadas al comportamiento de las entradas a TransMilenio, su distribución horaria a lo largo de los diferentes días y su diferencia con respecto a cada día de la semana; a su vez, se describe su comportamiento espacial por medio de la visualización de la distribución de entradas en

cada una de las estaciones, según la línea de la entrada y hora; y se obtuvo el coeficiente de asimetría de la distribución horaria, para describir la concentración de los datos hacia horas de la mañana y horas de la tarde por línea.

Con la información de programación de TransMilenio, se construyó la matriz de distancias entre paradas para aquellos pares origen-destino (OD) que se pueden conectar de forma directa a través de una sola ruta; esta información se complementó con las velocidades que se obtienen de la operación de la semana seleccionada, validando por medio de una prueba de hipótesis de diferencia de medias, que para varias franjas horarias la velocidad es la misma. Debido a que no todos los pares OD entre estaciones y/o portales se pueden completar con una sola ruta, se elaboró un modelo de redes para obtener la información de los tiempos faltantes.

Sobre la base de datos de etapas del viaje de la EODH – 2019, se realizó la codificación de estaciones y portales, y se complementó con la información asociada al viaje, como la hora de inicio y finalización, motivo y medio principal usado. A partir de esta información, se realizó un análisis espacial y temporal, diferenciado para cada uno de los motivos de viaje, revisando el comportamiento de las variables de hora de inicio del viaje, duración del viaje y duración de la actividad.

A su vez, se presenta un análisis espacial de tipología de suelo y equipamientos alrededor de estaciones de TransMilenio, para determinar las características espaciales, haciendo uso de la información geográfica contenida en las bases de IDECA, sobre los metros cuadrados de uso de suelo asociados a cada estación y portal.

Finalmente, se presentan una discusión y conclusiones sobre el método de procesamiento y minería de datos.

2.2 Datos del sistema TransMilenio

2.2.1 Validaciones de los Sistemas de Recaudo Automático

La base de datos utilizada del uso de tarjetas inteligentes en las estaciones y portales de TransMilenio, se encuentra almacenada en 334 archivos, cada uno para un día de

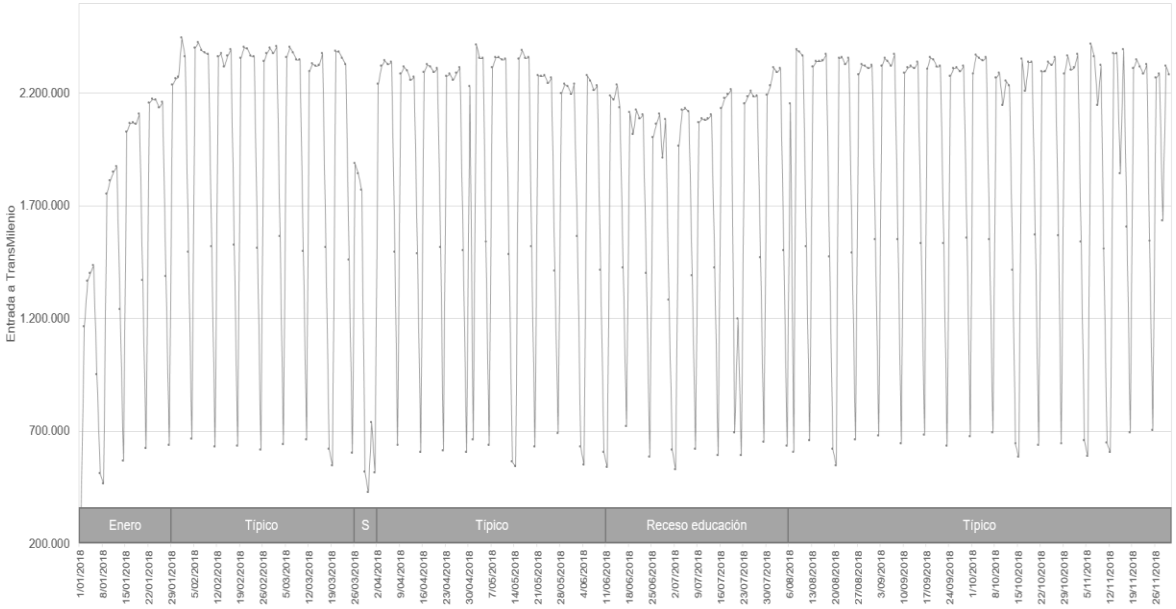
operación, con registros que van desde el primero de enero al 30 de noviembre del 2018, en promedio con cerca de 1,8 millones de registros por día, que en conjunto conforman una base de más de 609 millones de usos de tarjeta del componente troncal del SITP (TransMilenio).

La base incluye 14 variables, las cuales registran para cada uno de los validadores[2], información del uso de cada una de las tarjetas inteligentes asociadas a la fecha y hora, el emisor y operador de la tarjeta, el nombre de la línea, estación, validador y código del dispositivo, tipo de tarjeta, nombre del perfil del usuario, número único de identificación de la tarjeta, información sobre saldo previo, valor pagado y saldo después del uso.

Esta información sirve para definir la semana de datos a utilizar, para ello, se toma el número de usos de las tarjetas inteligentes en los validadores de TransMilenio. La Figura 2-3 muestra los cambios puntuales en los ingresos al sistema para las semanas del año 2018; se aprecia la variación asociada a cada una de las semanas, con reducciones cíclicas de la demanda, producidas por cambios de día hábil pasando de viernes a sábado, domingo y festivos. De igual manera se observan días donde se presentaron anomalías en la movilidad de la ciudad producidas por cierres de vías, marchas, y disturbios, entre otros.

[2] Validadores, dispositivos electrónicos que registran el uso de las tarjetas inteligentes por medio de contacto, ubicados en estaciones y portales del sistema troncal del SITP (TransMilenio).

Figura 2-3: Histórico de comportamiento de la demanda 2018



Fuente: elaboración propia.

Así mismo, se observa un crecimiento de la demanda progresiva en el mes de enero hasta comienzos del mes de febrero, desde este punto se mantiene constantes hasta el mes de marzo, donde se presenta una reducción de la demanda en la Semana Santa, un retorno a la demanda habitual hasta la mitad de junio, desde este punto se presenta una nueva reducción hasta la mitad de julio, en los periodos de receso escolar y de universidades, para retomar la demanda típica hasta el mes de noviembre.

Debido a la cantidad de datos y que se observan comportamientos cíclicos, se realizan agrupaciones, primero considerando tendencias similares entre los diferentes tipos de día de una semana en: hábiles (lunes a viernes), sábado, domingo y festivo.

Se corrobora que entre los grupos las entradas al sistema difieren entre días, mediante la aplicación de una prueba de hipótesis de medias, donde la hipótesis nula es igual a cero:

$$H_o: \mu_1-\mu_2 = 0$$

El estadístico de la prueba es (Kumar Molugaram Rao, 2017):

$$Z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Donde:

X: media

S: desviación estándar

n: tamaño

Para este caso se rechaza la hipótesis nula a un nivel de confianza del 95%, cuando el valor de Z es mayor que el valor de la tabla, con un valor de $Z_{\alpha/2} = 1{,}96$, lo que implica que se rechaza la hipótesis nula cuando el valor de Z estimado (Z*) es mayor al Z de tabla:

$$Z^* > Z$$

Al aplicar la prueba paramétrica, sobre la cual se corrobora la normalidad de los datos en la Figura 2-4, se rechaza la hipótesis nula para la diferencia de medias entre todos los tipo día, debido a que los Z*, son mayores al Z de referencia, como se presenta en la Tabla 2-1, donde se puede observar que la mayor diferencia entre comportamiento de días ocurre entre un día hábil y un domingo; por otra parte, la menor diferencia se presenta entre un domingo y un día Festivo. Por lo tanto, hay una diferencia significativa entre las medias de las muestras para los diferentes tipos de día.

Tabla 2-1 Valores de Z* entre tipo día

Z*	$\overline{X}$	S	n	Hábil	Sábado	Domingo	Festivo
Hábil	2.244.054	182.670	224	0,0	30,9	119,1	73,1
Sábado	1.452.387	154.393	47		0,0	35,2	30,3
Domingo	634.686	38.879	45			0,0	4,3
Festivo	545.186	83.917	18				0,0
Total	1.824.265	665.424	334				

Fuente: elaboración propia.

De igual forma, se aplica la prueba de hipótesis para las diferentes estacionalidades del año, aplicada solamente a los días hábiles diferenciando en días típicos, enero, semana santa y receso de educación, obteniendo que se rechaza la hipótesis nula para la mayoría de los casos, como se observa en la Tabla 2-2, excepto para enero y semana santa.

Tabla 2-2 Valores de Z* entre diferentes estacionalidades para días hábiles

Estacionalidades	$\overline{X}$	S	n	Típico	Enero	Semana Santa	Receso educación
Típico	2.332.688	43.310	149	0,0	5,7	17,8	14,2

Estacionalidades	$\overline{X}$	S	n	Típico	Enero	Semana Santa	Receso educación
Enero	1.930.296	321.903	21		0,0	1,3	-3,2
Semana Santa	1.833.250	48.202	3			0,0	-10,6
Receso educación	2.155.155	85.519	51				0,0
Hábil	2.244.054	182.670	224				

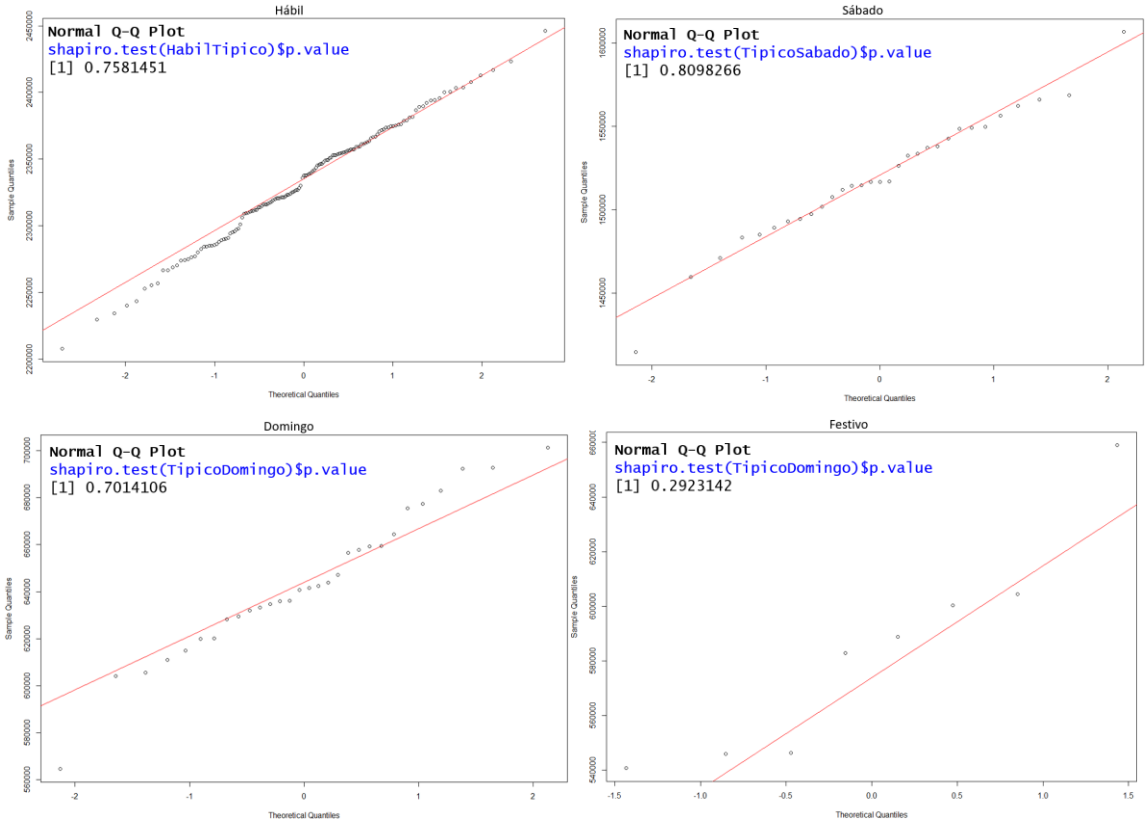
Fuente: elaboración propia.

Considerando que la prueba paramétrica de hipótesis de diferencia de medias supone que los datos se distribuyen de forma normal, el gráfico Q-Q Plot y la prueba de Shapiro-Wilk permiten corroborar que la forma de distribución de los datos se asemeja con la forma teórica de la distribución normal, en donde:

Ho: la muestra proviene de una población con distribución normal

Obteniendo los resultados que se presentan en la Figura 2-4, aplicada para días con estacionalidad típica en los diferentes días de la semana, donde no se rechaza la hipótesis nula, ya que el valor de p-value es mayor a 0,05, lo que significa que los datos se distribuyen de forma normal.

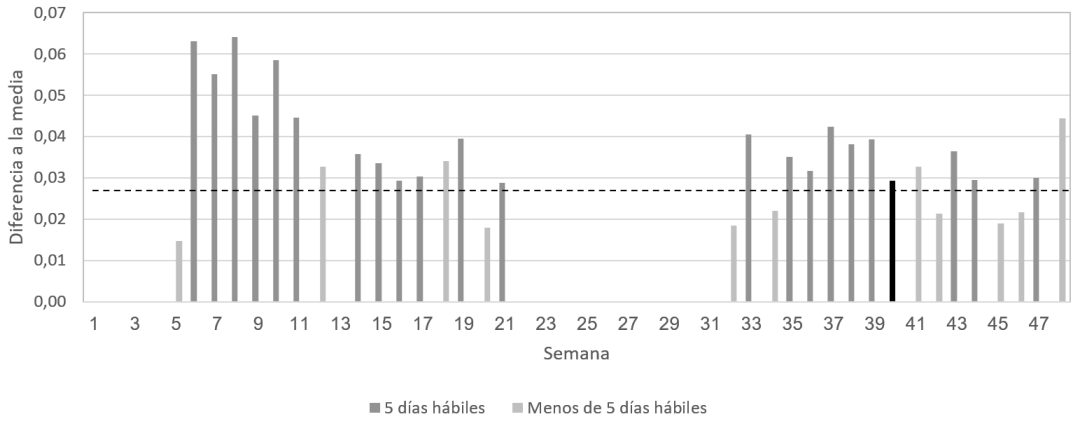
Figura 2-4 Prueba de normalidad – Q-Q Plot y Shapiro-Wilk



Fuente: elaboración propia.

Identificados los conjuntos de días con estacionalidad típica, se realiza la selección de la semana de demanda sobre la cual se realizan los diferentes análisis, para esto se conoce la media del conjunto de días con estacionalidad típica y se compara con la media de ingresos para cada una de las semanas, obteniendo la Figura 2-5, donde se observa que para las semanas completas, de 5 días (sin días festivos), la menor diferencia se presenta en la semana 40, comprendida entre el primero (1) al siete (7) de octubre de 2018.

Figura 2-5 Semana con menor variación a la media

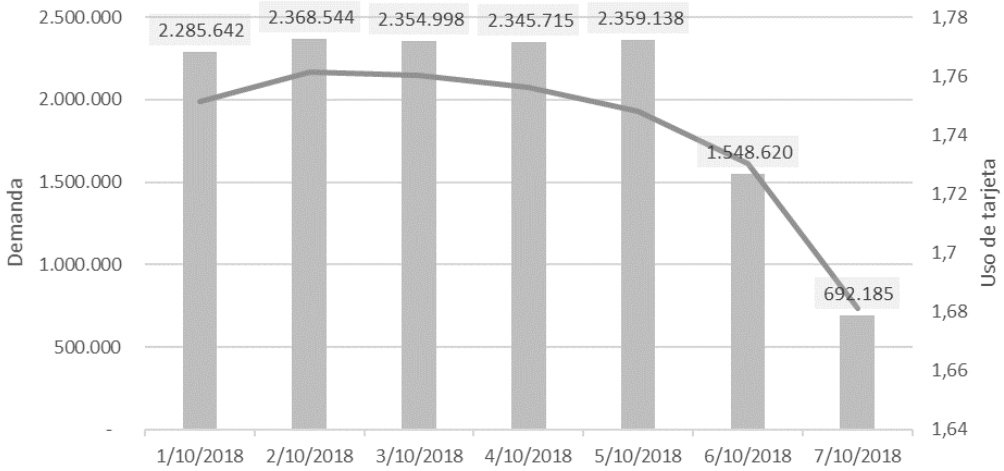


Fuente: elaboración propia.

2.2.2 Análisis temporal del uso de tarjetas inteligentes

De la semana de análisis seleccionada en el numeral anterior, se cuenta con una base con 13,9 millones de registros de usos de tarjetas inteligentes en TransMilenio, con la cual se construye el histograma de entradas desde el primero (1) de octubre al siete (7) de octubre que se presenta en la Figura 2-6, observándose que para días hábiles las entradas están en cerca de 2,3 millones, para el sábado es de 1,5 millones y para el domingo de 692 mil; a su vez, se obtiene que en promedio la cantidad de veces que se usa una misma tarjeta en el día, oscila entre 1,67 y 1,77 veces.

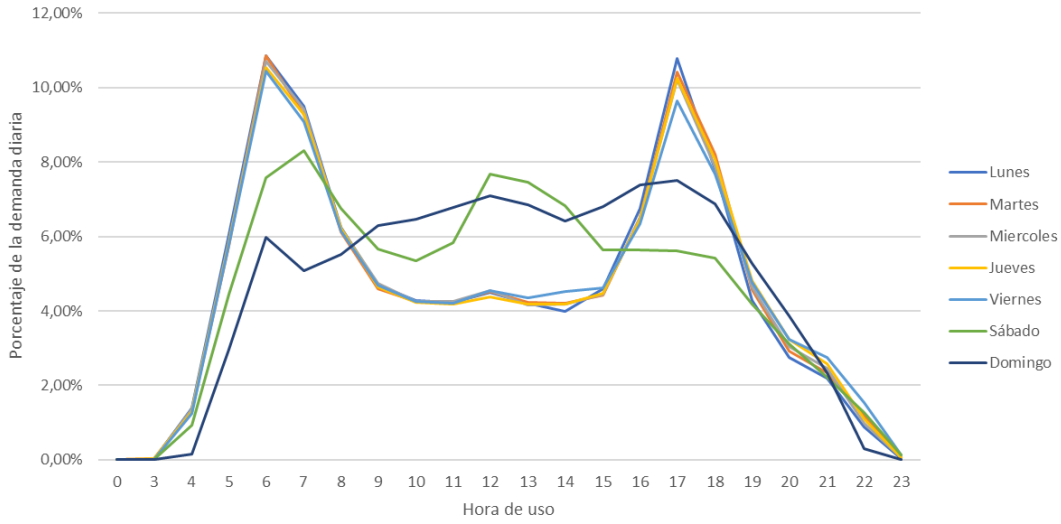
Figura 2-6 Histograma semana de análisis de uso de tarjetas inteligentes



Fuente: elaboración propia.

Por otra parte, la Figura 2-7 presenta el comportamiento horario de la demanda para cada uno de los días de la semana, para el cual se observa que el comportamiento horario es similar en los días hábiles de lunes a viernes, donde se distinguen dos picos marcados a las 6:00 horas y a las 17:00 horas, cada uno con más del 10% de la demanda diaria. El comportamiento de la demanda del sábado tiene su principal pico en horas de la mañana y un pico adicional al medio día, concentrando cerca del 8% de la demanda del día en las dos horas; para el domingo se observa una distribución uniforme sin picos pronunciados.

Figura 2-7 Histograma de demanda diaria del uso de tarjetas inteligentes

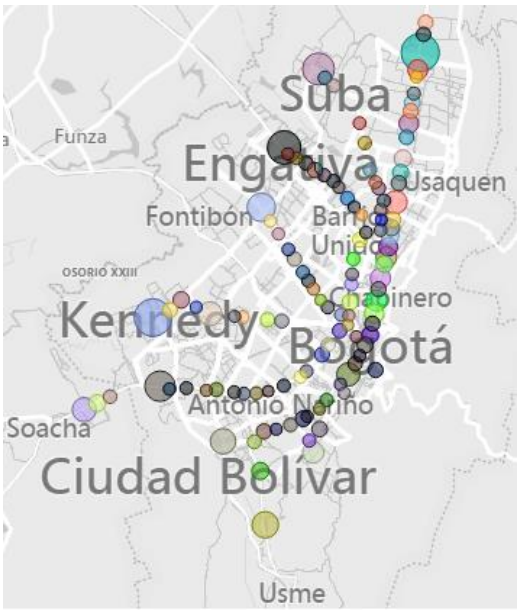


Fuente: elaboración propia.

2.2.3 Comportamiento espacial del uso de tarjetas inteligentes

El uso de las tarjetas inteligentes se registra en 1.184 dispositivos diferentes, ubicados en 378 accesos de estaciones, distribuidos en 154 estaciones y/o portales, a lo largo de las 12 líneas que conforman el sistema TransMilenio. Para la semana de información seleccionada, se presenta el análisis espacial de concentración de pasajeros en estaciones de la Figura 2-8, por medio del número de entradas por estación. Las estaciones y portales con mayor número de entradas son el Portal Norte y Portal Américas con cerca de 95 mil ingresos al día cada una, seguido por cabecera Calle 80, Portal Suba y Portal Sur. Las estaciones con menor número de ingresos reportados son las estaciones tipo corral[3] de la carrera 77 y Avenida ciudad de Cali, seguido por San Martín y Gratamira, estas últimas con un poco más de mil ingresos al día.

Figura 2-8: Número de entradas en estaciones en un día



Número	Recuento	Estacion
1	96501	(02000) Cabecera Autopista Norte
2	94636	(05000) Portal Américas
3	80261	(04000) Cabecera Calle 80
4	74712	(03000) Portal Suba
5	69509	(07000) Portal Sur
6	62855	(06000) Portal Eldorado
7	55705	(09000) Cabecera Usme
8	52337	(08000) Portal Tunal
9	47344	(07503) SAN MATEO
10	46805	(05100) Banderas P. Central
11	44651	(09110) Avenida Jimenez
12	44104	(02300) Calle 100
13	38266	(10000) Portal 20 de Julio
14	37231	(09118) Marly
15	36204	(09122) Calle 72
150	1741	(09108) Hospital
151	1484	(03004) GRATAMIRA
152	1127	(03014) SAN MARTIN
153	449	(50006) Corral Carrera 77
154	19	(50004) Corral Avenida Ciudad de Cali

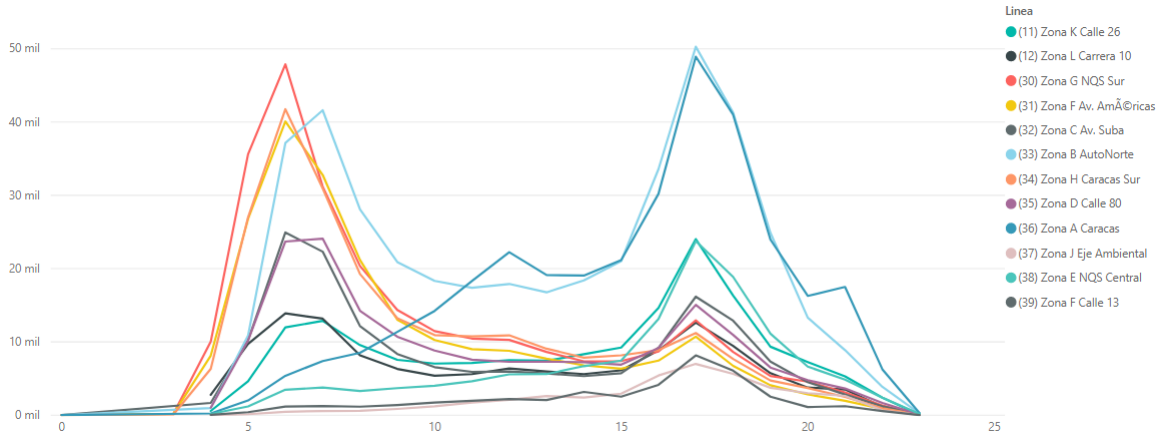
Fuente: elaboración propia.

[3] Las estaciones tipo corral, son estaciones asociados a una estación principal, donde llegan rutas alimentadoras, las estaciones de este tipo son: Calle 40 Sur, Molinos, Avenida Ciudad de Cali, Carrera 77 y General Santander.

Para evidenciar si existe concentración o diferencias en la distribución de entradas a TransMilenio, se emplea el cálculo del coeficiente Gini como medida de concentración o desconcentración del número de entradas por estación (Gujarati and Porter, 2010), obteniendo un valor de 0,49 lo que indica que no es perfectamente concentrado (0), ni perfectamente desconcentrado (1); siendo esto que, el 80% de las entradas se concentra en el 47% de las estaciones (73 estaciones) y el 95% de la demanda se concentra en el 77% de las estaciones (118 estaciones).

El histograma de la demanda horaria diferenciada por tipo de línea, se observa el comportamiento que se muestra en la Figura 2-9, en donde algunas líneas presentan una mayor concentración en horas de la mañana, como ocurre para la línea NQS Sur y la línea Av. Américas, y en otras concentraciones en horas de la tarde como ocurre con la línea de la Av. Caracas.

Figura 2-9 Histograma por línea



Fuente: elaboración propia.

Para identificar sí la concentración de los datos se da en horas de la mañana (izquierda) o en horas de la tarde (derecha), se utiliza el coeficiente de asimetría de la distribución calculado para el promedio de la hora de ingreso por línea, obteniendo la Tabla 2-3, donde un coeficiente mayor que cero, muestra una concentración de los datos en hora de la mañana, y menor que cero, una concentración de los datos en horas de la tarde; a su vez, una mayor distancia absoluta del valor de cero, refleja mejor concentración de la demanda en una franja horaria; por ejemplo, la línea Av. Américas que registra 225 mil ingresos en el día, tiene una media de ingreso a las 10 horas, una mediana a las 8,2 horas y un

coeficiente de asimetría de 0,9, el valor más alejado del cero en el rango positivo, lo que significa que su pico se concentra en horas de la mañana.

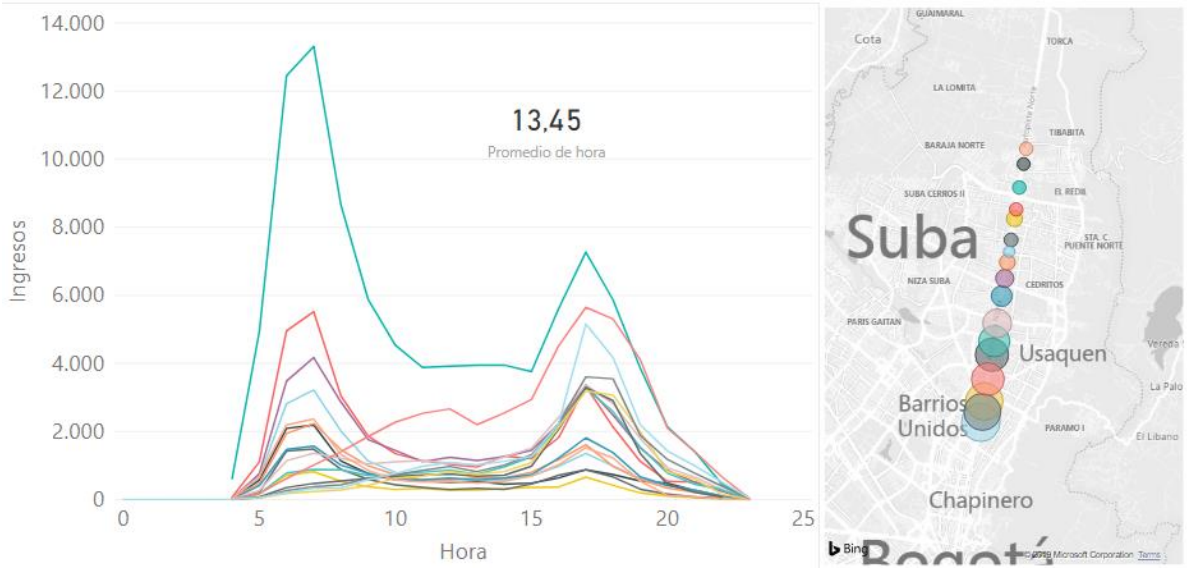
Tabla 2-3: Coeficiente de asimetría por línea

Línea	Ingreso	$\overline{X}$, h	Mediana, h	Coeficiente de asimetría	σ, h	Núm. estaciones
(11) Zona K Calle 26	173.487	14,0	15,2	-0,3	4,8	14
(12) Zona L Carrera 10	132.655	12,5	12,4	0,1	5,1	11
(30) Zona G NQS Sur	260.489	10,2	8,3	0,8	4,7	18
(31) Zona F Av. Américas	225.001	10,0	8,2	0,9	4,5	10
(32) Zona C Av. Suba	170.173	12,0	10,7	0,3	5,1	14
(33) Zona B Auto Norte	428.873	13,4	14,2	-0,1	4,8	17
(34) Zona H Caracas Sur	237.714	10,3	8,6	0,8	4,6	18
(35) Zona D Calle 80	182.753	12,0	10,8	0,4	4,9	16
(36) Zona A Caracas	329.933	15,7	16,7	-0,5	3,9	14
(37) Zona J Eje Ambiental	45.703	16,4	17,1	-0,7	3,5	2
(38) Zona E NQS Central	138.134	16,1	17,3	-0,8	4,1	13
(39) Zona F Calle 13	43.629	15,3	16,7	-0,7	3,9	7

Fuente: elaboración propia.

Adicionalmente, cuando se analiza espacialmente el promedio de la hora de entrada para cada una de las estaciones, se observa el comportamiento de la Figura 2-10 para la línea de la Autopista Norte (Zona B), donde a medida que las estaciones se van acercando al centro de la ciudad la hora promedio de ingresos es mayor (círculos más pequeños), debido a que en las zonas más alejadas del centro de la ciudad, son principalmente suelos asociados a viviendas, lo que implica que sus primeras entradas se dan en horas de la mañana y a medida que se acercan al centro de la ciudad, donde se desarrollan las principales actividades económicas de la ciudad, las entradas se dan en horas de tarde y noche (círculos más grandes) para regresar a las viviendas.

Figura 2-10: Hora promedio de ingreso a estaciones de la línea de la Autopista Norte

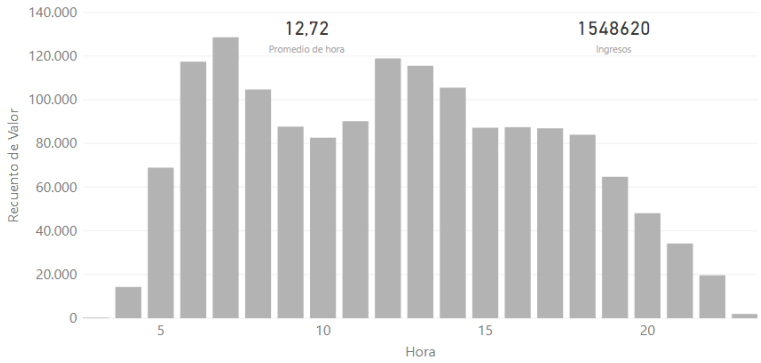


Fuente: elaboración propia.

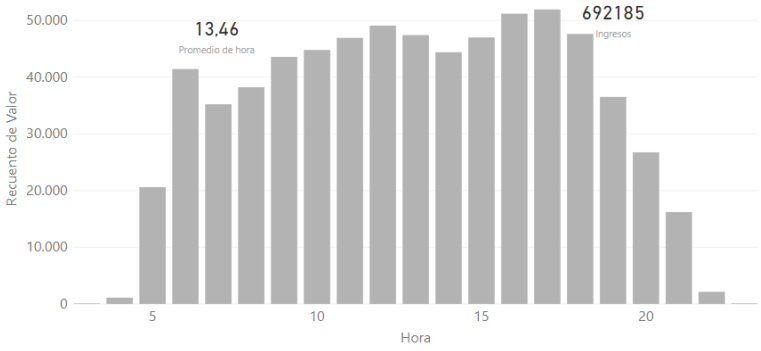
De igual forma se observa un cambio en el comportamiento para los días sábados (Figura 2-11 – a), donde el histograma diario muestra un comportamiento más uniforme a lo largo del día, similar a lo que ocurre para el día domingo (Figura 2-11 – b) y donde ya no es tan visible la diferenciación en el patrón de los viajes, asociado a la hora media de viaje.

Figura 2-11: Histogramas horarios - sábado y domingo

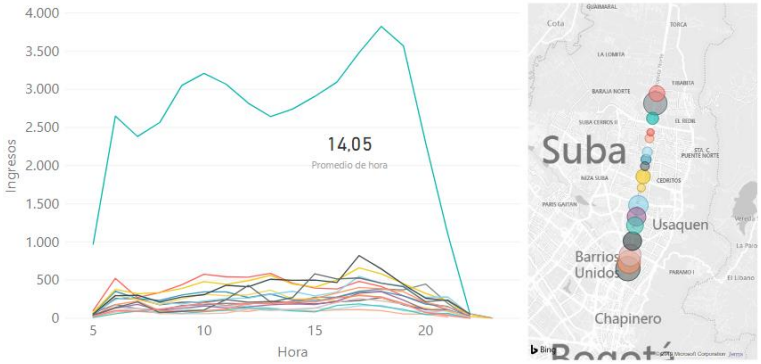
(a) Diario - sábado



(b) Diario – Domingo



(c) Línea – Autopista Norte



Fuente: elaboración propia.

2.2.4 Matriz de duración del viaje de TransMilenio

Como fue expuesto en la sección 1.2 Estado del arte, una de las variables de interés para la obtención de matrices de viaje por motivo de viaje, es la duración del viaje; debido a que esta característica no está registrada de forma directa con el uso de las tarjetas inteligentes, por esta razón, se utiliza la información de programación y operación del sistema TransMilenio.

Por una parte, se cuenta con la información de la matriz de distancia acumulada de las rutas ofertadas en el año 2018 (ver Tabla 2-4.), que contiene la información de la longitud acumulada entre cada una de las paradas, para cada una de las rutas del sistema.

Tabla 2-4 Muestra de base de datos de la matriz de distancias vigentes TransMilenio

Línea	Sublínea	Ruta	Nodo	Nombre Nodo	Posición	Tipo Servicio
1	856	[1169] C15 PORTAL SUBA	1808	Portal Tunal T3	0	[1] TRONCAL
1	856	[1169] C15 PORTAL SUBA	1805	Parque A - 4	1293	[1] TRONCAL
1	856	[1169] C15 PORTAL SUBA	1774	Biblioteca A - 3	1874	[1] TRONCAL
1	856	[1169] C15 PORTAL SUBA	1785	Calle 40 S. C - 3	3573	[1] TRONCAL
1	856	[1169] C15 PORTAL SUBA	1800	Olaya A - 3	4918	[1] TRONCAL

Fuente: TransMilenio S.A.

Considerando que la variable de interés es la duración del recorrido entre estaciones, es necesario usar la base de datos de viajes realizados del sistema, que registra para cada uno de los viajes de TransMilenio, información de la hora de inicio y finalización del viaje para cada uno de los días de la semana. A partir de la unificación de estas dos bases se obtiene la velocidad promedio para cada uno de los viajes realizados.

Debido a que los cambios de demanda a lo largo del día pueden provocar variaciones en la velocidad de viaje, se aplica una prueba de hipótesis de diferencias de medias entre la duración de la ruta para diferentes franjas horarias, con el fin de corroborar si es suficiente con obtener una velocidad por ruta para todo el día o se requieren diferenciar las franjas horarias.

Al aplicar la prueba de hipótesis de diferencia de medias, se obtienen resultados como los que se presentan en la Tabla 2-5, para la ruta G12, donde el valor de Z* solamente es más grande que Z (1,96) para las 23 horas y que es grande para después de las 16 y 17 horas, este comportamiento se repite para las 86 rutas que son programadas en la semana del primero (1) al siete (7) de octubre; de esta forma se corrobora que la media de las velocidades no varía de forma significativa entre diferentes franjas horarias para una misma ruta.

Tabla 2-5 Prueba de hipótesis de diferencia de medias para duración de viaje entre horas del día

Hora	x	s	n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
5	18,4	9,4	21	0	1,176	1,505	0,397	0,937	1,449	1,148	1,101	1,229	0,892	1,141	2,045	2,472	1,637	1,865	1,405	1,345	1,476	-2,56
6	15,6	6,3	23		0	0,337	-1,05	-0,29	0,32	-0,1	-0,12	0,024	-0,39	-0,06	1,09	1,691	0,481	0,833	0,204	0,155	0,341	-4,63
7	15,0	4,9	26			0	-1,55	-0,66	0,009	-0,48	-0,49	-0,33	-0,79	-0,42	0,903	1,672	0,15	0,591	-0,16	-0,19	0,026	-5,45
8	17,5	5,9	22				0	0,731	1,439	1,027	0,954	1,135	0,67	1,008	2,34	3,106	1,788	2,09	1,408	1,306	1,485	-3,76
9	16,1	5,7	17					0	0,616	0,214	0,178	0,328	-0,09	0,234	1,388	1,992	0,813	1,141	0,524	0,465	0,642	-4,28
10	15,0	5,1	19						0	-0,45	-0,46	-0,31	-0,74	-0,39	0,797	1,43	0,122	0,517	-0,15	-0,19	0,015	-5,13
11	15,7	5,2	21							0	-0,03	0,129	-0,32	0,033	1,309	2,025	0,656	1,03	0,335	0,276	0,475	-4,8
12	15,8	5,6	20								0	0,153	-0,28	0,06	1,268	1,918	0,649	1,004	0,349	0,293	0,482	-4,62
13	15,5	5,1	19									0	-0,43	-0,09	1,143	1,817	0,491	0,867	0,191	0,139	0,336	-4,85
14	16,3	5,7	20										0	0,333	1,562	2,233	0,968	1,305	0,651	0,582	0,767	-4,34
15	15,7	5,8	21											0	1,191	1,825	0,572	0,928	0,28	0,227	0,416	-4,65
16	13,8	4,4	23												0	0,629	-0,84	-0,33	-1,05	-1,04	-0,8	-6,15
17	13,1	3,7	35													0	-1,74	-1,04	-1,83	-1,74	-1,46	-7,05
18	14,8	4,9	40														0	0,498	-0,32	-0,35	-0,11	-5,82
19	14,2	4,3	23															0	-0,74	-0,75	-0,51	-5,96
20	15,2	4,6	22																0	-0,05	0,17	-5,3
21	15,3	5,4	25																	0	0,206	-5,09
22	15,0	4,6	17																		0	-5,22
23	25,1	4,2	7																			0

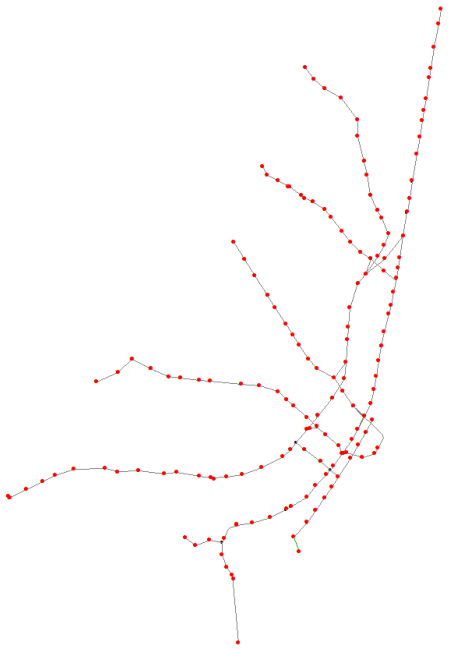
Rojo: valor de Z* mayor a 1,96. Azul: valor de Z* menor a 1,96.

Fuente: elaboración propia.

Con esta información es posible obtener la duración de viajes para los más de 7 mil pares origen destino (OD) que utilizan solamente una ruta para conectarse; no obstante, la matriz de viajes entre todos los pares OD, se construye con la unión de varias rutas, ya que TransMilenio cuenta con 154 estaciones y/o portales, para una matriz de más de 24 mil pares OD posibles. Por esta razón, para completar la información de los pares que usan más de una ruta, se construye el modelo de redes que se presenta en la Figura 2-12 y se

calcula la velocidad promedio de todas las rutas para obtener la matriz de duración del viaje entre todas las estaciones y/o portales faltantes.

Figura 2-12 Modelo de redes y matriz de distancias

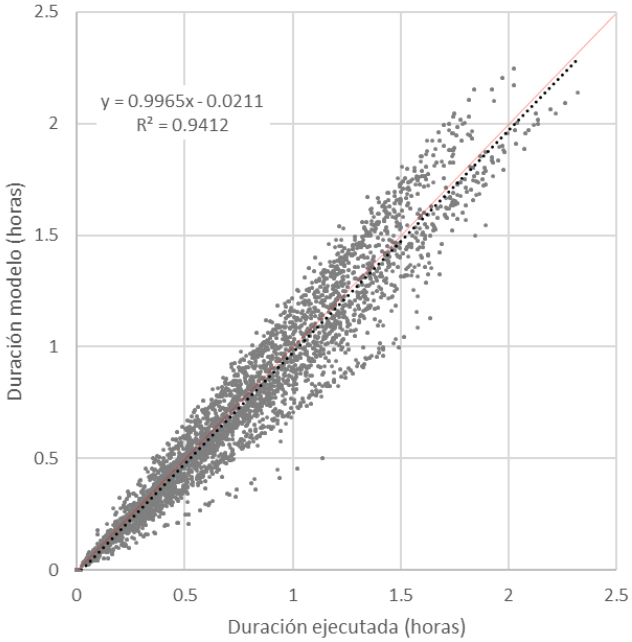


	2000	2001	2101	2102	2103	2104	2105
2000	0.00	0.97	0.90	1.28	2.16	2.65	3.07
2001	0.97	0.00	1.87	2.25	3.13	3.62	4.04
2101	0.90	1.87	0.00	0.38	1.26	1.75	2.17
2102	1.28	2.25	0.38	0.00	0.88	1.37	1.79
2103	2.16	3.13	1.26	0.88	0.00	0.49	0.91
2104	2.65	3.62	1.75	1.37	0.49	0.00	0.42
2105	3.07	4.04	2.17	1.79	0.91	0.42	0.00
2200	3.75	4.72	2.85	2.47	1.59	1.10	0.68
2201	4.48	5.45	3.57	3.19	2.31	1.83	1.40
2202	5.61	6.58	4.71	4.33	3.45	2.96	2.54
2204	6.32	7.29	5.42	5.04	4.16	3.67	3.25
2205	6.96	7.93	6.05	5.67	4.79	4.31	3.89
2300	7.95	8.92	7.04	6.66	5.78	5.30	4.87
2302	8.87	9.84	7.96	7.59	6.71	6.22	5.80
2303	9.29	10.26	8.39	8.01	7.13	6.64	6.22
2304	9.72	10.69	8.82	8.44	7.56	7.07	6.65
2502	1.60	0.63	2.50	2.88	3.76	4.25	4.67
3000	20.15	21.12	19.25	18.87	17.99	17.50	17.08
3001	19.58	20.55	18.68	18.30	17.42	16.93	16.51
3002	18.95	19.92	18.05	17.67	16.79	16.30	15.88
3003	18.17	19.14	17.27	16.89	16.01	15.52	15.10
3004	17.05	18.03	16.15	15.77	14.89	14.41	13.98
3005	16.36	17.34	15.46	15.08	14.20	13.72	13.29
3006	15.29	16.26	14.39	14.01	13.13	12.64	12.22
3007	14.73	15.70	13.83	13.45	12.57	12.08	11.66
3009	13.89	14.86	12.98	12.60	11.72	11.24	10.81
3010	13.17	14.14	12.27	11.89	11.01	10.52	10.10
3011	12.80	13.77	11.90	11.52	10.64	10.15	9.73
3012	12.09	13.07	11.19	10.81	9.93	9.45	9.02
3013	11.59	12.56	10.68	10.31	9.43	8.94	8.52
3014	11.06	12.03	10.16	9.78	8.90	8.41	7.99

Fuente: elaboración propia.

Para validar los datos del modelo de redes, se utilizan las duraciones de viaje para los 7 mil pares con tiempos conocidos y se comparan con la duraciones del viaje obtenida con el modelo de redes, obteniendo los ajustes de la recta que se presentan en la Figura 2-13, con un R^2 del 94%, un intercepto muy cercano a cero y una pendiente de 0,99 muy cerca a uno.

Figura 2-13 Tiempos entre estaciones observados versus estimados



Fuente: elaboración propia.

2.3 Viajes en TransMilenio según encuesta origen destino en hogares

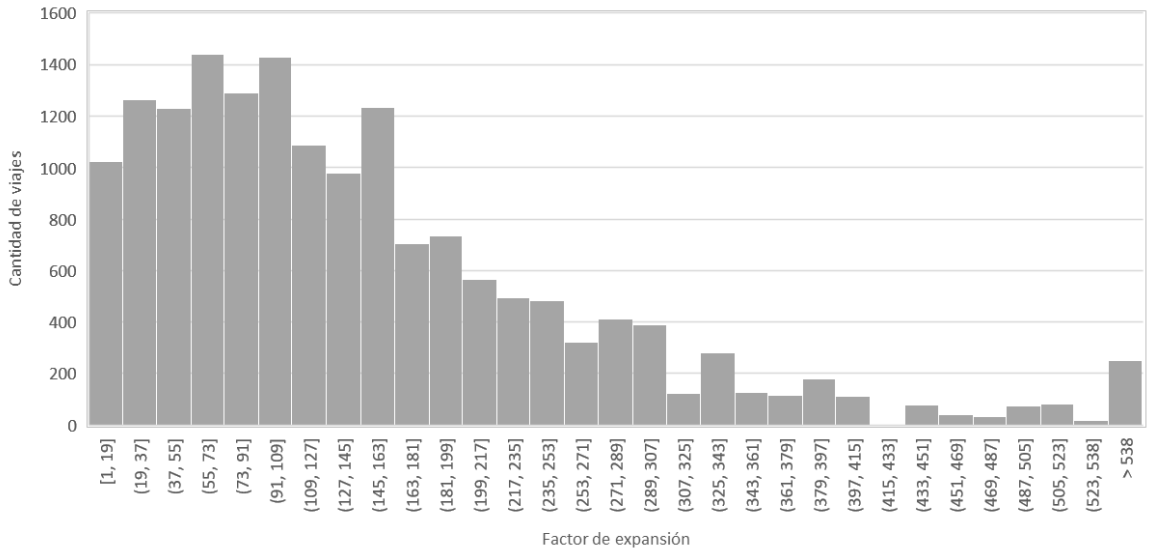
La ciudad de Bogotá D.C. desde la Secretaría Distrital de Movilidad, realizó una inversión en el año 2018 de 5,3 mil millones de pesos (Secop II, 2018), con el fin de contratar una consultoría, que dentro de sus alcances, empleara métodos tradicionales de encuestas de hogares para la recolección de datos de viajes, y así poder tener una caracterización de la demanda de transporte y las MOD.

Como resultado de esta contratación en el año 2019 se entrega la encuesta origen destino en hogares EODH del año 2019 de Bogotá D.C.[4]., la cual es la fuente de información más reciente que recoge los datos asociada a una muestra de 21.828 hogares, con datos de las personas, viajes y sus etapas.

[4] Fuente: https://www.simur.gov.co/portal-simur/datos-del-sector/encuestas-de-movilidad/, consulta realizada el 9/5/2020. Encuesta de movilidad 2019

Sobre esta información se hace la revisión y extracción de los datos asociados a TransMilenio. De los cerca de 19 millones de viajes que se realizan para Bogotá y los municipios cercanos, la encuesta reporta que se aplicaron 16.589 encuestas, donde el principal medio usado fue TransMilenio, presentando el histograma de factores de expansión de la Figura 2-14, que en total suman 2’489.738 viajes. A pesar de que las base indica que la fecha del viaje incluye sábados y domingo, se confirma en el Anexo A, que toda la información se asocia a la demanda de un día hábil.

Figura 2-14 Histograma de factores de expansión medio de transporte principal TransMilenio



Fuente: elaboración propia, a partir de datos de la EODH.

Por otra parte, debido a que la encuesta no está estandarizada a nivel de estaciones para las etapas del viaje, se realizó el proceso de codificación de los nombres de las estaciones de la base EODH, según los código y nombres de la información de la base de las validaciones de TransMilenio, que cuenta con un listado de 154 estaciones; tomando el nombre de la estación en la que abordó el vehículo, con medio de transporte TransMilenio, donde se tienen 4.410 diferentes registros en el origen y 4.406 en el destino (ver Figura 2-15), con cero coincidencias con el nombre original, para solucionar esto, se apoya la codificación con el geocodificador de Google Maps, que permite armonizar las bases.

Figura 2-15 Muestra de estación de entrada y salida base etapas EODH

p20_Estacion_abordo_vehic
AV JIMENEZ
0
1 de mayo
1 de mayo
1 de mayo estaciÃ³n trasmilenio
10
100
100 autopista
100cl autopista norte
106
116
11-jun
124 con 7
127

p25_lugar_descensoD
EST.UNIVERSIDADES
:estaciÃ³n molinos
0
1
1 de mayo
1 mayo con 10
10 con 24
10 con 94
100
100 cl autopista norte
100 CON AU TM
100 CON AUTO NORTE
106
106 autopista al frente sin

Fuente: elaboración propia.

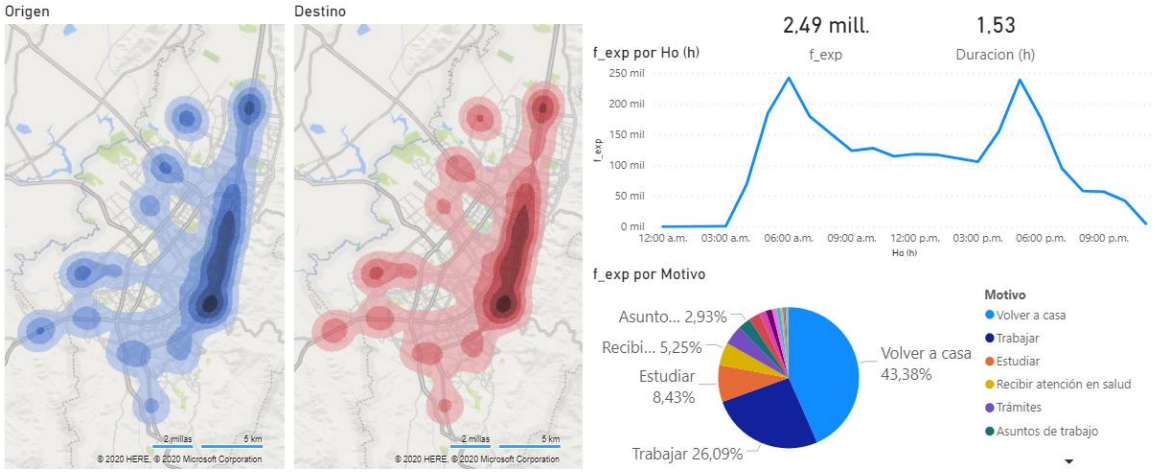
Esta información de las etapas del viaje es complementada con los datos asociados a hora de inicio, finalización y motivo del viaje; adicionalmente, se complementa la información de la encuesta con el cálculo de la variable duración de la actividad, la cual se calcula como la diferencia entre la hora de inicio del viaje siguiente menos la hora del viaje actual.

2.3.1 Información de etapas del viaje en TransMilenio

De los datos de los viajes de la EODH que se realizan en TransMilenio, es posible obtener información asociada a los minutos caminados antes de tomar el primer medio de transporte del viaje y la cantidad de cuadras caminadas, teniendo para TransMilenio en promedio 9 minutos y 5 cuadras caminadas.

Como se presentan en la Figura 2-16, la demanda de los viajes que usan TransMilenio, se dan en dos picos marcados a las 6 a.m. y a las 5 p.m., adicionalmente se observan los mapas de isócronas de viajes que usan TransMilenio, tanto para orígenes como para destino; con una hora promedio de un viaje de 1,5 horas; a su vez, se presenta un diagrama de torta donde el principal motivo de viaje es volver a casa con un 43,4%.

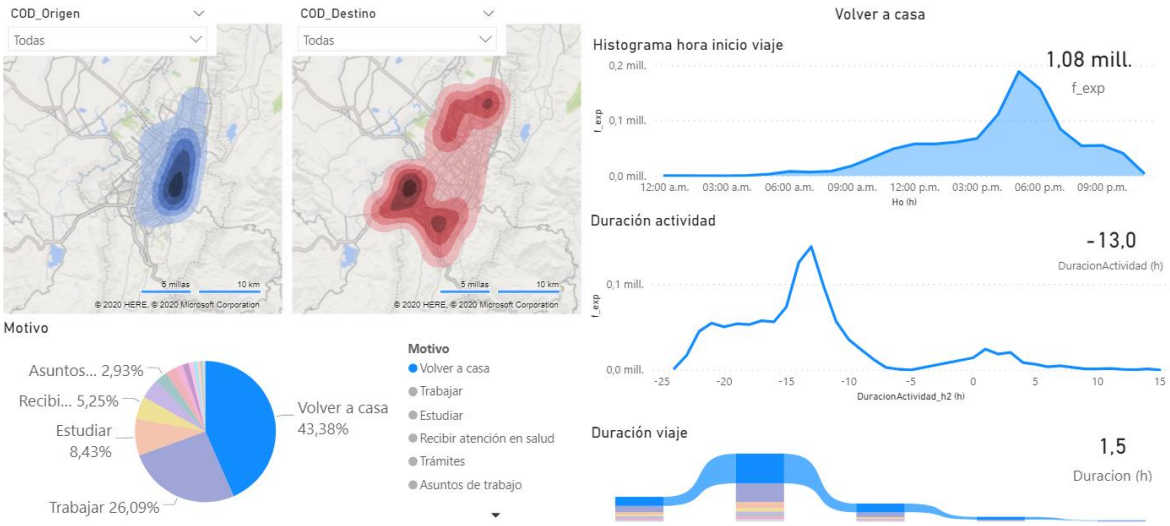
Figura 2-16 Isócronas para ingresos al modo TransMilenio EODH



Fuente: elaboración propia a partir de Geocodificación de estación en la que abordó el vehículo, con información EODH

Para el caso de la demanda de viajes con motivo “volver a casa”, se presenta en la Figura 2-17, que esta demanda representa el 43,4% de los viajes, con una hora de inicio del viaje concentrado a las 6 p.m.; a su vez, las entradas a TransMilenio con este motivo, se dan sobre la línea Av. Caracas, en el centro de la ciudad, y las salidas en los portales en la periferia de la ciudad, sumado a esto, se observa que la duración de la actividad del motivo “volver a casa” tiene una duración de 13,0 horas, en este caso el signo negativo representan el cambio de día, ya que gran parte de los viajes con este motivo, no tienen un viaje posterior en el mismo día; finalmente estos viajes tienen una duración promedio de 1,5 horas con poca variación con respecto a los demás motivos.

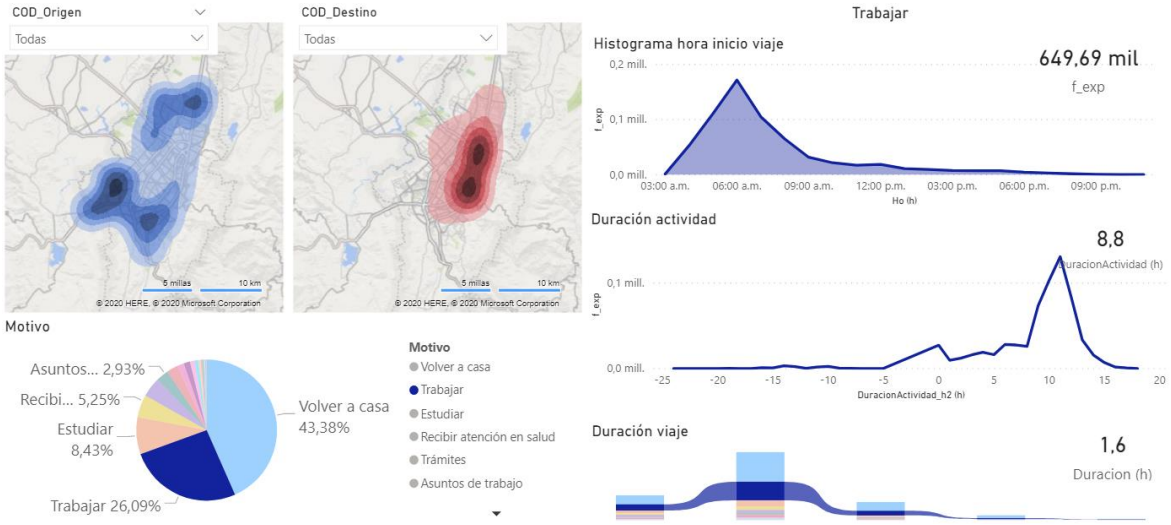
Figura 2-17 Isócronas, histogramas por motivo viaje – Motivo volver a casa



Fuente: elaboración propia a partir de Geocodificación de estación en la que abordó el vehículo, con información EODH

Al realizar el mismo ejercicio, para el motivo “trabajar”, el comportamiento es casi opuesto al de motivo volver a casa, ya que el pico de inicio del viaje en este caso se concentra principalmente en horas de la mañana (Figura 2-18), alcanzando un valor máximo a las 6 a.m., donde existe una mayor concentración en entradas en los portales del sistema, y los destinos están principalmente sobre la línea Av. Caracas, en el centro y el centro extendido hasta cerca de la calle 127. Los viajes con motivo trabajo, representan el 27% de la demanda de viajes, con una duración promedio de 1,6 horas, la duración promedio de la actividad está en 8,8 horas, con una moda de 11 horas entre viajes, lo que concuerda con las horas laborales de la ciudad.

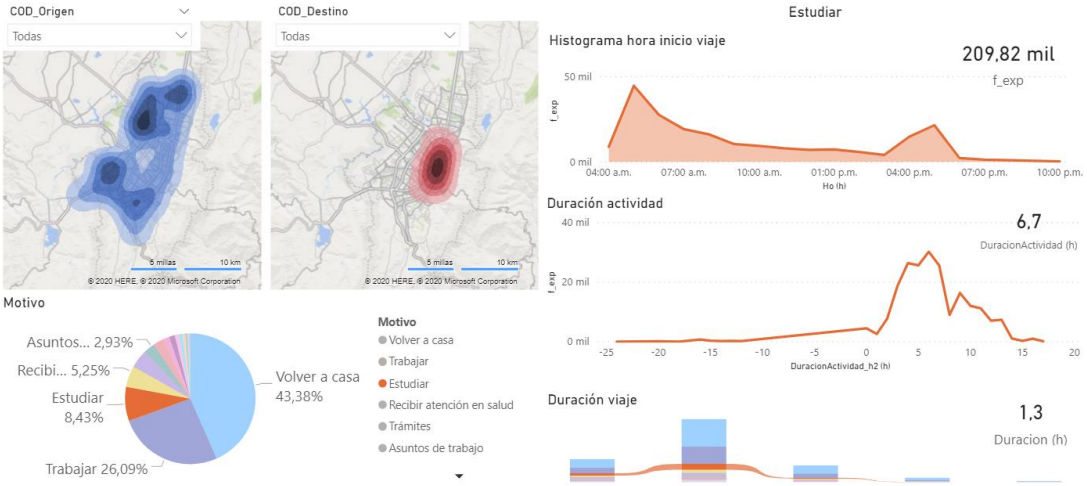
Figura 2-18 Isócronas, histogramas por motivo viaje – Trabajar



Fuente: elaboración propia a partir de Geocodificación de estación en la que abordó el vehículo, con información EODH

Para el caso del motivo de viaje “estudiar”, la Figura 2-19 muestra dos picos, uno en la mañana a las 5 a.m. y uno de menor concentración en la tarde a las 5 p.m., con un comportamiento más similar al presentado para el caso del motivo trabajo en cuanto el comportamiento de los orígenes de viajes que se concentran en los portales; no obstante, la concentración de los destinos de viajes es diferente, ya que se concentran de forma más puntual, en las estaciones como Universidad Nacional, Las Aguas, Calle 45, entre otras. El motivo estudiar representa el 8,3% de la demanda diaria de viajes que usa TransMilenio, con una duración de actividad promedio de 6,7 horas y moda de 6 horas, y una duración del viaje es de 1,3 horas, en ambos casos menor a los motivos de viaje volver a casa y trabajar, reducciones que pueden estar asociada a la flexibilidad de los estudiantes a jornadas más cortas y a la posibilidad de localizarse alrededor de su lugar de estudio.

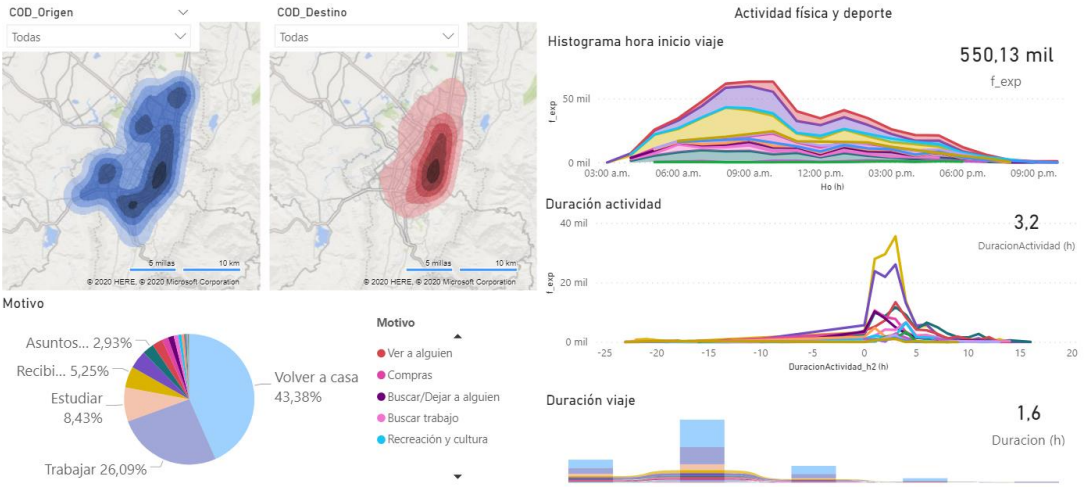
Figura 2-19 Isócronas, histogramas por motivo viaje – Estudiar



Fuente: elaboración propia a partir de Geocodificación de estación en la que abordó el vehículo, con información EODH

En el caso del resto de motivos, que representan menos de una cuarta parte de la demanda, el comportamiento de los viajes varía, ya que no se evidencian concentraciones puntuales con centralidades o con portales (ubicación residencial), para el caso de los orígenes. A su vez, se observa que la duración de la actividad se reduce considerablemente a 3,2 horas menos de la mitad del motivo estudiar, con una duración media del viaje de 1,6 horas.

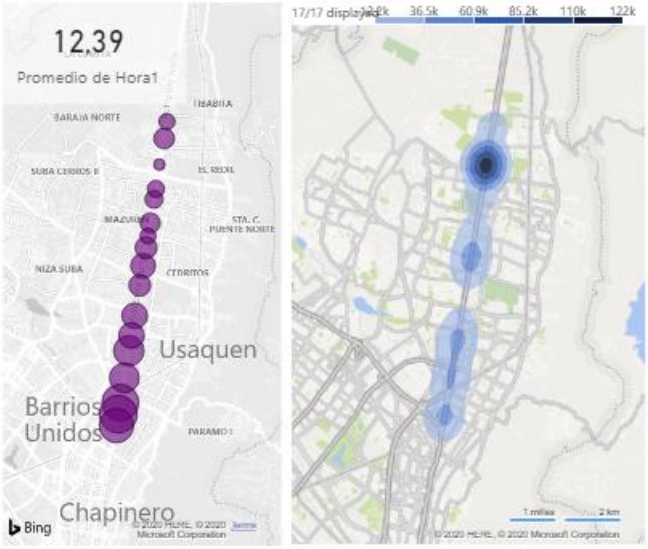
Figura 2-20 Isócronas, histogramas por motivo viaje – Otros



Fuente: elaboración propia a partir de Geocodificación de estación en la que abordó el vehículo, con información EODH

Para cada una de las líneas se tiene la información de la distribución espacial y la hora promedio de inicio del viaje, como se presentan en la Figura 2-21, para la línea de la Auto Norte, con un valor promedio de hora de entrada a las 12,39.

Figura 2-21 Hora promedio de ingreso en Zona B Auto Norte



Fuente: elaboración propia.

2.4 Uso del suelo alrededor de estaciones y portales de TransMilenio

La información asociada al uso del suelo alrededor de estaciones y portales de TransMilenio, se obtiene mediante un área de influencia alrededor de cada una de las estaciones y portales, considerando lo obtenido de la EODH, donde se obtuvo que en promedio se caminan 5 cuadras y/o duran un tiempo de 9 minutos, considerando esto y una velocidad promedio de caminata de 1,2 m/s (Guío Burgos, 2009), se define un área de influencia alrededor de estaciones de 500 metros, como se presenta en el mapa de la Figura 2-22.

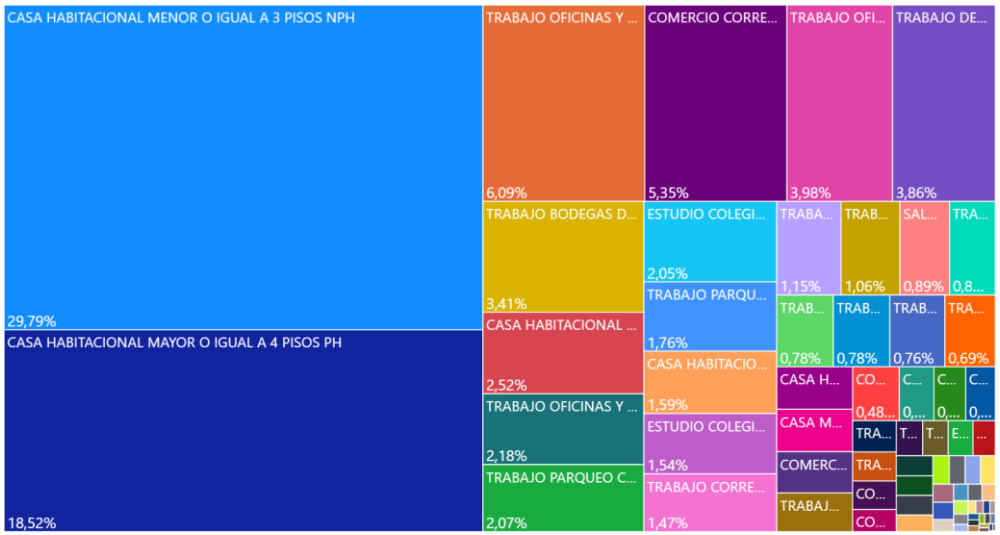
Figura 2-22 Área de influencia alrededor de estaciones y portales de TransMilenio

Fuente: elaboración propia.

Definida el área de influencia alrededor de estaciones y portales, se obtiene para cada una de ellas, el área asociada a los diferentes tipos de uso de suelo, tomando como referencia la información por lote, que permite conformar una base de 240 mil lotes alrededor de estaciones y portales de TransMilenio.

Con esta base, se obtiene la información del usos del suelo que se presenta en la Figura 2-23, en donde en conjunto para los porcentajes de distribución en el área de influencia de 500 metros, el 29,8% corresponde a uso habitacional menor o igual a 3 pisos, 18,5% a casa habitacional mayor o igual a 4 pisos, el 6,1% a trabajo oficina y consultorios, y el 2% a estudio colegios y universidades, así la distribución para las 100 tipologías de uso de suelo.

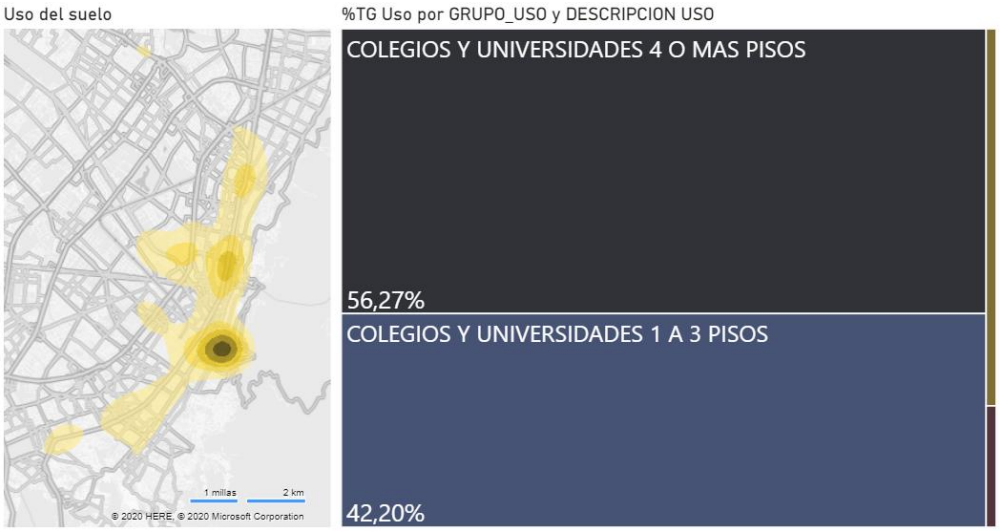
Figura 2-23 Distribución de usos alrededor de estaciones y portales de TransMilenio



Fuente: elaboración propia.

A partir de esta información, se pueden realizar consultas de forma individual, ya sea asociadas a una estación específica o a un grupo de usos de suelo, como se presenta en la Figura 2-24, donde se tienen los usos de suelo de Colegios y universidades con 4 o más pisos, y Colegios y universidades 1 a 3 pisos, que muestran una concentración en estaciones como la Universidad Nacional, Las Aguas, Calle 45 entre otras.

Figura 2-24 Uso de suelo asociado a educación



Fuente: elaboración propia.

2.5 Discusión y conclusiones

El procesamiento y la minería de datos, presentada en este capítulo, han generado hallazgos interesantes sobre la información de viajes de TransMilenio, por ejemplo, cuando se aplica la prueba de hipótesis de diferencia de medias entre tipos días, se observa que la menor diferencia se presenta entre un día domingo y festivo, y en este sentido se podría pensar reducir el nivel de confianza para encontrar que no se rechaza la hipótesis nula para los días domingo y festivo, permitiendo su agrupación en un mismo tipo día.

Por otra parte, al aplicar la prueba de diferencia de media entre las diferentes estacionalidades del año, refleja que los días de semana santa y enero, presentan comportamientos similares, que sugieren su agrupación.

Adicionalmente, cuando se analiza el comportamiento espacial de la distribución de las entradas sobre las líneas TransMilenio, se observa que líneas como la NQS Sur y la Av. Américas, presentan un mayor número de entradas en hora de la mañana, lo que puede estar asociado principalmente a la localización de viviendas alrededor de las estaciones que conforman estas líneas. Por otra parte, concentraciones como las que se presentan sobre la línea de la Av. Caracas, donde las entradas se dan principalmente en horas de la tarde, se pueden asociar a que el uso del suelo alrededor de estas estaciones es principalmente de tipo dotacional.

A su vez, el hecho de que la hora promedio de entrada a las estaciones sea mayor a medida que se van acercando al centro, es un indicio de que existe una correlación entre la ubicación de las estaciones y portales, y su dinámica temporal de entrada y salidas; a pesar de que este comportamiento, no se mantiene para los sábados y domingos.

El hecho de haber encontrado que no existen variaciones significativas entre las velocidades para diferentes franjas horarias se asocia a que TransMilenio es un sistema de transporte público tipo BRT, con carril exclusivo, corroborando que sus cambios están dados principalmente por las intersecciones semaforizadas y por las paradas en estaciones, las cuales demuestran ser constantes a lo largo del recorrido, dado que el tiempo total de viaje no varía significativamente.

A partir de los datos de programación y ejecución de TransMilenio, se obtiene que solamente el 30% de los pares origen destino, pueden ser atendidos sin transbordos dentro del sistema. Por otra parte, existe un buen ajuste entre la duración del viaje ejecutado, frente a la modelación de redes.

La simetría que existe entre los mapas de isócronas del origen y el destino que se presentaron en la Figura 2-16, son consistentes con la simetría que presentan las matrices de viajes para un día, ya que se espera que los viajes que se realizan en la mañana tengan su par en horas de la tarde. A su vez, el histograma que muestra la distribución de la demanda a lo largo del día presenta los mismos picos que las entradas del uso de tarjetas inteligentes, como para las obtenidas de la EODH, a las 6 a.m. y a las 5 p.m.

Cuando se realiza el ejercicio de comparar la hora de inicio promedio en las estaciones, los comportamientos son similares, tanto para los resultados obtenidos de los usos de tarjetas inteligentes, como los obtenidos de la EODH, donde la hora promedio de viaje de entrada es mayor a medida que se acerca al centro de la ciudad.

El uso de suelo alrededor de estaciones, refleja que puede existir una relación con los motivos del viaje, considerando que para el caso de las concentraciones de viajes con motivo estudiar de la Figura 2-19, la concentración de destinos en las estaciones de Universidad, Las Aguas y Calle 45, se repite para los usos de suelo asociados al estudio de la Figura 2-24.

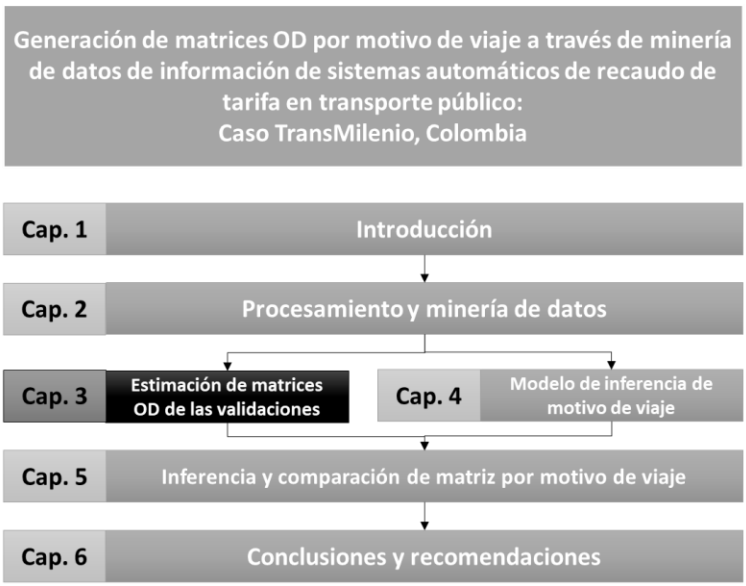
Finalmente, se puede concluir que el procesamiento y minería de datos, permite evidenciar que los datos de diferentes fuentes de información mantienen una consistencia, y que el comportamiento espacial y temporal, se puede asociar a diferencias en los motivos del viaje.

3. Estimación de matrices OD a partir de datos de validación de entradas al sistema

3.1 Introducción

El método de encadenamiento de viajes y el uso de las heurísticas es el método seleccionado para la estimación de las matrices de viajes del componente troncal del SITP (TransMilenio), a partir de los datos de uso de tarjetas inteligentes.

Figura 3-1 Metodología general – Capítulo 3



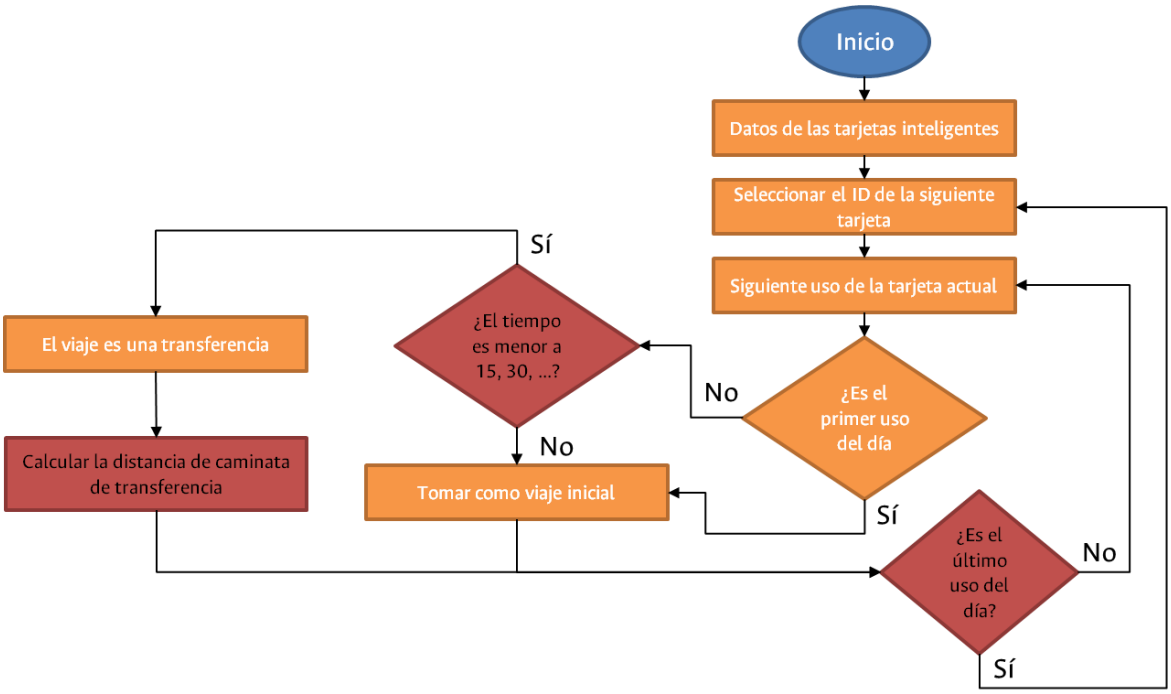
Fuente: elaboración propia.

En el caso particular del componente troncal del SITP, el método de encadenamiento de viajes busca la obtención del destino del viaje, para cada uno de los usos de las tarjetas, mediante la aplicación de diferentes heurísticas y supuestos. Las heurísticas empleadas se asocian principalmente a la frecuencia y ubicación del uso de las tarjetas.

Los supuestos del encadenamiento, consideran las reglas recogidas en la investigación de (Zhao, Rahbee and Wilson, 2007), y aplicadas al uso de tarjetas de TransMilenio, en donde el primero supone que entre dos viajes que se registran en TransMilenio, no existe un medio intermedio, lo que permite realizar una secuencia de encadenamiento, el segundo supuesto está asociado a que las transferencias se realizan dentro del sistema y consisten en transferencias entre rutas troncales, y el tercer supuesto considera un cierre en la cadena de viaje lo que indica que el destino del último viaje, es el comienzo del primero.

La Figura 3-2, contiene el flujograma del método de encadenamiento de viajes para la obtención de las matrices OD, el cual es una adaptación del procedimiento usado por (Alsger, 2017), en donde para cada una de las tarjetas diferenciadas con un ID de identificación, se construye la secuencia del viaje, a partir del orden que es definido por la hora de uso.

Figura 3-2 Método de encadenamiento del viaje



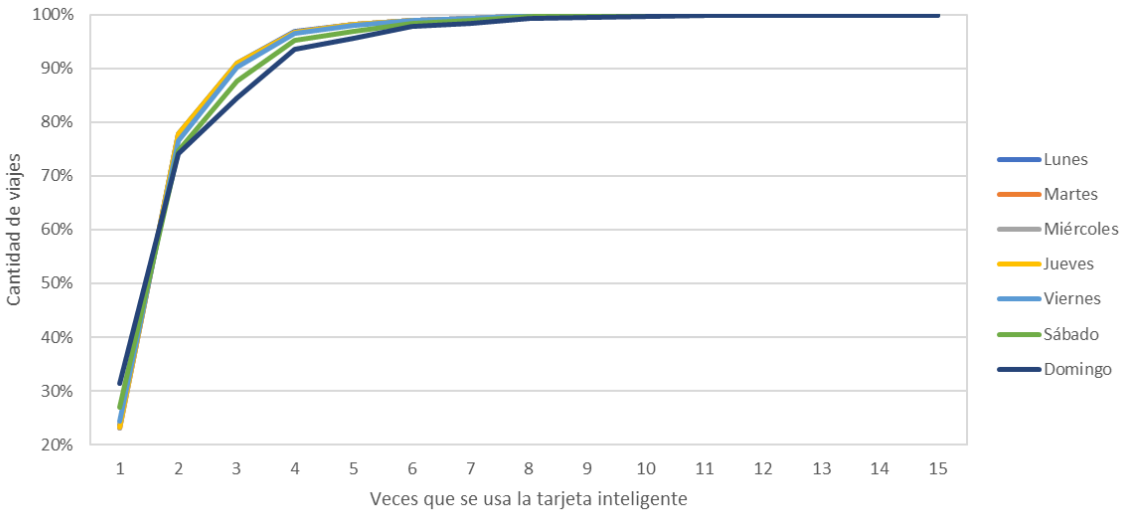
Fuente: Adaptado de (Alsger, 2017)

3.2 Heurísticas del modelo de encadenamiento

La obtención de las matrices de viaje está limitada por la forma en que son usadas las tarjetas inteligentes, ya que un solo uso de la tarjeta en un día impide realizar el encadenamiento del viaje, a su vez, muchos usos de una misma tarjeta pueden estar asociadas a otros usos operativos, que no obedecen a un patrón de viaje de una sola persona.

Analizando la frecuencia acumulada del uso de tarjetas para la semana de análisis, se observa que a pesar de que hay tarjetas que se usan muchas veces durante el día, el 99% de las tarjetas tienen una frecuencia de uso menor a 8 veces, cómo lo muestra la Figura 3-3.

Figura 3-3: Frecuencia acumulada uso de tarjetas



Fuente: elaboración propia.

Cuando el uso de las tarjetas se da en una misma estación, es decir que el origen y el destino quedarían asociados a una misma parada, no es posible realizar el encadenamiento del viaje.

A partir de estas tres heurísticas, se obtienen los resultados que se presentan en la Tabla 3-1, donde se reporta la afectación secuencial en el número de viajes que resulta al aplicar cada una de las reglas, obteniendo así el número de viajes que pueden ser encadenados.

Para los días hábiles el promedio del porcentaje de viajes que pueden ser encadenados es del 67,6%, con una variación del 1% para el lunes y un 2% para el viernes. Para el sábado tan solo un poco más de la mitad de los viajes pueden ser encadenados y el domingo menos de la mitad.

Tabla 3-1: Heurísticas y viajes encadenados por día de la semana

Heurísticas	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Total de datos	2.285.642	2.368.544	2.354.998	2.345.715	2.359.138	1.548.620	692.185
Más de una validación	1.746.251	1.821.936	1.810.582	1.798.351	1.783.486	1.132.478	475.524
Menos de 8 validaciones	1.729.421	1.804.415	1.792.679	1.779.694	1.765.173	1.114.746	464.463
OD diferentes	1.537.886	1.615.571	1.607.005	1.596.308	1.557.930	899.596	324.508
Porcentaje matriz final	**67,3%**	**68,2%**	**68,2%**	**68,1%**	**66,0%**	**58,1%**	**46,9%**

Fuente: elaboración propia.

3.3 Obtención de matrices

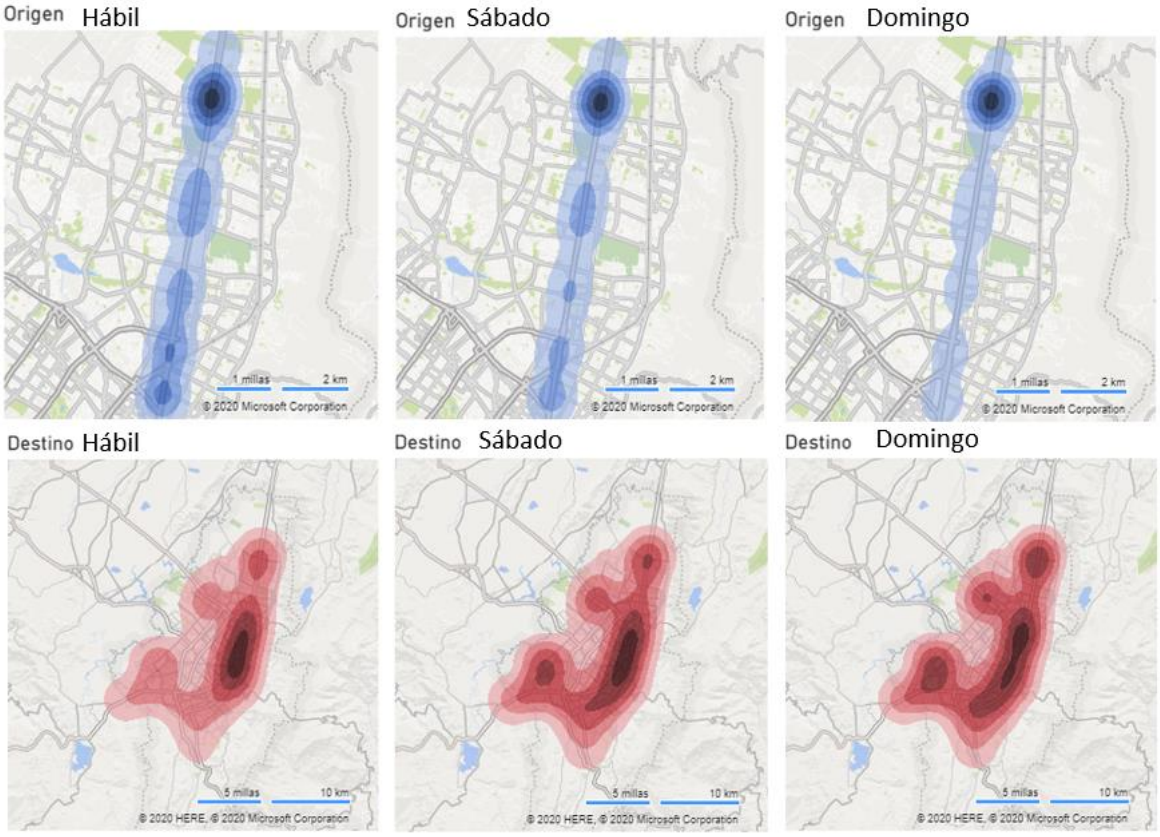
Aplicado el encadenamiento de viajes descrito en la sección anterior, para la semana del 1 al 7 de octubre de 2018, se obtiene para cada uno de los registros de validación la hora de ingreso al sistema, la estación de ingreso al sistema (origen) y la estación de salida del sistema (destino), manteniendo para cada uno las características del usuario que realiza el viaje.

La aplicación de las heurísticas sobre la base de datos con 13’954.842 registros, se realizó con el software SQL Server, en donde cada uno de los días y cada una de las tarjetas se ordenan según la hora del viaje, se enumeran con un código de secuencia de viajes para el viaje actual (n) y el siguiente viaje (n+1), al unir estos dos códigos se obtiene el encadenamiento de viajes; posterior a esto se aplican las heurísticas definidas, mediante la aplicación de filtros sobre el total de la base de datos.

La Figura 3-4 muestra la información geográfica de los viajes encadenados para la línea de la Autopista Norte, donde se observa que la mayor cantidad de entradas (origen) se presenta en el Portal Norte y la distribución de las salidas (destino), se da principalmente sobre las líneas de la Av. Caracas y en los Portales de la Calle 80, las Américas y Portal

del Sur, para los diferentes días de la semana. Las figuras para las demás líneas troncales se presentan en el Anexo B.

Figura 3-4: Localización de orígenes Troncal Norte y destinos de viajes resto del sistema



Fuente: elaboración propia.

3.4 Comparación de matrices inferidas y EODH

Con el objetivo de analizar la racionalidad de los resultados obtenidos por el método de encadenamiento de viajes, que permitió la obtención de las matrices de viaje a nivel de estaciones para el componente troncal del SITP, se presenta en esta sección la comparación de los resultados con aquellos obtenidos a partir de la encuesta origen destino en hogares, EODH para el modo TransMilenio.

Para el caso de la MOD a nivel de estaciones, estimadas a partir de la EODH, fue necesario realizar un proceso de geo codificación de estaciones de origen y destino, de 15.015 registros de viajes. Posteriormente, se obtuvo una matriz que representa 2,2 millones de viajes. Cabe aclarar que se mantuvieron los factores de expansión de la EODH y que no

fue posible realizar la codificación del total de la base, debido a la ausencia de una descripción adecuada que permitiera asociar el origen o destino a una estación. El resumen de este ejercicio se presenta en la Tabla 3-2.

Tabla 3-2 Cantidad de viajes y registros modo TransMilenio EODH

Atributos	F. expansión	Registros
Base total	2’489.738	16.589
Pares completos	2’238.520	15.015
Diferencia	-251.218	-1.574
% diferencia	-10,09%	-9,49%

Fuente: elaboración propia.

3.4.1 Comparación OD total de información

Como primera etapa en el proceso de comparación de resultado de las matrices OD, se calcula el total de ingresos en cada una de las estaciones, tanto para la matriz estimada a partir de la EODH, como para la matriz estimada por el método de encadenamiento de viajes; en la Tabla 3-3, se presentan estadísticos descriptivos asociados a los orígenes y destinos que ingresan por una misma estación.

Tabla 3-3 Medidas de tendencia total para ingresos y salidas totales por estación

	Origen		Destino	
Parámetro	**Mat. EODH**	**Mat. encadenamiento**	**Mat. EODH**	**Mat. encadenamiento**
$\overline{X}$	14.549	14.946	14.659,7	14.946,1
S	17.298.8	16,781.6	17,332.2	16,781.6
n	156	156	156	156
Viajes	2,269,648.1	2,331,593.0	2,286,907.6	2,331,593.0

Fuente: elaboración propia.

Para corroborar si existe similitud entre los ingresos y egresos de las estaciones, obtenidos por los dos métodos, EODH y encadenamiento de viajes, se construye un diagrama de dispersión con su recta de regresión lineal y se determina el coeficiente de correlación de R^2, usado para comparar la variabilidad del error estimado con la variabilidad de los valores originales (Gujarati and Porter, 2010), en donde:

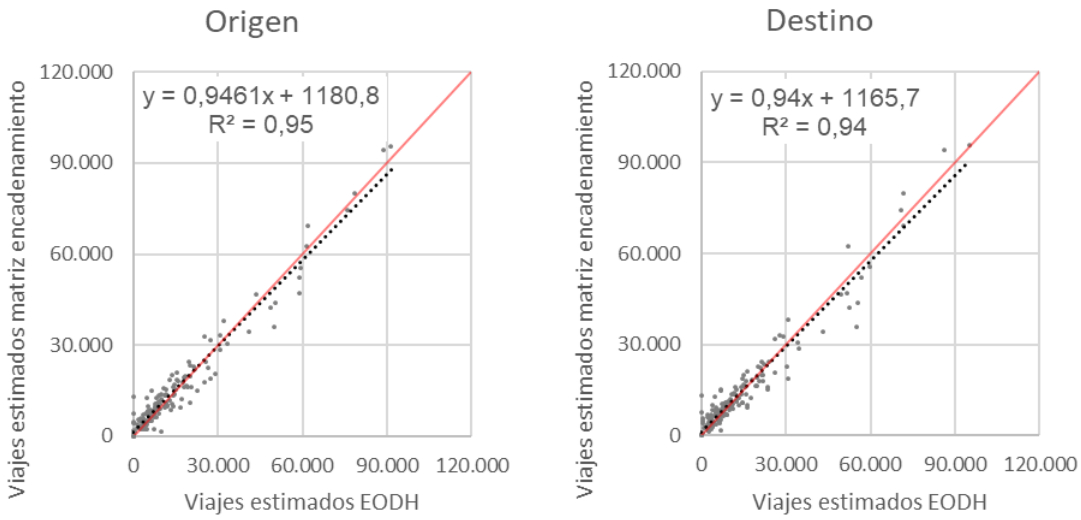
$$R^2 = 1 - \frac{SSe}{SSt}$$

$$SSe = \sum (yi - \bar{y}_i)^2$$

$$SSt = \sum (yi - \bar{y})^2$$

Los resultados obtenidos se presentan de forma gráfica en la Figura 3-5, donde se observa una buena correlación entre los ingresos y salidas construido a partir de la información de etapas de la EODH, con respecto a la información de la matriz estimadas por el método de encadenamiento de viajes.

Figura 3-5 Recta de correlación total origen y destino totales por estación



Fuente: elaboración propia.

Para probar si existe una diferencia estadística en la media de las estimaciones realizadas por los dos métodos, se realizó una prueba de hipótesis de medias, cuyos resultados se presentan junto con los de la regresión lineal en la Tabla 3-4. Para un nivel de confianza del 90%, se contrasta con un valor de $Z_{\alpha/2} = 1,96$ lo que implica que para los dos casos se acepta la hipótesis nula, considerando que Z* no es muy mayor Z, lo que indica que las medias de los valores de viajes obtenidos por los dos métodos no son estadísticamente diferentes. Incluso, las estimaciones de los destinos por los dos métodos son estadísticamente iguales al 95% de confianza.

Tabla 3-4 Comparación total de origen y destino por estación

EODH/TM	Origen	Destino
Z*	-0,20	-0,15
R²	0,95	0,94
Pendiente	0,95	0,94
Intercepto	1.180,82	1.165,70

Fuente: elaboración propia.

A su vez, se reporta un valor de correlación R^2 muy cercano a uno, con una pendiente cercana a 1. Sin embargo, el intercepto debería estar cercano de cero. Se considera que el valor es aceptable, teniendo en cuenta que la media de los viajes estimados por estación es superior a los 14 mil 500 ingresos o salidas del sistema.

3.4.2 Comparación OD en estaciones sin valores similares

Considerando que la aplicación del método de estimación de matrices por encadenamiento de viajes reduce los pares OD encadenados, en este numeral se comparan los valores de viajes estimados que pudieron ser encadenados de acuerdo con el procedimiento descrito en la sección 3.2 Heurísticas del modelo de encadenamiento. Se aclara que en el numeral anterior la matriz OD incluyó todos los pares origen destino. Como se muestra en la Tabla 3-5, existe una reducción significativa tanto en el número de viajes en los orígenes y destinos para el caso de la información tomada de las matrices estimadas por el método de encadenamiento.

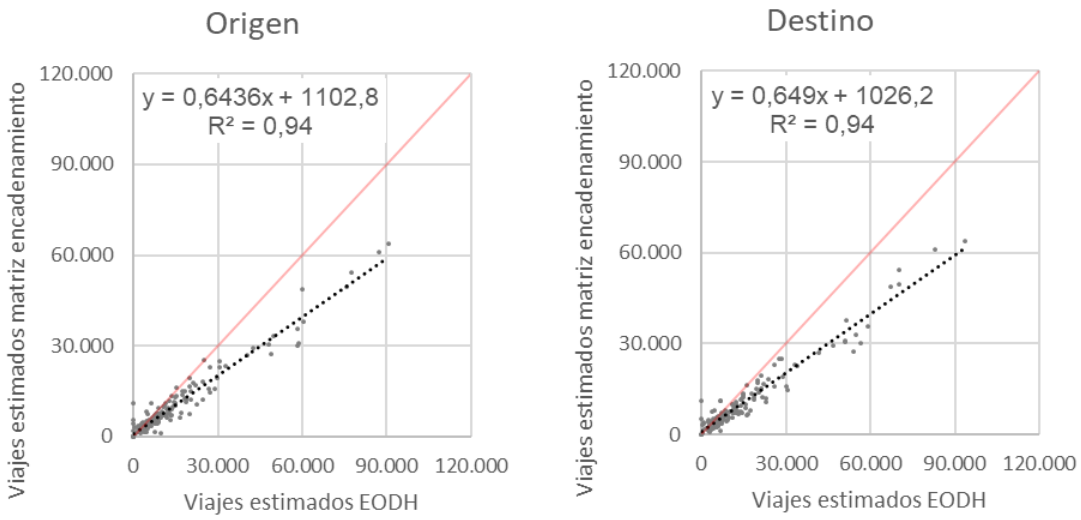
Tabla 3-5 Medidas de tendencia total para origen y destino sin duplicado en validaciones

	Origen		Destino	
Parámetro	**Mat. EODH**	**Mat. Encadenamiento**	**Mat. EODH**	**Mat. Encadenamiento**
$\overline{X}$	14.349,5	10.338,3	14.349,5	10.338,3
S	17.060,1	11.306,3	16.928,4	11.306,3
n	154	154	154	154
Viajes	2.238.520	1.612.778	2.238.520	1.612.778

Fuente: elaboración propia.

Al realizar el cálculo de la recta de regresión entre los orígenes y destino, existe una tendencia lineal, como se presenta en la Figura 3-6.

Figura 3-6 Recta de correlación total origen y destino sin duplicados en validaciones



Fuente: elaboración propia.

De igual forma, que en el proceso de comparación de la sección 3.4.1, se aplica la prueba de hipótesis de diferencia de medias, donde se encuentra un valor de Z*, es superior a 1,96 por lo que se rechaza la hipótesis nula. Lo anterior implica que existe diferencia entre las medias de los valores de la matriz estimada por el método de encadenamiento de viajes y los valores estimado de la EODH al incluir únicamente los pares OD de la matriz de encadenamiento sin duplicados en validaciones ; por otra parte, existen valores altos de correlación de 0,94, esto a pesar de que el intercepto se está alejando del valor de 1, como se presenta en la Tabla 3-6.

Tabla 3-6 Comparación de origen y destino por estación sin duplicados en validaciones

Mat. EODH / Mat. encadenamiento	Origen	Destino
Z*	2,43	2,45
R²	0,94	0,94
Pendiente	0,64	0,65
Intercepto	1.102,79	1.026,20

Fuente: elaboración propia.

3.4.3 Comparación OD a nivel de líneas del sistema

Considerando que existe una clara agrupación de estaciones asociadas a las líneas del sistema, se realiza una comparación entre las matrices OD no a nivel de estaciones, sino agrupado en líneas[5], incluyendo solamente aquellos pares origen destino (OD) que han podido ser reconstruidos en su totalidad, tanto para la información de las matrices estimadas a partir de la EODH, como para las matrices estimadas por el método de encadenamiento de viajes.

En la Tabla 3-7, se presentan las medidas de tendencia central obtenidas de las matrices a nivel de líneas, una matriz de 144 pares para las 12 líneas del sistema.

Tabla 3-7 Medidas de tendencia central para matrices a nivel de líneas del sistema

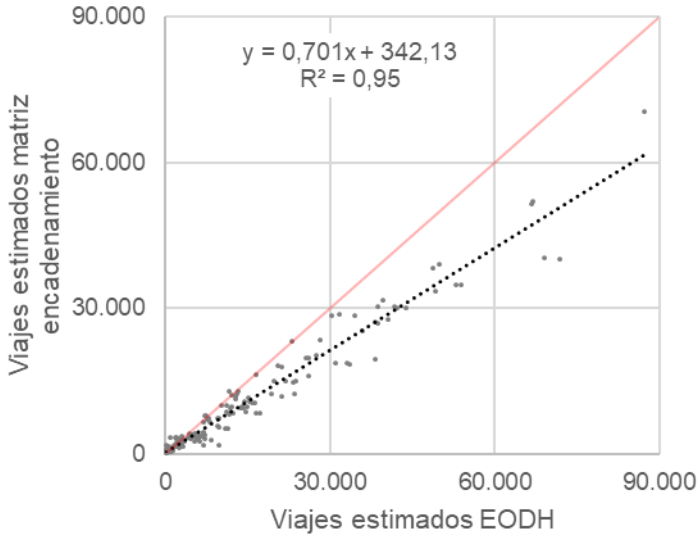
Parámetro	Mat. EODH	Mat. Encadenamiento
$\overline{X}$	15.499,5	11.199,8
S	17.141,0	12.293,0
n	144	144
Viajes	2.231.933	1.612.778

Fuente: elaboración propia.

Al construir la recta de regresión lineal para evaluar la similaridad de las matrices estimadas a partir de la EODH, con respecto a las matrices estimadas por el método de encadenamiento de viajes, se observa en la Figura 3-7 que existe una relación lineal.

[5] Las 12 líneas son: Carrera 10, Calle 13, Eje Ambiental, AutoNorte, Av.Suba, Calle 80, Av. Américas, Calle 26, NQS Sur, NQS Central, Caracas Sur y Caracas.

Figura 3-7 Recta de correlación para pares OD a nivel de líneas del sistema



Fuente: elaboración propia.

Para la comparación de los viajes estimados por los dos métodos a niveles de pares OD por líneas, se reporta un coeficiente de correlación 0,95, el cual es mayor al reportado a nivel de estaciones. La pendiente es de 0,7, mayor que en el caso de las estimaciones a nivel de estaciones y el intercepto es más pequeño. La pendiente de 0,7 sugiere que las estimaciones a partir de la EODH son menores a nivel de línea, comparada con la estimación de viajes por el método de encadenamiento. Como se observa en la Tabla 3-8, al evaluar la diferencia estadística de medias se encuentra que el valor calculado de Z* es mayor a $Z_{\alpha/2}$, lo que implica que se rechaza la hipótesis nula, es decir que existe diferencias significativas entre las medias de las estimaciones.

Tabla 3-8 Comparación de pares OD a nivel de líneas del sistema

Parámetro	Mat. EODH/ Mat. Encadenamiento
Z*	2,45
R²	0,95
Pendiente	0,70
Intercepto	342,13

Fuente: elaboración propia.

3.4.4 Comparación a nivel de par OD de estación

Considerando la cantidad de pares origen destino del análisis a nivel de estaciones que resulta de las codificaciones, se ha llegado a un total de 154 estaciones, esta combinación

(154*154) podría generar una matriz de más de 24 mil pares origen destino. A continuación, se presenta la comparación únicamente para aquellos pares que contienen información tanto del origen y del destino (valor mayor a cero) y en los que el origen y destino son diferentes.

En la Tabla 3-9, se presentan los estimadores de la media, desviación estándar, cantidad de registros y total de viajes asociados a cada una de las matrices. Se observa que, en este caso, el valor de media para cada par origen destino, difiere considerablemente entre ambas matrices, y que la variabilidad de las estimaciones por el método de encadenamiento de viajes es menor.

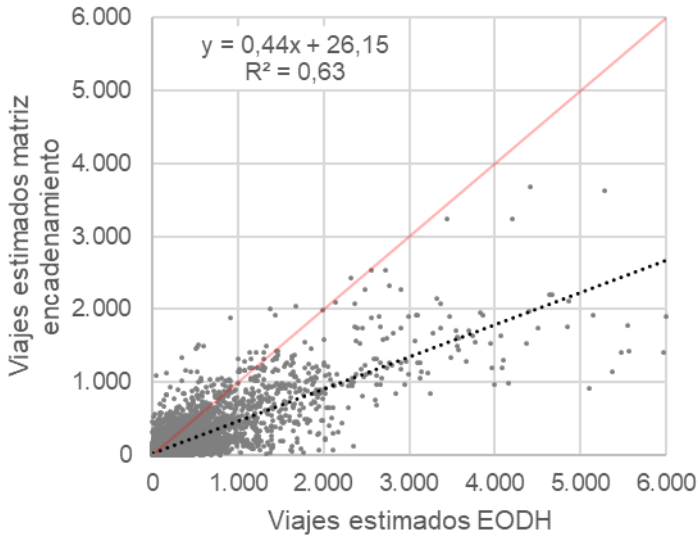
Tabla 3-9 Medidas de tendencia central para matrices OD a nivel de estación

Parámetro	**Mat. EODH**	**Mat. Encadenamiento**
$\overline{X}$	355,9	182,6
S	533,4	295,9
n	6.264	6.264
Viajes	2.231.933	1.612.778

Fuente: elaboración propia.

La Figura 3-8 contiene el diagrama de dispersión, donde se presenta la relación entre cada uno de los pares origen destino de la matriz estimada a partir de la EODH, y su comparación con los viajes OD de la matriz estimada por el método de encadenamiento.

Figura 3-8 Recta de correlación para pares OD a nivel de líneas del sistema



Fuente: elaboración propia.

En este caso, el coeficiente de correlación R^2 se ha reducido considerablemente con un valor de 0,63 y la pendiente está muy alejado de uno, a pesar de que el intercepto está más cercano a 1. Adicionalmente, en este caso la hipótesis nula que señala que las medias son iguales se debe rechazar, considerando que el valor del Z* estimado, es muy mayor al valor de $Z_{\alpha/2}$ al 95%, como se presenta en la Tabla 3-10.

Tabla 3-10 Comparación de pares OD a nivel de estación

Parámetro	Mat. EODH / Mat. encadenamiento
Z*	22,50
R²	0,63
Pendiente	0,44
Intercepto	26,15

Fuente: elaboración propia.

3.5 Discusión y conclusiones

Como ha sido presentado, con los datos del uso de tarjetas inteligentes fue posible estimar las matrices OD a nivel de estación por el método de encadenamiento de viajes, utilizando dos terceras partes de los registros de validación del sistema TransMilenio de un día hábil. Esto implicó la reconstrucción de cerca de 1,6 millones de pares origen destino, casi 100 veces más la cantidad de viajes que fueron encuestados en la EODH, donde se reportaron 16.589 viajes del medio principal TransMilenio, que con factores de expansión permiten

llegar a los 2,49 millones de viajes y de los cuales pudieron ser reconstruidos en su totalidad 2,2 millones para un total de 15 mil registros.

Por otra parte, se pudieron comparar los viajes que son estimados por el método de encadenamiento de viajes con aquellos estimados a partir de la EODH, a nivel de entradas y salidas en estaciones, y a nivel de pares origen y destino a nivel de líneas.

Se reconoce que esta comparación no permite estimar errores ni sesgos, pues los dos métodos son aproximaciones a los valores reales de viajes entre estaciones, cada uno con sus supuestos estadísticos de soporte. Sin embargo, se resalta que existen similitudes entre las estimaciones de viajes a nivel de línea, ya que se obtiene un valor de R^2 del 95%, con una pendiente cercana a 1 y un intercepto cercano a cero.

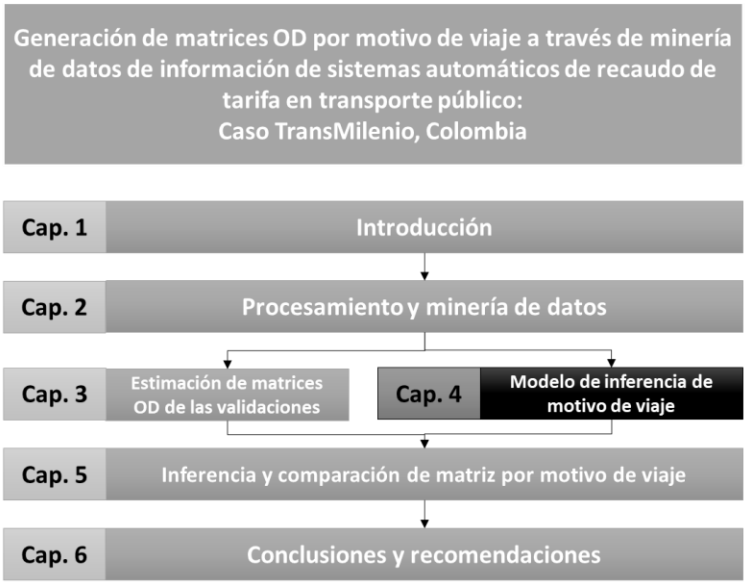
En cuanto a las diferencias entre las estimaciones de viajes de los orígenes y destinos a nivel de estación y/o portal, se concluye que estas diferencias pueden ser explicadas debido a que la información de la EODH obedece a resultados de un diseño muestral a nivel de Zona de Análisis de Transporte (ZAT), y no a nivel de estación y/o portal de TransMilenio, resultando en que el número de pares contenidos en la matriz estimada de la EODH sea inferior a la matriz estimada por el método de encadenamiento de viajes.

4. Modelo de inferencia de motivo de viaje

4.1 Introducción

A pesar de que el uso de las tarjetas inteligentes permite la obtención de información valiosa para la planeación y operación de sistemas de transporte como las matrices OD, hay una parte importante de la información que no se puede estimar directamente, y ese es el motivo del viaje. En este capítulo se presenta una exploración para la obtención de los motivos de viaje de las matrices obtenidas a partir de los datos de las tarjetas inteligentes, utilizando varias fuentes de información. El modelo de inferencia de motivo de viaje se construye mediante el uso de las variables obtenidas en el procesamiento y minería de datos (ver Figura 4-1), y a partir del uso de modelos logit multinomial, utilizando características temporales y espaciales que permiten hacer la inferencia en el sistema troncal del SITP (TransMilenio).

Figura 4-1 Metodología general – Capítulo 4



Fuente: elaboración propia.

La Figura 4-2, resume el método utilizado, donde se hace una exploración de las variables espaciales y temporales, con el fin de hacer reducción de la cantidad de variables y de las categorías incluidas cómo motivos del viaje. La Tabla 4-1, presenta las variables dependientes de tipo categórico que son tratadas como variables cuantitativa discretas asociados al motivo del viaje con 17 tipologías de viaje obtenidas de la EODH, 4 variables temporales independientes de tipo cuantitativo continuo obtenido de la EODH y 100 tipologías de uso del suelo como variables espaciales asociadas al origen y destino del viaje.

Tabla 4-1 Variables el procesamiento y minería de datos

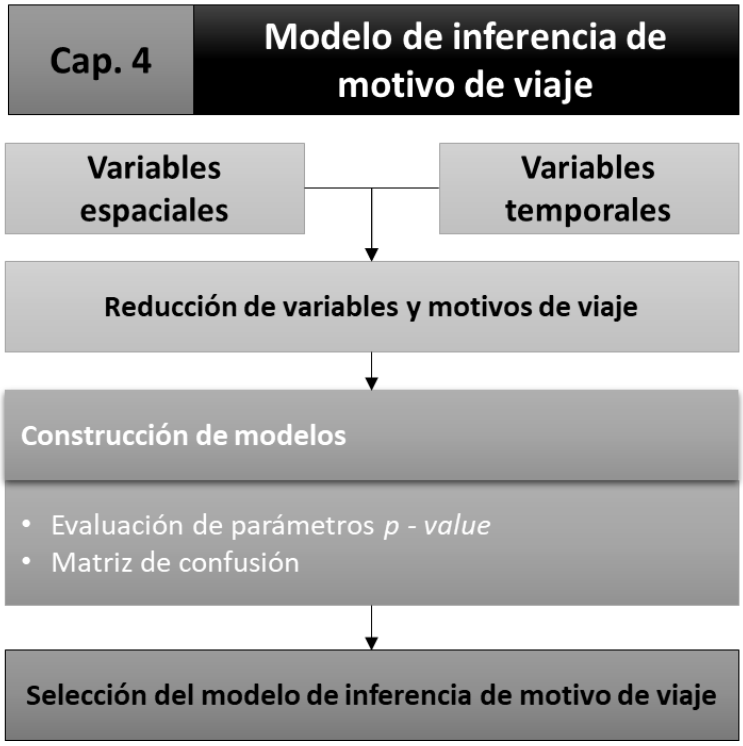
Variable	Grupo	Tipo de variable	Unidades	Variables	Fuente
Dependiente	Motivo del viaje	Cuantitativa discreta	Sin	Trabajar, Asuntos de trabajo, Estudiar, Recibir atención en salud, Ver a alguien, Volver a casa, Buscar/Dejar a alguien, Buscar/Dejar algo, Comer/Tomar algo, Compras, Trámites, Recreación y cultura, Buscar trabajo, Actividades con fines religiosos, Cuidado de personas, Actividad física y deporte, Otro.	Secretaría Distrital de Movilidad
Independiente	Temporal	Cuantitativa continua	Horas	Hora inicio del viaje, hora fin del viaje, duracción de la actividad, duracción del viaje.	Secretaría Distrital de Movilidad
	Espacial	Cuantitativa continua	Metros cuadrados (m2)	Habitacional Menor O Igual A 3 Pisos Nph, Habitacional Mayor O Igual A 4 Pisos Nph, Comercio Puntual Nph, Corredor Comercial Nph, Estaciones De Servicio, Centro Comercial Mediano Nph, Centro Comercial Grande Nph, Bodega Comercial Nph, Industria Artesanal, Industria Mediana, Industria Grande, Institucional Puntual, Colegios Y Universidades 1 A 3 Pisos, Iglesias, Oficinas Y Consultorios (Oficial) Nph, Colegios Y Universidades 4 O Mas Pisos, Clinicas, Hospitales, Centros Medicos, Instalaciones Militares, Industria Artesanal, Oficinas Y Consultorios Nph, Hoteles Nph, Depositos De Almacenamiento Nph, Teatros Y Cinemas Nph, Edificios De Parqueo Nph, Bodegas De Almacenamiento Nph, Moteles, Amoblados, Residencias Nph, Moteles, Amoblados, Residencias Ph, Industria Mediana Ph, Parques De Diversion, Clubes Mayor Extension,	Infraestructura de Datos Espaciales para el Distrito Capital

Variable	Grupo	Tipo de variable	Unidades	Variables	Fuente
				Piscinas En Nph, Coliseos, Bodega Economica, Industria Grande Ph, Colegios En Ph, Parques De Diversion En P.H., Habitacional Menor O Igual A 3 Pisos Ph, Habitacional Mayor O Igual A 4 Pisos Ph, Comercio Puntual Ph, Corredor Comercial Ph, Centro Comercial Mediano Ph, Centro Comercial Grande Ph, Centros Medicos En Ph, Institucional Ph, Oficinas Y Consultorios Ph, Hoteles Ph, Teatros Y Cinemas Ph, Parqueo Libre Ph, Parqueo Cubierto Ph, Edificios De Parqueo Ph, Deposito (Lockers) Ph, Piscinas En Ph, Iglesia Ph, Cementerios, Restaurantes Nph, Restaurantes Ph, Aulas De Clase, Clubes Pequenos, Plazas De Mercado, Museos, Enramadas, Cobertizos, Caneyes, Galpones, Gallineros, Establos, Pesebreras, Cocheras, Marraneras, Porquerizas, Beneficiaderos, Secaderos, Kioskos, Silos, Oficinas En Bodegas Y/O Industrias, Oficina Bodega Y/O Industria Ph, Oficinas Operativas(Estaciones Servicio, Lote En Propiedad Horizontal, Bodega Comercial Ph, Oficinas Y Consultorios (Oficial) Ph, Bodegas De Almacenamiento Ph, Centro Comercial Pequeno Nph, Centro Comercial Pequeno Ph, Parqueo Cubierto Nph, Bodega Economica(Serviteca,Esta.Servic., Deposito Almacenamiento Ph.	

Fuente: elaboración propia

Las variables espaciales que inicialmente se caracterizan con 100 tipos de uso de suelo son agrupadas en 8 variables, revisando la correlación entre variables a través de la construcción de una matriz de correlación de Pearson, estas variables son transformadas relacionando el uso de suelo del origen y el destino.

Figura 4-2 Método – Modelo de inferencia de motivo de viaje



Fuente: elaboración propia.

Sobre los 17 motivos de viaje contenidos en la EODH, se definen categorías agrupadas analizando el comportamiento de las cuatro variables temporales: duración de la actividad, hora inicio, hora fin y duración del viaje, mediante una prueba de hipótesis de diferencia de medias.

Una vez determinadas las variables espaciales y temporales y las categorías de motivos de viaje, se estiman los parámetros para los modelos logit multinomial, que son evaluados con el uso de la prueba de significancia estadística, a partir de la inspección del p-value, a un 95% de confianza. Se prueban más de 150 modelos, los cuales son evaluados mediante matrices de confusión. Los modelos se estiman usando el 70% de los datos y son validados con el 30% de los datos restantes.

4.2 Reducción de variables y motivos de viaje

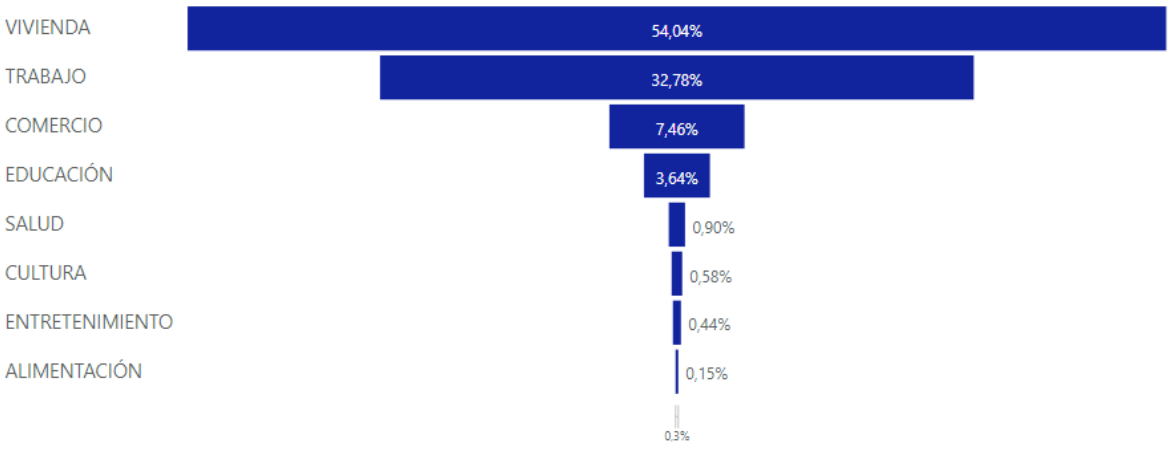
Como fue expuesto en la sección 1.2 Estado del arte, autores como (Alsger, 2017) han evidenciado la importancia de los atributos espaciales y temporales, para la inferencia de matrices de viaje por motivo. Para ello, se toman las variables obtenidas principalmente de la EODH y las características del uso del suelo. La selección de aquellas variables de interés para la inferencia de motivo de viaje, deben ser variables que puedan ser obtenidas, del uso de las tarjetas inteligentes e información adicional que permita complementar su respuesta (ver Tabla 4-1).

4.2.1 Análisis de variables de uso del suelo

La base de datos geográfica de usos del suelo de Bogotá, D.C. (Unidad Administrativa Especial de Catastro Distrital, 2019) define más de 100 tipos diferentes de usos. Estos fueron agrupados en 8 categorías según la descripción del uso del suelo.

Las 8 categorías son: vivienda, trabajo, comercio, educación, salud, cultura, entretenimiento, alimentación y otros. La participación de cada tipo de uso de suelo se presenta en la Figura 4-3. La mayor participación en porcentaje de uso de suelo está representada por vivienda con un 54%, seguido de trabajo con un 33%, comercio con un 7,5% y educación con un 3,6%.

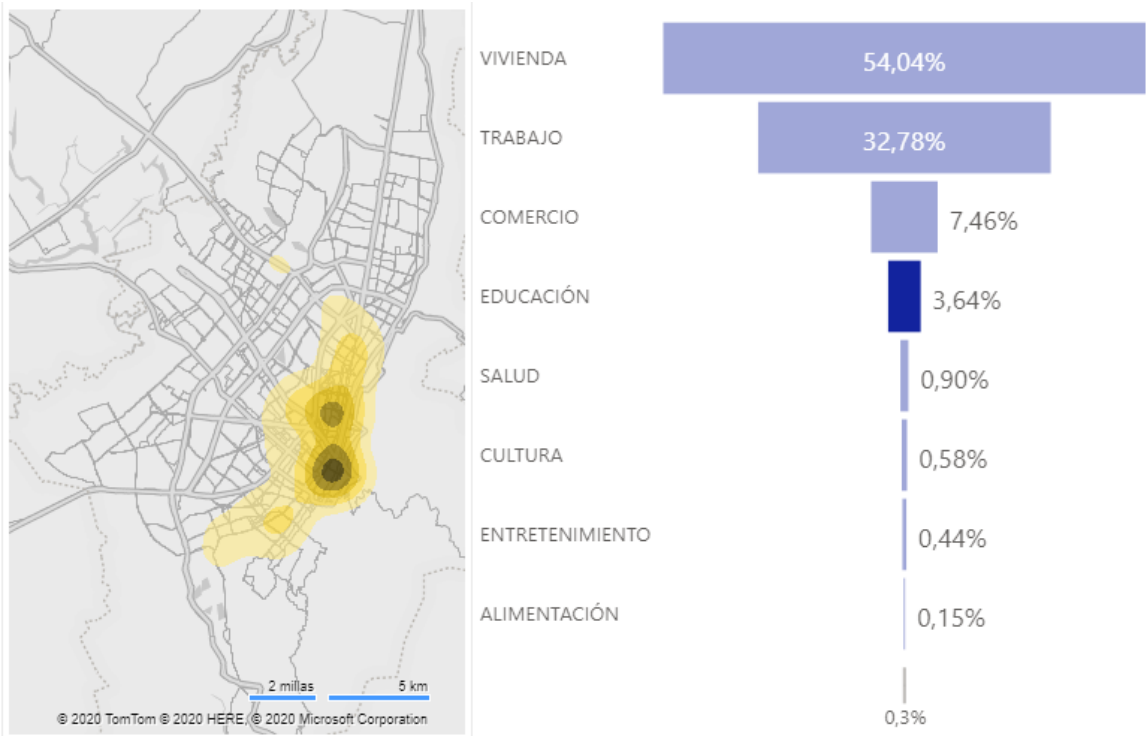
Figura 4-3 Grupos tipología uso de suelo



Fuente: Elaboración propia.

La Figura 4-4 muestra la distribución geográfica de las 8 categorías de usos del suelo. Para el caso del uso de suelo dedicado a educación, se observa que las áreas de universidades y colegios se concentran principalmente alrededor de las estaciones Universidad Nacional, Las Aguas, y Calle 45, entre otras.

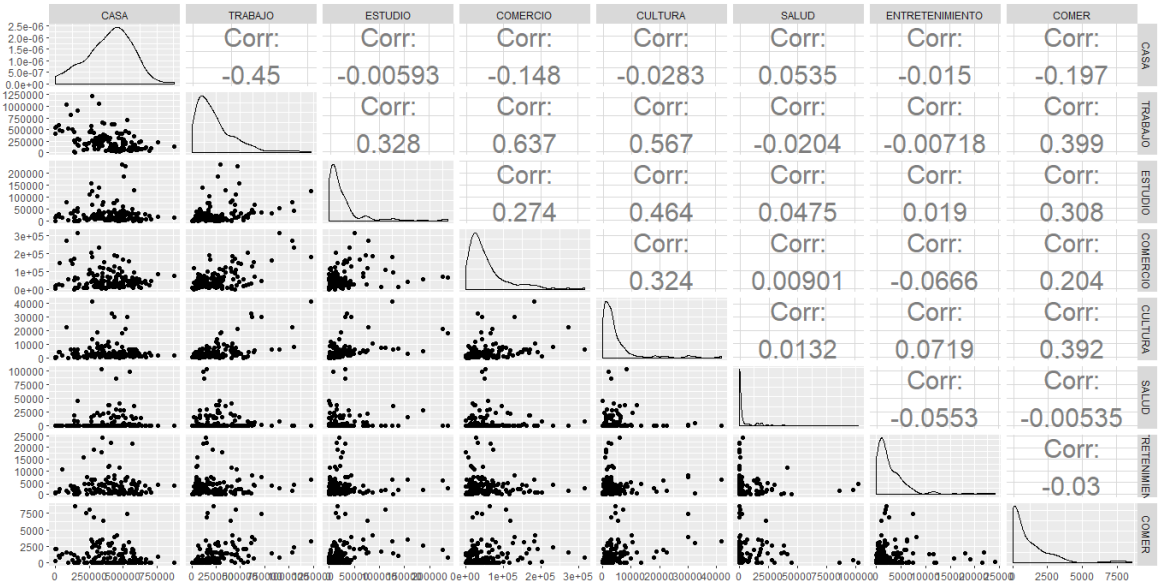
Figura 4-4 Concentración de área por grupo de uso de suelo estudio



Fuente: elaboración propia.

Es importante validar que no existe colinealidad entre las 8 variables de uso de suelo que representan las categorías agrupadas, ya que este es un supuesto para el uso de los modelos logit que se utilizarán para estimar el motivo del viaje. La Figura 4-5, muestra la matriz de correlación de Pearson. Se reporta que las correlaciones tienen valores inferiores a 0,6 absoluto, por lo cual su inclusión conjunta dentro del modelo logit es válida.

Figura 4-5 Matriz de correlación de Pearson, 8 grupos de uso de suelo



Fuente: elaboración propia.

4.2.2 Análisis de motivos de viaje

Es importante analizar las variables temporales asociadas al motivo de viaje: la hora de inicio del viaje, la duración del viaje y la duración de la actividad. Para validar si existe diferencia entre las medias de estas variables asociadas para cada motivo, se aplica una prueba de hipótesis de diferencia de medias.

La Figura 4-6 muestra la estadística descriptiva y valor de Z estimado para la variable tiempo de inicio de viaje de cada uno de los motivos. Adicionalmente, en la parte diagonal superior de la figura se presentan de forma matricial los resultados de la prueba de diferencia de medias entre todos los motivos de viaje, es decir el resultado del valor de Z el parámetro de la prueba de hipótesis de diferencia de medias. En la diagonal inferior de la figura se muestra a partir de ceros y unos, donde el valor de cero significa que las medias no son estadísticamente diferentes, para un valor menor o igual de 1,95*1,5, lo que implicaría que se acepta la hipótesis.

Figura 4-6 Estadística descriptiva y valor de Z estimado, hora de inicio por motivo del viaje

Id	Motivo viaje	$\bar{X}$	S	n	Volver a casa	Trabajar	Estudiar	Salud	Trámites	Buscar trabajo	Ver a alguien	Asuntos de trabajo	Recreación y cultura	Comer/Tomar algo	Compras	Buscar/Dejar alguien	Buscar/Dejar algo	Act. Religiosa	Act. física y deporte	Cuidado de personas	Otro
					6	1	3	4	11	13	5	2	12	9	10	7	8	14	16	15	77
6	Volver a casa	16.35	3.86	7102	0.00	130.51	53.22	50.88	47.65	33.47	17.10	37.60	6.89	3.05	19.65	14.03	13.03	5.57	6.91	9.22	3.86
1	Trabajar	7.66	3.15	4280	1	0.00	14.48	16.03	23.21	5.11	26.19	14.47	17.90	14.71	18.41	16.35	10.02	8.91	8.95	5.53	5.62
3	Estudiar	9.53	4.37	1335	1	1	0.00	1.31	6.02	2.86	14.83	2.76	12.00	10.63	9.23	9.17	4.87	5.70	5.43	2.32	3.56
4	Salud	9.75	3.52	844	1	1	0	0.00	4.63	3.74	13.80	1.64	11.38	10.18	8.32	8.42	4.29	5.33	5.03	1.94	3.32
11	Trámites	10.52	3.21	803	1	1	1	1	0.00	6.89	10.65	2.24	9.35	8.68	5.38	5.98	2.35	4.08	3.67	0.67	2.49
13	Buscar trabajo	8.81	2.94	178	1	1	0	1	1	0.00	13.97	4.66	12.46	11.31	9.73	9.85	6.05	6.58	6.37	3.36	4.25
5	Ver a alguien	12.92	4.12	447	1	1	1	1	1	1	0.00	11.26	2.54	3.69	3.56	1.70	3.51	0.13	0.62	3.24	0.12
2	Asuntos de trabajo	10.08	3.51	479	1	1	0	0	0	1	1	0.00	10.08	9.32	6.50	6.96	3.34	4.72	4.36	1.38	2.94
12	Recreación y cultura	13.93	4.08	138	1	1	1	1	1	1	0	1	0.00	1.54	5.01	3.57	4.89	1.35	2.12	4.42	1.15
9	Comer/Tomar algo	14.86	3.37	48	1	1	1	1	1	1	1	1	0	0.00	5.58	4.48	5.57	2.40	3.14	5.16	1.97
10	Compras	11.87	3.58	257	1	1	1	1	1	1	1	1	1	1	0.00	1.31	0.98	1.78	1.18	1.51	1.01
7	Buscar/Dejar alguien	12.34	4.40	244	1	1	1	1	1	1	0	1	1	1	0	0.00	1.92	1.01	0.37	2.18	0.50
8	Buscar/Dejar algo	11.44	3.77	102	1	1	1	1	0	1	1	1	1	1	0	0	0.00	2.22	1.70	0.75	1.39
14	Act. Religiosa	13.00	4.67	61	1	1	1	1	1	1	0	1	0	0	0	0	0	0.00	0.54	2.49	0.17
16	Act. física y deporte	12.56	4.40	65	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0.00	2.05	0.23
15	Cuidado de personas	10.92	3.98	46	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0.00	1.74
77	Otro	12.81	4.19	21	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0.00

Fuente: elaboración propia.

El recuadro en la parte inferior de la Figura 4-6, señala variables sobre las cuales se acepta la hipótesis nula de la prueba de diferencia de medias, lo que sugiere que el valor de inicio de viaje presenta un comportamiento muy similar para los motivos de viaje: compras, buscar/dejar alguien, buscar/dejar algo, actividad religiosa, actividad física y deporte, cuidado de personas y otro. Esto sugiere que la agrupación de estos motivos en un mismo conjunto sería adecuada.

Este mismo ejercicio se desarrolló para la variable temporal duración de la actividad. Los resultados se presentan en la Figura 4-7, que incluye de igual forma la información de las estadística descriptiva (media, desviación estándar y cantidad de datos) y valor de Z estimado; en este caso, se acepta la hipótesis nula, en donde un valor cero en la diagonal inferior de la figura, para una mayor cantidad de motivos de viaje, lo que indica que las media de los valores de viajes obtenidos, no son estadísticamente diferentes, marcados dentro del recuadro.

Figura 4-7 Estadísticas descriptivas y valor de Z estimado, duración actividad por motivo del viaje

Id	Motivo viaje	X̄	S	n	Volver a casa	Trabajar	Estudiar	Salud	Trámites	Buscar trabajo	Ver a alguien	Asuntos de trabajo	Recreación y cultura	Comer/Tomar algo	Compras	Buscar/Dejar alguien	Buscar/Dejar algo	Act. Religiosa	Act. física y deporte	Cuidado de personas	Otro
					6	1	3	4	11	13	5	2	12	9	10	7	8	14	16	15	77
6	Volver a casa	-7.68	5.45	7102	0.00	169.3	124.2	106.5	96.00	61.55	36.63	64.07	31.22	13.39	76.64	50.34	41.06	39.33	42.49	17.25	9.49
1	Trabajar	8.87	4.80	4280	1	0.00	17.52	49.92	46.22	23.96	21.95	22.97	14.29	8.91	43.65	29.37	24.50	18.44	19.59	7.27	6.98
3	Estudiar	6.75	3.52	1335	1	1	0.00	26.52	25.25	12.49	14.13	11.35	8.35	6.03	26.04	18.43	15.67	10.81	11.36	4.11	4.86
4	Salud	3.40	2.37	843	1	1	1	0.00	0.51	3.94	2.65	5.38	0.71	1.54	4.51	3.19	2.92	0.57	0.82	0.83	1.53
11	Trámites	3.33	2.68	803	1	1	1	0	0.00	4.14	2.39	5.54	0.88	1.45	3.86	2.82	2.62	0.78	1.04	0.92	1.47
13	Buscar trabajo	4.18	2.42	178	1	1	1	1	1	0.00	4.68	1.06	1.30	2.53	6.60	5.47	5.05	1.86	1.78	0.32	2.28
5	Ver a alguien	2.64	5.80	447	1	1	1	0	0	1	0.00	5.54	2.26	0.49	0.35	0.23	0.02	2.36	2.60	1.82	0.76
2	Asuntos de trabajo	4.45	3.89	479	1	1	1	1	1	0	1	0.00	1.99	2.89	7.95	6.54	5.98	2.69	2.66	0.71	2.55
12	Recreación y cultura	3.66	4.20	138	1	1	1	0	0	0	0	0	0.00	1.71	2.42	2.30	2.33	0.21	0.09	0.40	1.69
9	Comer/Tomar algo	2.26	5.12	48	1	1	1	0	0	0	0	0	0	0.00	0.65	0.61	0.51	1.66	1.74	1.71	0.32
10	Compras	2.75	1.92	257	1	1	1	1	1	1	0	1	0	0	0.00	0.11	0.35	2.70	3.07	1.78	0.88
7	Buscar/Dejar alguien	2.72	3.06	244	1	1	1	1	0	1	0	1	0	0	0	0.00	0.22	2.48	2.78	1.77	0.85
8	Buscar/Dejar algo	2.65	2.45	102	1	1	1	0	0	1	0	1	0	0	0	0	0.00	2.47	2.73	1.83	0.77
14	Act. Religiosa	3.56	2.17	61	1	1	1	0	0	0	0	0	0	0	0	0	0	0.00	0.15	0.54	1.64
16	Act. física y deporte	3.62	2.08	65	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0.00	0.47	1.70
15	Cuidado de personas	3.96	4.55	46	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0.00	1.74
77	Otro	1.86	4.59	21	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00

Fuente: elaboración propia.

En el caso de la duración del viaje, que se presenta en la Figura 4-8, para la mayor cantidad de motivos de viaje, se acepta la hipótesis nula, donde se rechaza únicamente para los motivos de buscar trabajo con una duración media de viaje de 2 horas y comer o tomar algo con una duración media de 1,3 horas.

Figura 4-8 Estadísticas descriptivas y valor de Z estimado, duración por motivo del viaje

Id	Motivo viaje	X̄	S	n	Volver a casa	Trabajar	Estudiar	Salud	Trámites	Compras	Asuntos de trabajo	Buscar trabajo	Buscar/Dejar alguien	Buscar/Dejar algo	Ver a alguien	Act. física y deporte	Act. Religiosa	Recreación y cultura	Comer/Tomar algo	Cuidado de personas	Otro
					6	1	3	4	11	10	2	13	7	8	5	16	14	12	9	15	77
6	Volver a casa	1.55	1.84	7102	0.00	0.3	6.8	2.6	1.06	2.70	0.09	5.27	1.41	0.51	0.39	0.72	0.10	1.02	3.00	0.62	0.10
1	Trabajar	1.56	1.03	4280	0	0.00	8.19	2.56	0.95	2.88	0.23	5.25	1.61	0.45	0.49	0.79	0.05	1.13	3.12	0.68	0.14
3	Estudiar	1.35	0.74	1335	1	1	0.00	7.06	4.89	0.14	3.48	7.45	2.17	2.34	1.85	0.88	1.53	1.32	0.91	0.88	1.09
4	Salud	1.67	1.19	844	0	0	1	0.00	0.99	3.95	1.87	3.74	3.01	0.54	1.61	1.61	0.72	2.23	3.99	1.46	0.78
11	Trámites	1.61	1.39	803	0	0	1	0	0.00	2.98	0.86	4.22	1.91	0.00	0.92	1.11	0.29	1.49	3.27	0.99	0.42
10	Compras	2.05	1.22	178	1	1	1	1	1	6.04	4.73	0.00	5.46	3.08	4.19	3.79	2.87	4.70	5.98	3.57	2.66
2	Asuntos de trabajo	1.52	1.85	447	0	0	0	0	0	0	0.30	4.19	0.43	0.66	0.00	0.37	0.30	0.43	1.97	0.30	0.10
13	Buscar trabajo	1.55	1.16	479	0	0	1	0	0	0	0	4.73	1.01	0.51	0.30	0.64	0.13	0.84	2.64	0.55	0.06
7	Buscar/Dejar alguien	1.46	0.99	138	0	0	0	0	0	0	0	1	0.08	1.06	0.43	0.02	0.63	0.00	1.60	0.03	0.38
8	Buscar/Dejar algo	1.26	0.65	48	1	1	0	1	1	0	0	1	0	2.41	1.97	1.27	1.81	1.60	0.00	1.26	1.41
5	Ver a alguien	1.36	1.09	257	0	0	0	1	1	0	0	1	0	0	1.41	0.71	1.33	0.96	0.84	0.73	0.97
16	Act. física y deporte	1.47	0.82	244	0	0	0	1	0	0	0	1	0	0	0	0.09	0.63	0.08	1.95	0.03	0.36
14	Act. Religiosa	1.61	1.10	102	0	0	0	0	0	0	0	1	0	0	0	0	0.24	1.06	2.41	0.82	0.36
12	Recreación y cultura	1.57	1.09	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0.63	1.81	0.51	0.14
9	Comer/Tomar algo	1.46	1.00	65	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1.27	0.04	0.36
15	Cuidado de personas	1.47	0.90	46	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.00	0.31
77	Otro	1.54	0.77	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00

Fuente: elaboración propia.

Como resultado del análisis presentado en esta se definen cómo objeto de análisis de los motivo de viaje: volver a casa, trabajar, estudiar, buscar trabajo, compras y otros.

4.2.3 Estimación de modelos

De acuerdo con el análisis presentado en la sección anterior, se trabajará con 6 motivos de viaje:

I. Volver a casa
II. Trabajar
III. Estudiar
IV. Buscar trabajo
V. Compras
VI. Otros

Los motivos de viaje se tratarán como variables cuantitativas discretas y constituyen las variables respuesta de interés. Se definieron 4 variables temporales explicativas de tipo cuantitativa continua:

I. Hora inicio del viaje (hora/24),
II. Hora final del viaje (hora/24),
III. Duración de la actividad (horas, hora inicio siguiente viaje – hora inicio viaje actual),
IV. Duración del viaje (horas)

Y 8 variables espaciales explicativas asociadas al uso del suelo en la zona de influencia de las estaciones de Transmilenio, de tipo cuantitativas continuas:

I. Vivienda (m^2)
II. Trabajo (m^2)
III. Comercio (m^2),
IV. Alimentación (m^2),
V. Educación (m^2),
VI. Cultura (m^2),
VII. Entretenimiento (m^2), y
VIII. Salud (m^2)

Como método de estimación del motivo del viaje se utiliza el modelo de tipo logit multinomial, considerando que éste permite realizar la estimación de elecciones según un conjunto de características explicativas, en donde la probabilidad de elección de la alternativa i (Pi) está dada por: (Greene, 2003)

$$P_i = \Pr ob(Y_i = j) = \frac{e^{(\beta'_j x_i)}}{\sum_{k=0}^{J} e^{(\beta'_k x_i)}}, \text{ para j} = 0,1,2,3,4$$

Donde:

Pi: Probabilidad de elección del motivo i

β: Vector de parámetro estimado de la función

x: Vector de variables

j: Alternativas

En este caso, el modelo estimado nos dará la probabilidad para un motivo de viaje comparado con los demás motivos del viaje (probabilidad del motivo i, respecto al motivo j), considerando que la sumatoria de las probabilidades debe ser igual a uno (1).

Los modelos logit multinomial se estiman a partir de una base de datos con 14.300 registros de viajes que provienen de la EODH. La base de datos incluye para cada registro de viaje, las variables explicativas espaciales y temporales y el motivo de viaje. El conjunto de datos se separa en dos grupos, uno para estimar el modelo (70% de los datos) y otro para la evaluación de los modelos (30% de los datos).

La estimación de los modelos se realiza inicialmente considerando de forma independiente las variables temporales, luego las variables espaciales y finalmente una combinación de variables; esto con el fin de entender el aporte de cada una de ellas. Los modelos son estimados utilizando en el software estadístico de R, con el paquete “multinomial”.

4.2.4 Modelo logit multinomial variables temporales

Para el modelo logit multinomial que considera las variables temporales: hora inicio del viaje, hora final del viaje, duración de la actividad y duración del viaje, se seleccionan aleatoriamente de la base total 9.154 registros para la estimación (70%) y los 5.152 registros adicionales constituyen la muestra para la validación del modelo (30%).

En la Tabla 4-2 se presentan los resultados de la estimación de parámetros del modelo logit multinomial usando las cuatro variables temporales, incluyendo tanto el estimador del parámetro de cada variable, cómo el p-value para cada parámetros asociado con cada variable independiente, que acompaña la prueba estadística que permite evaluar la significancia del parámetro dentro del modelo, donde se espera que el p-value sea inferior a 0,05 para un nivel de confianza del 95%.

Tabla 4-2 Parámetros y p-value, modelo logit multinomial todas las variables temporales

	Motivo del viaje				
Variable	**Trabajar**	**Estudiar**	**Compras**	**Buscar Trabajo**	**Otro**
(Intercept)	0,082	-0,958	-2,018	-1,006	0,670
	(0,6865)	(0)	(0)	(0,0191)	(0)
Ho	-1,792	-0,795	-0,363	-2,632	-0,771
	(0)	(0)	(0,219)	(0)	(0)
Hf	-1,802	-0,808	-0,370	-2,622	-0,772
	(0)	(0)	(0,21)	(0)	(0)
DuracionActividad_h	0,652	0,558	0,320	0,316	0,328
	(0)	(0)	(0)	(0)	(0)
Duracion_h	-0,241	-0,313	-0,157	0,235	-0,015
	(0)	(0)	(0,004)	(0)	(0,649)
Residual Deviance:	14710,4				
AIC:	14750,4				

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

Para los resultados del modelo de la Tabla 4-2, se obtienen que la totalidad de los parámetros para los motivos estudiar y buscar trabajo, son estadísticamente significativos considerando que su p-value es inferior a 0,05; para el motivo compras, las variables hora de inicio del viaje y hora de fin del viaje, tiene parámetros que no son estadísticamente significativos, con valor de p-value superior a 0,05.

El Residual Desviance y el criterio de información Akaike - AIC, son medias relativas de la calidad del modelo cuando se utilizan las mismas variables, y de utilidad para comparar diferentes tipos de modelos, en donde el modelo con el valor del AIC más pequeño es el mejor.

Para evaluar la calidad del modelo se calculó una matriz de confusión aplicada sobre la totalidad de los datos de entrenamiento (70%) y los datos de evaluación (30%), con los

parámetros obtenidos del modelo de la Tabla 4-2, estimando los motivos de viajes con los valores de las variables para todos los registros, obteniendo la probabilidad de cada uno de los motivos de viaje, con respecto al motivo volver a casa, seleccionando cómo motivo estimado aquel que presenta la mayor probabilidad de ocurrencia.

La Tabla 4-3, presenta los resultados de estimación del motivo a partir del modelo tanto para los datos de entrenamiento, como de evaluación. La matriz de confusión que compara la predicción de los datos en el eje izquierdo, frente a los motivos de viaje observados en la parte superior, donde se obtiene que al utilizar el modelo logit multinomial con las variables temporales, el 24% de las estimaciones son erróneas; a su vez, el modelo es incapaz de predecir los motivos estudiar, compras y buscar trabajo.

Tabla 4-3 Matriz de confusión todas las variables temporales

	Clasificaciones	Entrenamiento: 76,0%						Evaluación: 76,0%					
Id.	**Motivo**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	Volver a casa	3663	93	12	10	1	133	2065	58	11	8	3	76
2	Trabajar	82	1847	457	11	18	268	37	1036	263	4	13	126
3	Estudiar	3	0	0	0	0	0	1	0	0	0	0	0
4	Compras	0	0	0	0	0	0	0	0	0	0	0	0
5	Buscar trabajo	0	0	0	0	0	0	0	0	0	0	0	0
6	Otro	330	304	270	129	78	1444	192	171	165	66	40	817

Fuente: elaboración propia.

A partir de la matriz de confusiones también se observa que el motivo uno (1) volver a casa, el cual considera 2.295 registros de viajes por este motivo de acuerdo con EODH (suma columna), el modelo estima únicamente 2.221 registros con este motivo (suma fila), lo que significaría que un 3% de las predicciones del modelo son erróneas; no obstante, las coincidencias exactas de la matriz son de 2.065, lo que implica un error del 10% para el motivo volver a casa.

Considerando que la duración del viaje (Duración_h), está estimado como la diferencia entre la hora final del viaje y la hora de inicio del viaje, y que estas dos son variables independientes dentro del modelo, se calcula un nuevo modelo, sin considerar la variable explicativa hora del fin del viaje, considerando su similitud en magnitud de parámetros con la hora inicio del viaje. Los resultados se presentan en la Tabla 4-4, que incluye los parámetros estimados y en la parte inferior el resultado del p-value en paréntesis, los valores en gris representan aquellos p-value superiores al 0,05.

Tabla 4-4 Parámetros y p-value, modelo logit multinomial variables temporales

	Motivo del viaje				
Variable	**Trabajar**	**Estudiar**	**Compras**	**Busca trabajo**	**Otro**
(Intercept)	0,082	-0,958	-2,018	-1,006	0,670
	(0,6861)	(0)	(0)	(0,0191)	(0)
Ho	-3,595	-1,603	-0,733	-5,255	-1,543
	(0)	(0)	(0,215)	(0)	(0)
DuraciónActividad_h	0,651	0,558	0,320	0,316	0,328
	(0)	(0)	(0)	(0)	(0)
Duracion_h	-0,316	-0,346	-0,172	0,126	-0,048
	(0)	(0)	(0,002)	(0,055)	(0,165)
Residual Deviance:	14710,4				
AIC:	14750,4				

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

La Tabla 4-4 muestra que solamente el motivo estudiar presenta significancia para un p-value de 0,05 en todos sus parámetros, debido que esto no se cumple para el parámetro del intercepto para el motivo trabajo, ni para el parámetro de la hora de inicio para el motivo compras y ni para la duración del viaje para los motivos buscar trabajo y otros; adicionalmente, la duración de la actividad, tiene significancia en sus parámetros para todos los motivos de viaje.

Sobre los resultados obtenidos, se genera la matriz de confusión y se obtiene la cantidad de aciertos en la estimación del modelo, que se presenta en la Tabla 4-5, la cual es idéntica a la obtenida en el modelo anterior, lo cual indica que la ausencia de la variable hora final del viaje en el modelo no modifica el poder de predicción y su inclusión en el modelo no es necesaria.

Tabla 4-5 Matriz de confusión variables temporales

	Clasificaciones	Entrenamiento: 76,0%						Evaluación: 76,0%					
Id.	**Motivo**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	Volver a casa	3663	93	12	10	1	133	2065	58	11	8	3	76
2	Trabajar	82	1847	457	11	18	268	37	1036	263	4	13	126
3	Estudiar	3	0	0	0	0	0	1	0	0	0	0	0
4	Compras	0	0	0	0	0	0	0	0	0	0	0	0
5	BuscarTrabajo	0	0	0	0	0	0	0	0	0	0	0	0
6	Otro	330	304	270	129	78	1444	192	171	165	66	40	817

Fuente: elaboración propia.

Con el propósito de explorar la gran cantidad de combinaciones de variables temporales para modelar el motivo de viaje, se estima para las posibles combinaciones de variables,

su modelo asociado y se obtiene los porcentajes de valores no predichos tanto para la muestra de entrenamiento, como para la de evaluación, evaluando para el caso de las variables temporales 15 combinaciones de variables, como se presentan en la Tabla 4-6. Se observa que los primeros cuatro modelos presentan los menores errores en clasificación del motivo de viaje, los cuales involucran todas las variables.

Tabla 4-6 Comparación de modelos temporales según matriz de confusión valores no predichos

Ho	Hf	DuracionActividad_h	Duracion_h	Núm, Variables	Entrenamiento	Evaluación
Ho	Hf	DuracionActividad_h		3	24,02%	23,95%
Ho	Hf	DuracionActividad_h	Duracion_h	4	24,02%	23,95%
Ho		DuracionActividad_h	Duracion_h	3	24,02%	23,95%
	Hf	DuracionActividad_h	Duracion_h	3	24,02%	23,95%
		DuracionActividad_h	Duracion_h	2	23,93%	24,20%
	Hf	DuracionActividad_h		2	24,16%	24,38%
Ho		DuracionActividad_h		2	24,45%	24,69%
		DuracionActividad_h		1	24,27%	24,79%
Ho				1	36,77%	36,10%
Ho	Hf			2	37,07%	36,61%
Ho			Duracion_h	2	37,07%	36,61%
Ho	Hf		Duracion_h	3	37,07%	36,61%
	Hf		Duracion_h	2	37,07%	36,61%
	Hf			1	37,52%	36,76%
			Duracion_h	1	55,62%	55,67%

Fuente: elaboración propia.

4.2.5 Modelo logit multinomial variables espaciales

Para estimar el modelo logit multinomial con variables espaciales, se asocia a cada uno de los viajes de la EODH los metros cuadrados de cada tipo de uso del suelo tanto en el origen como en el destino y se crea una variable adicional que asocia el porcentaje de cada uno de los tipos de suelo en el origen y en el destino. La estimación de este modelo parte entonces de considerar 16 variables, 8 tipo de uso del suelo asociada al origen del viaje y 8 asociadas con los usos en el destino.

La Tabla 4-7, muestra los resultados de errores en la clasificación de la matriz de confusión tanto para los datos de entrenamiento (70%), como para los datos de evaluación (30%). De acuerdo con estos resultados no es posible predecir los motivos de estudiar, comprar, buscar trabajo ni otros, cuando sólo se incluyen las variables de uso en el origen. En contraste, las variables de usos del suelo en el destino muestran un mejor poder de predicción, ya que muestran repuesta a los motivos estudio y otros.

Por otra parte, al comparar las predicciones que utilizan las variables de uso del suelo en metros cuadrados (m²) y en porcentaje, se obtiene que el error de la matriz de confusión es menor cuando las variables se incluyen en metros cuadrados (m²).

Tabla 4-7 Resultados errores de matriz de confusión

Variable espacial	Núm. Variables	Entrenamiento	Evaluación	Nota
Todos los orígenes m^2	8	47,76%	47,88%	No predice motivos de 3 a 6
Todos los orígenes porcentaje	8	49,67%	49,55%	No predice motivos de 3 a 6
Todos los destinos m^2	8	47,07%	45,85%	No predice motivos 4 y 5
Todos los destinos porcentaje	8	48,47%	47,13%	No predice motivos 4 y 5
Todas origen y destino m^2	16	44,91%	44,58%	No predice motivos 4 y 5
Todas destino y destino porcentaje	16	45,96%	44,83%	No predice motivos 4 y 5

Fuente: elaboración propia.

Adicionalmente, se exploró el poder explicativo de incluir variables que relacionan el uso del suelo del origen y el uso del suelo del destino, asociado a cada motivo de viaje, ya que se distingue la presencia de viajes pendulares casa-trabajo o casa-estudio. En la Tabla 4-8 se muestran las relaciones entre el uso del suelo y el motivo de viaje.

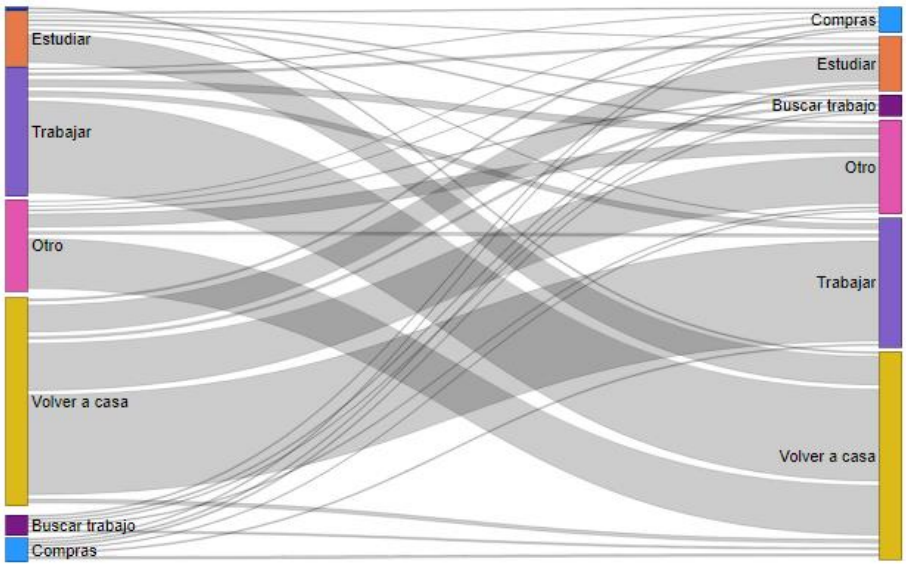
Tabla 4-8 Relación uso del suelo y motivo del viaje

Uso de suelo	Motivo del viaje
Vivienda	Volver a casa
Trabajo	Trabajar
Educación	Estudiar
Comercio	Compras
Cultura	Otro
Salud	Otro
Entretenimiento	Otro
Comer	Otro

Fuente: elaboración propia.

La Figura 4-9 presenta al lado derecho el motivo del viaje y al lado izquierdo el motivo del viaje anterior. Gran parte de los motivos volver a casa, llegan principalmente de los motivos estudiar, trabajar y otros; y para el caso del motivo estudiar, principalmente llegan de un motivo del viaje previo a volver a casa.

Figura 4-9 Motivo anterior y motivo del viaje



Fuente: elaboración propia.

Considerando lo anterior, la Tabla 4-9 presenta para cada uno de los motivos del viaje en la primera columna, su porcentaje del motivo del viaje anterior, lo cual muestra que el 50% de los viajes con motivo anterior trabajo, tienen motivo de viaje volver a casa. Las celdas marcadas en gris muestran las relaciones seleccionadas entre el área del suelo del origen (motivo del viaje) y el área del uso del suelo del destino (motivo del viaje anterior), tomando los motivos que permiten representar más del 85% de los viajes.

Tabla 4-9 Porcentaje motivo del viaje y motivo del viaje anterior

		Motivo del viaje anterior						
	Motivo	**Volver a casa**	**Trabajar**	**Estudiar**	**Buscar trabajo**	**Compras**	**Otro**	**Total**
Motivo del viaje	Volver a casa	2,8%	50,1%	15,9%	1,7%	2,0%	27,5%	100,0%
	Trabajar	89,2%	6,2%	0,8%	0,0%	0,2%	3,7%	100,0%
	Estudiar	80,7%	10,6%	4,3%	0,3%	0,1%	4,0%	100,0%
	Compras	50,3%	13,8%	4,6%	3,0%	5,0%	23,3%	100,0%
	Buscar trabajo	79,6%	0,0%	1,6%	10,6%	0,1%	8,2%	100,0%
	Otro	64,8%	11,2%	3,7%	0,5%	0,9%	19,0%	100,0%

Fuente: elaboración propia.

De esta forma, se seleccionan y calculan 6 variables que relacionan el uso del suelo en el origen y el destino, calculadas de forma similar a como se presenta para la relación del uso de suelo trabajo, usando la siguiente expresión:

$$Relación\ uso\ Trabajo\_DO = \frac{Trabajo\ (m2)}{Vivienda\ (m2)}$$

El modelo con las variables explicativas que representa las relaciones del área de uso del suelo del destino y el origen (6 variables) se presenta en la Tabla 4-10, con un error del 44,9% para los datos de evaluación, muy cercano al menor error del modelo la Tabla 4-7, con la ventaja adicional de lograr reducir el número de variables, pasando de 16 a 6 variables.

Tabla 4-10 Resultados errores de matriz de confusión, relación entre usos

Variable espacial	Núm, Variables	Entrenamiento	Evaluación	Nota
Relación entre usos	6	45,28%	44,90%	No predice motivos 4 y 5

Fuente: elaboración propia.

En la Tabla 4-11 se presentan los parámetros estimados y los de p-value, para el modelo que usa variables de relación de usos del suelo, en donde la mayoría de las variables presentan un valor de significancia mayor al 95%, exceptuando aquellas que se resaltan en color gris; las variables asociadas a los motivos compras y buscar trabajo son aquellas que presentan menor significancia en su parámetro estimado.

Tabla 4-11 Parámetros y p-value, modelo logit multinomial todas las variables espaciales relacionadas

	Motivos de viaje				
Variable	**Trabajar**	**Estudiar**	**Compras**	**Buscar trabajo**	**Otro**
(Intercept)	-2,446	-3,573	-4,626	-5,565	-2,450
	(0)	(0)	(0)	(0)	(0)
Casa_DO	0,199	0,129	0,050	0,208	0,175
	(0)	(0)	(0,215)	(0)	(0)
Trabajo_DO	-0,066	-0,115	-0,053	-0,073	-0,048
	(0,003)	(0)	(0,119)	(0,178)	(0,03)
Estudio_DO	0,577	6,886	-1,126	-0,912	1,509
	(0,187)	(0)	(0,42)	(0,521)	(0,001)
Comercio_DO	-1,571	0,154	3,785	0,076	-0,050
	(0)	(0,769)	(0)	(0,931)	(0,912)
BuscarTrabajo_DO	3,226	2,539	2,081	3,170	2,477
	(0)	(0)	(0)	(0)	(0)
Otros_DO	10,801	4,885	7,287	6,356	13,478
	(0)	(0,009)	(0,045)	(0,082)	(0)
Residual Deviance:	21089,24				
AIC:	21159,24				

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

Sobre los resultados de la matriz de confusión de la Tabla 4-12, se obtiene la cantidad de aciertos en la estimación del modelo, evidenciándose que menos de la mitad de las veces se obtiene un error en la clasificación, tanto para los grupos de datos de entrenamiento, como para los de evaluación.

Tabla 4-12 Matriz de confusión todas las variables espaciales relacionadas

	Clasificaciones	Entrenamiento: 54,7%						Evaluación: 55,1%					
Id.	**Motivo**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	Volver a casa	3684	944	258	69	42	948	2085	505	156	40	19	540
2	Trabajar	310	1083	296	68	46	683	163	629	173	36	31	363
3	Estudiar	37	77	81	1	3	54	18	44	46	0	3	37
4	Compras	0	0	0	0	0	0	0	0	0	0	0	0
5	Buscar trabajo	0	0	0	0	0	0	0	0	0	0	0	0
6	Otro	47	140	104	12	6	161	29	87	64	2	3	79

Fuente: elaboración propia.

De igual forma que para el caso de las variables temporales, se pueden obtener para las variables espaciales diferentes clasificaciones, según las variables que se consideran en cada uno de los modelos, por esta razón en el Anexo C se presentan los resultados obtenidos para los 52 modelos que pueden ser construidos con las variables espaciales. La Tabla 4-13, presenta los primeros 10 modelos con menores errores de clasificación obtenidos a partir de las matrices de confusión tanto para entrenamiento como para evaluación. En este caso, los errores entre el modelo con menos y más errores de clasificación puede llegar a ser del 10%.

Tabla 4-13 Comparación de modelos espaciales según matriz de confusión valores no predichos

Casa	**Trabajo**	**Estudio**	**Comercio**	**Buscar trabajo**	**Otros**	**Núm. Variables**	**Entrenamiento**	**Evaluación**
Casa	Trabajo	Estudio	Comercio	BuscarTrabajo	Otros	6	45,28%	44,90%
Casa		Estudio	Comercio	BuscarTrabajo	Otros	5	45,25%	44,97%
Casa		Estudio		BuscarTrabajo	Otros	4	45,43%	44,86%
Casa		Estudio	Comercio	BuscarTrabajo		4	45,63%	44,95%
Casa			Comercio	BuscarTrabajo	Otros	4	45,50%	45,13%
		Estudio		BuscarTrabajo	Otros	3	45,97%	44,72%
Casa	Trabajo	Estudio	Comercio	BuscarTrabajo		5	45,70%	44,99%
	Trabajo	Estudio		BuscarTrabajo	Otros	4	46,00%	44,88%
Casa		Estudio		BuscarTrabajo		3	45,78%	45,11%
Casa				BuscarTrabajo	Otros	3	45,95%	44,97%

Fuente: elaboración propia.

Del total de modelos, se observa que el modelo con menores errores en clasificación emplea todas las variables, no obstante, esto puede ocasionar dificultades en el momento de conseguir los datos y también en el análisis de los modelos. Basado en el principio de parsimonia se recomienda el uso del primer modelo de cuatro (4) variables, considerando que los errores no varían en más de 0,2% con respecto al primer modelo de 6 variables. El cual considera las relaciones de área del uso de suelo de casa, estudio, otros y buscar trabajo.

En la Tabla 4-14, se presentan los parámetros de las cuatro variables empleadas para obtener el motivo del viaje, en este sentido la mayoría de las variables presenta un buen ajuste, exceptuando los parámetros de estudio y otros, en los casos de estimación de los motivos compras y buscar trabajo, que presentan un p-value superior a 0,05.

Tabla 4-14 Parámetros y p-value, modelo logit multinomial variables espaciales

	Motivos de viaje				
Variable	**Trabajar**	**Estudiar**	**Compras**	**Buscar trabajo**	**Otro**
(Intercept)	-2,435	-3,588	-4,859	-5,589	-2,451
	(0)	(0)	(0)	(0)	(0)
Casa_DO	0,195	0,138	0,082	0,208	0,175
	(0)	(0)	(0,028)	(0)	(0)
Estudio_DO	0,230	6,690	-0,061	-0,788	1,538
	(0,595)	(0)	(0,96)	(0,563)	(0)
BuscarTrabajo_DO	2,891	2,586	3,305	3,225	2,470
	(0)	(0)	(0)	(0)	(0)
Otros_DO	11,399	4,211	3,551	5,422	13,038
	(0)	(0,023)	(0,301)	(0,126)	(0)
Residual Deviance:	21183,4				
AIC:	21233,4				

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

Por otra parte, la matriz de confusión tanto para el entrenamiento como la evaluación arrojan los resultados que se presentan en la Tabla 4-15, donde cerca del 55% de las clasificaciones no presentan errores, a su vez no hay estimaciones para el motivo compras y buscar trabajo.

Tabla 4-15 Matriz de confusión variables espaciales

	Clasificaciones	Entrenamiento: 54,6%						Evaluación: 55,1%					
Id.	**Motivo**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	Volver a casa	3683	968	263	66	44	960	2081	511	155	40	19	546
2	Trabajar	329	1119	315	79	46	724	182	664	188	36	32	384
3	Estudiar	38	71	78	1	3	47	16	41	42	0	3	35
4	Compras	0	0	0	0	0	0	0	0	0	0	0	0
5	Buscar trabajo	0	0	0	0	0	0	0	0	0	0	0	0
6	Otro	28	86	83	4	4	115	16	49	54	2	2	54

Fuente: elaboración propia.

4.2.6 Estimación del modelo con variables temporales y espaciales

A partir de la exploración de los modelos que usan variables temporales o espaciales, se observa que la inclusión de tres (3) variables temporales permiten clasificaciones con un acierto del 76% y el modelo con cuatro (4) variables espaciales permite una clasificación con un acierto del 55%. A partir de estas 7 variables se estima un modelo logit multinomial, en donde los efectos conjuntos en la estimación de los motivos de viaje y los parámetros se muestran en la Tabla 4-16.

Tabla 4-16 Parámetros y p-value, modelo logit multinomial variables espaciales y temporales

	Motivos de viaje				
Variable	**Trabajar**	**Estudiar**	**Compras**	**Buscar trabajo**	**Otro**
(Intercept)	-1,272	-2,475	-3,268	-2,376	-0,653
	(0)	(0)	(0)	(0)	(0,0004)
Ho	-3,033	-1,056	-0,271	-4,739	-1,037
	(0)	(0,009)	(0,662)	(0)	(0)
DuracionActividad_h	0,627	0,539	0,297	0,289	0,306
	(0)	(0)	(0)	(0)	(0)
Duracion_h	-0,305	-0,330	-0,151	0,133	-0,028
	(0)	(0)	(0,009)	(0,048)	(0,437)
Casa_DO	0,071	0,027	0,003	0,098	0,077
	(0)	(0,183)	(0,926)	(0)	(0)
Estudio_DO	-1,010	5,153	-1,727	-2,505	-0,115
	(0,107)	(0)	(0,177)	(0,081)	(0,839)
BuscarTrabajo_DO	1,785	1,562	2,397	2,236	1,549
	(0)	(0)	(0)	(0)	(0)
Otros_DO	9,958	2,841	0,869	2,793	10,813
	(0)	(0,218)	(0,816)	(0,457)	(0)
Residual Deviance:	13884,91				
AIC:	13964,91				

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

La matriz de confusión, que se presenta en la Tabla 4-17, muestra que el 76% de las clasificaciones son correctas. Sin embargo, a pesar de incluir tanto las variables temporales y espaciales, no se obtiene respuesta asociada a los motivos compras, ni buscar trabajo, es decir que el modelo no tiene capacidad de estimar esos motivos de viaje.

Tabla 4-17 Matriz de confusión variables todas las variables espaciales y temporales

	Clasificaciones	Entrenamiento: 76,0%						Evaluación: 76,1%					
Id.	**Motivo**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	Volver a casa	3754	104	14	29	5	236	2130	55	9	13	4	143
2	Trabajar	79	1796	387	13	20	261	36	1011	221	7	11	121
3	Estudiar	9	56	87	1	0	33	4	41	51	0	3	26
4	Compras	0	0	0	1	1	1	0	0	0	0	0	2
5	Buscar trabajo	0	0	0	0	0	0	0	0	0	0	0	0
6	Otro	236	288	251	106	71	1314	125	158	158	58	38	727

Fuente: elaboración propia.

Debido a que el modelo estimado, no tiene poder de estimación para los motivos buscar trabajo y compras, se opta por agruparlo en la categoría, otros motivos. De esta manera, se estima nuevamente el modelo para predecir cuatro motivos de viaje:

- Volver a casa
- Trabajar
- Estudiar
- Otros

Para estos cuatro motivos de viaje, se estiman los parámetros del modelo logit multinomial que se presentan en la Tabla 4-18. Este modelo, mantienen siete (7) variables independientes, hora de inicio del viaje, duración de la actividad, duración del vieja, relación del uso del suelo en el origen y destino para casa, estudio, buscar trabajo y otros.

Se observa que la variable espacial asociada al uso de suelo de estudio (Estudio_DO), presenta valores de p-value, superiores al 0,05 para predecir viajes con motivo trabajo u otro. A su vez, las variables duración del viaje no es significativa a la hora estimar el motivo de viaje otro, al igual que ocurre con la variable hora de inicio del viaje H_o al estimar el motivo estudio.

Tabla 4-18 Parámetros y p-value, modelo logit multinomial todas las variables espaciales y temporales y cuatro motivos del viaje

	Motivo de viaje		
Variable	**Trabajar**	**Estudiar**	**Otro**
(Intercept)	-1,414	-2,849	-0,633
	(0)	(0)	(0,0009)
Ho	-3,042	-0,573	-1,076
	(0)	(0,166)	(0)
DuracionActividad_h	0,637	0,556	0,314
	(0)	(0)	(0)
Duracion_h	-0,236	-0,286	-0,006
	(0)	(0)	(0,856)
Casa_DO	0,084	0,050	0,081
	(0)	(0,014)	(0)
Estudio_DO	-0,978	4,532	-0,096
	(0,126)	(0)	(0,868)
BuscarTrabajo_DO	1,846	1,612	1,702
	(0)	(0)	(0)
Otros_DO	10,209	5,578	10,779
	(0)	(0,016)	(0)
Residual Deviance:	11743,12		
AIC:	11791,12		

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

La matriz de confusión de este modelo se presenta en la Tabla 4-19, para los datos de entrenamiento y de evaluación, con una mejora en el poder de estimación de motivos de viaje, llegando a un 78% y 77% para datos de entrenamiento y evaluación respectivamente.

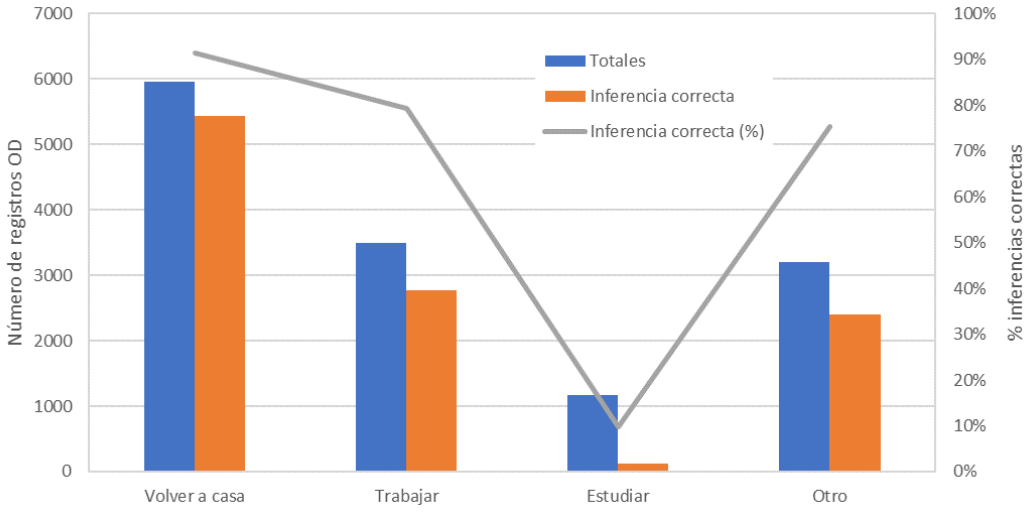
Tabla 4-19 Matriz de confusión variables espaciales y temporales

	Clasificaciones	**Entrenamiento: 78,0%**				**Evaluación: 77,2%**			
Id.	Motivo	1	2	3	6	1	2	3	6
1	Volver a casa	3506	87	13	217	1936	58	5	136
2	Trabajar	71	1811	376	238	31	958	219	146
3	Estudiar	5	49	72	39	3	37	43	16
6	Otro	258	314	277	1496	149	174	160	909

Fuente: elaboración propia.

De forma gráfica en la Figura 4-10, se presentan los porcentajes de inferencia correctas obtenidas de la matriz de confusión, en donde el motivo estudio que representa el 8,4% de los datos, solamente permite estimar cerca del 10% de las inferencias de forma correctas.

Figura 4-10 Inferencia 4 motivos, total de variables espaciotemporales



Fuente: elaboración propia.

De igual forma, que como se ha realizado para las variables temporales y espaciales, se estiman los modelos para las posibles combinaciones de las 7 variables independientes, con el fin de identificar aquella combinación que reportan los menores errores en la estimación, para los cuatro motivos de viaje. Se estiman y evalúan un total de 127 modelos, los cuales se han organizado de forma ascendente según su error en evaluación y entrenamiento, y se pueden consultar en el Anexo D. La Tabla 4-20 presenta los resultados de los 10 modelos con menores errores de estimación. El mejor modelo se ha obtenido solamente empleando las tres variables temporales, hora de inicio H_o, duración de la actividad y duración del viaje, para predecir los cuatro motivos del viaje.

Tabla 4-20 Comparación de los 10 mejores modelos espaciotemporales según matriz de confusión para valores no predichos

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
Ho	DuracionActividad_h	Duracion_h					Temporal	3	21.41%	22.11%
Ho	DuracionActividad_h	Duracion_h	Casa_DO				Espaciotemporal	4	21.94%	22.25%
	DuracionActividad_h	Duracion_h					Temporal	2	21.60%	22.75%
Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO			Espaciotemporal	5	21.98%	22.41%
Ho	DuracionActividad_h	Duracion_h		Estudio_DO			Espaciotemporal	4	21.96%	22.51%
Ho		Duracion_h				Otros_DO	Espaciotemporal	3	21.89%	22.65%
Ho	DuracionActividad_h	Duracion_h				Otros_DO	Espaciotemporal	4	21.89%	22.65%
Ho		Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	4	21.97%	22.65%
Ho	DuracionActividad_h	Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	5	21.97%	22.65%
Ho	DuracionActividad_h						Temporal	2	21.98%	22.65%

Fuente: elaboración propia.

Los parámetros del modelo con mejor clasificación para los cuatro motivos de viaje (volver a casa, trabajar, estudiar y otro), que utiliza tres (3) variables independientes, se presenta en la Tabla 4-21.

Tabla 4-21 Parámetros y p-value, modelo logit multinomial con mejor clasificación

	Motivo de viaje		
Variable	**Trabajar**	**Estudiar**	**Otro**
(Intercept)	-0,013	-1,211	0,751
	(0,9497)	(0)	(0)
Ho	-3,568	-1,125	-1,542
	(0)	(0,004)	(0)
DuracionActividad_h	0,661	0,572	0,335
	(0)	(0)	(0)
Duracion_h	-0,237	-0,299	-0,010
	(0)	(0)	(0,756)
Residual Deviance:	12468,04		
AIC:	12492,04		

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

La Tabla 4-22 presenta los parámetros del modelo que combina variables temporales y espaciales. En este caso se observa que la variable incluida presenta un valor de p-value significativo para todos los parámetros estimados.

Tabla 4-22 Parámetros y p-value, modelo logit multinomial mejor clasificación con variables temporales y espaciales

	Motivo de viaje		
Variable	**Trabajar**	**Estudiar**	**Otro**
(Intercept)	-0,387	-1,634	0,380
	(0,0769)	(0)	(0,0265)
Ho	-3,400	-0,928	-1,384
	(0)	(0,018)	(0)
DuracionActividad_h	0,658	0,569	0,332
	(0)	(0)	(0)
Duracion_h	-0,246	-0,306	-0,019
	(0)	(0)	(0,554)
Casa_DO	0,125	0,137	0,127
	(0)	(0)	(0)
Residual Deviance:	12376,05		
AIC:	12406,05		

Nota: el p-value para un nivel de confianza está escrito en paréntesis.

Fuente: elaboración propia.

4.3 Discusión y conclusiones

Se construyeron un total de 194 modelos para explorar el aporte de variables espaciales, variables temporales y efectos combinados para la predicción del motivo de viaje en el sistema troncal del SITP. Se utilizaron pruebas de diferencia de medias y matrices de correlación de Pearson, que ayudaron en la agrupación de las categorías de uso del suelo y del motivo de viaje. De igual manera, se analizó el aporte al modelo agrupando variables con características similares, en este caso asociadas a variables espaciales y a variables temporales. Esto ayudó a evidenciar efectos puntuales de cada uno de los grupos, que soportan la interpretación de los modelos saturados (uso de todas las variables) y la selección del mejor modelo.

Los modelos que incluyen únicamente variables temporales mostraron desde el principio una mejor respuesta en la clasificación de los motivos de viaje trabajo y volver a casa, en donde la hora de inicio del viaje y la hora final del viaje, presentaron valores muy similares en su parámetro, resultando en que el mejor modelo solamente incluyera una de estas dos variables. Al incluir variables espaciales, asociadas al uso de suelo de vivienda y de trabajo en el origen y uso de suelo educacional en el destino, se observó una mejora en la clasificación del modelo con motivo estudiar; no obstante, debido a la cantidad de variables, se optó por generar variables adicionales que permitieran relacionar el uso del suelo en el origen y en el destino, obteniendo significancia para el total de parámetros en los casos donde el uso de suelo estaba asociado a la vivienda y al trabajo.

El mejor modelo evaluado según la matriz de confusión, para la clasificación de los cuatro motivos de viaje (volver a casa, trabajar, estudiar y otro), se obtiene solamente al emplear tres variables independientes de tipo temporal (hora de inicio del viaje, duración de la actividad y duración del viaje). Este resultado es interesante ya que con pocas variables (menos datos) se obtienen buenas estimaciones. No obstante, es importante considerar que las diferencias entre clasificaciones de los primeros 10 modelos no varían en más del 1% de poder de clasificación, lo que implica que el uso de un modelo sobre otro no debería provocar grandes variaciones.

Sobre los parámetros del modelo, se reporta con un nivel de confianza del 95%, el intercepto del motivo trabajar y la duración en horas del motivo otro, superan el valor del p-value (0,05), lo cual implica que no son significativas para la estimación. Por otra parte, la mejor variable independiente para la clasificación del motivo del viaje, es la duración de la actividad, ya que el p-value es igual a cero en todos los casos (Tabla 4-21); adicionalmente, cuando se utiliza solamente esta variable independiente para obtener el motivo del viaje, el error es del 24,79%, y a su vez, se encuentra siempre en los modelos con menores errores en la estimación.

A pesar de que el mejor modelo permite estimar más de tres cuartas partes de los viajes (Tabla 4-19), se evidencia un bajo poder de predicción del motivo estudio (Figura 4-10), lo cual puede estar relacionado a la cantidad de datos que se utilizan para su estimación (8,4%); a pesar de esto, se considera como el mejor modelo debido a que es el único motivo diferente a volver a casa, trabajar y otros, que muestra alguna respuesta a las variables independientes utilizadas; adicionalmente y debido a que el modelo logit multinomial empleado, estima la probabilidad respecto al motivo volver a casa, se espera que su efecto sobre la clasificación general no afecte de forma significativa las estimaciones, considerando que los parámetros estimados del modelo, en su mayoría, tienen un p-value inferior a 0,05 (Tabla 4-18), lo que sugiere que los parámetros son sin duda importantes para el modelo.

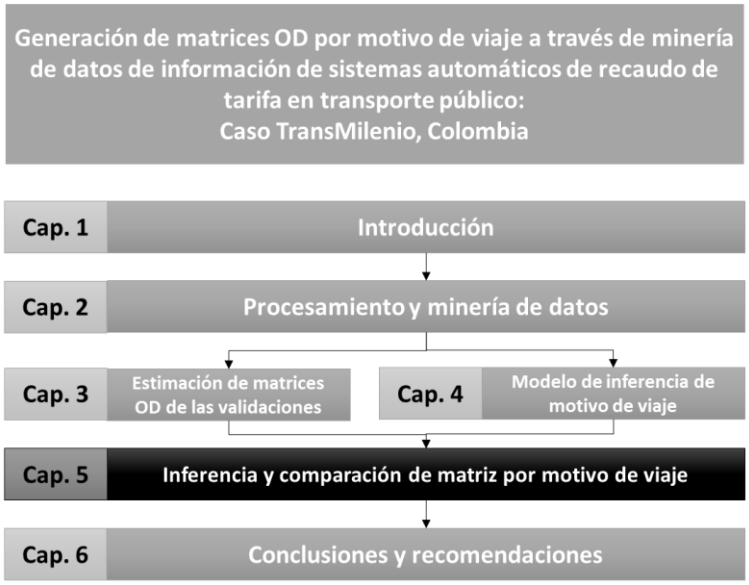
De esta forma, se concluye que el mejor modelo no solo es aquel que presenta los menores errores, sino aquel que con el menor número de variables permite realizar buenas estimaciones, ya que esto significa ahorro en recursos.

5. Inferencia y comparación de matrices por motivo de viaje

5.1 Introducción

La inferencia de matrices origen destino por motivo de viaje (ver Figura 5-1), es obtenida a partir de la aplicación del modelo de inferencia del motivo de viaje, sobre las matrices obtenidas por el método del encadenamiento de viajes, aplicado sobre los datos del uso de tarjetas inteligentes en el sistema troncal del SITP (TransMilenio).

Figura 5-1 Metodología general – Capítulo 5

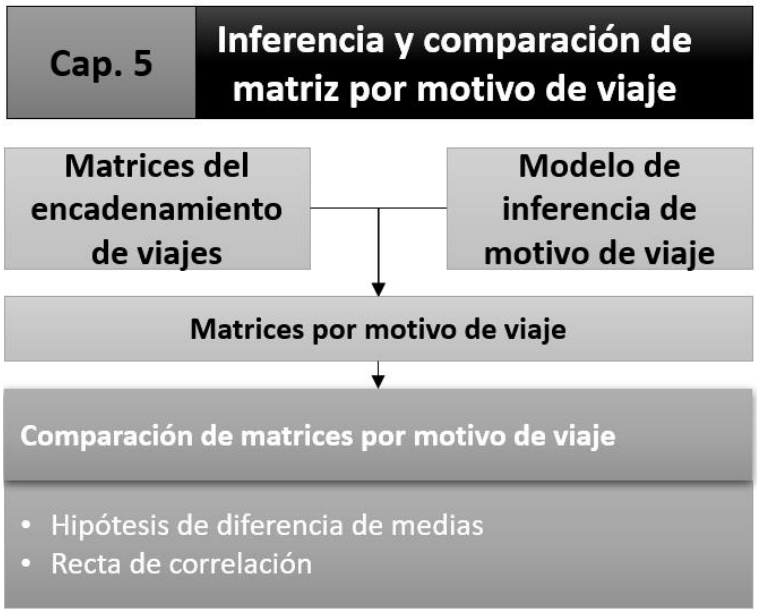


Fuente: elaboración propia.

El proceso de inferencia y validación de las matrices por motivo de viaje, comienza con la construcción de la base de datos que requiere, los valores de todas las variables usadas para aplicar el modelo logit multinomial estimado; en este caso incluyen la hora de inicio del viaje, la duración del viaje y la duración de la actividad.

La aplicación del modelo y los datos de las variables permiten estimar los motivos de viaje para cada una de las entradas a TransMilenio que pudieron ser encadenadas con el método de encadenamiento de viaje. De igual forma, con los resultados se analizan comportamientos de las variables asociadas a cada uno de los motivos predichos, y se contrastan por medio de pruebas de hipótesis de diferencias de medias y rectas de correlación frente a los datos de la EODH, proceso que se presentan en la Figura 5-2.

Figura 5-2 Método – Inferencia y comparación de matrices por motivo de viaje



Fuente: elaboración propia.

5.2 Inferencia de matrices con motivo de viaje

Luego de haber aplicado el encadenamiento de viajes sobre los datos del uso de tarjetas inteligentes de TransMilenio, es necesario complementar la información de aquellas variables que utiliza el modelo seleccionado, en este caso la hora de inicio del viaje, la duración de la actividad y la duración del viaje.

Para el caso de la hora de inicio del viaje, el uso de las tarjetas almacena de forma directa este atributo. Para la duración de la actividad, el destino del viaje es complementado con la hora en la que se realiza el siguiente viaje, y se calcula la diferencia entre la hora del siguiente viaje, menos la hora del viaje actual, para construir esta variable.

Para el caso de la duración del viaje, se toma la información del procesamiento y minería de datos de la sección 2.2.4, en donde se obtuvo la matriz de duración del viaje. Primero se calculan las duraciones del viaje que pueden ser completadas con solo una ruta, las cuales permiten llenar el 70% de las entradas. El 30% restante, es complementado con los tiempos obtenidos del modelo de redes.

Al aplicar el modelo de inferencia para el motivo de viaje, se obtienen los valores agregados en porcentajes que se presentan en la Tabla 5-1. Para los días hábiles (lunes a viernes), el motivo volver a casa representa más del 46%, el motivo trabajar más de 32% y el motivo otros más de un 20%; a pesar de que el modelo seleccionado consideró el motivo estudiar, no se obtiene respuesta, debido a que la probabilidad estimada para este motivo nunca es superior a los demás, lo cual puede derivarse de que el motivo estudiar según la EODH, solamente representa el 8,4% de los datos, provocando que la clasificación sea asignada entre los otros motivos; esto impide cualquier análisis posterior con este motivo.

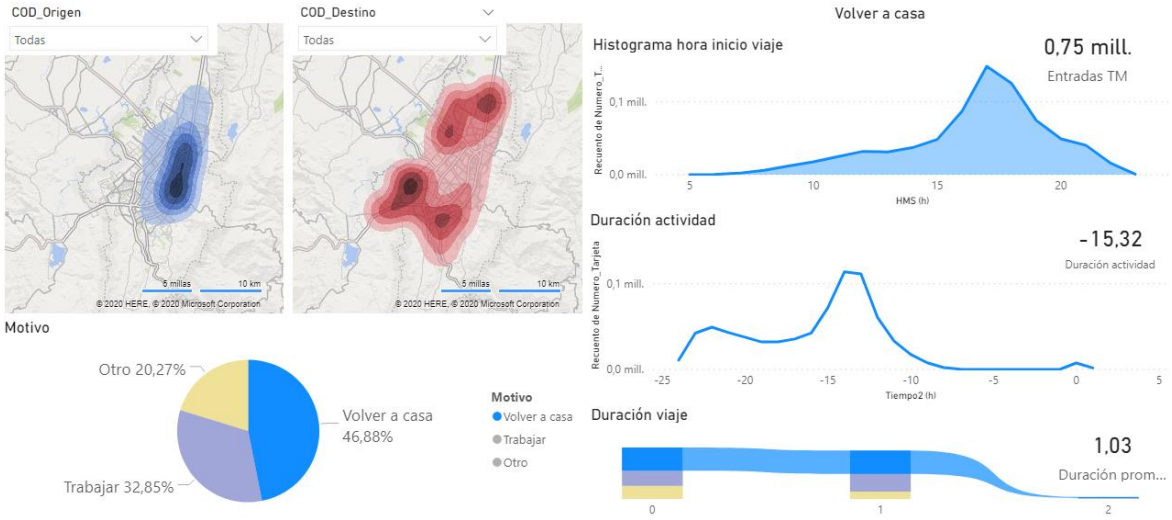
Tabla 5-1 Porcentaje de inferencia por motivo del viaje

Día	Volver a casa	Trabajar	Otros
Lunes	47,0%	32,8%	20,1%
Martes	46,9%	32,8%	20,3%
Miércoles	46,9%	32,9%	20,3%
Jueves	46,9%	32,9%	20,1%
Viernes	46,7%	32,3%	21,0%
Sábado	47,2%	27,9%	25,0%
Domingo	48,6%	26,5%	24,9%

Fuente: elaboración propia.

A su vez, se obtienen una caracterización similar a la de la EODH, presentada en la sección 2.3 Viajes en TransMilenio según encuesta origen destino en hogares. La Figura 5-3 muestra el comportamiento de la demanda para el motivo volver a casa, donde se observan tendencias similares entre los resultados obtenidos a partir de las dos matrices. Por ejemplo, los orígenes de los viajes se concentran en el centro de la ciudad, y los destinos en las periferias, cerca de los portales, la demanda del motivo volver a casa representa el 46,9% de la demanda diaria, con una concentración en el pico de la tarde a las 17 horas (5 p.m.); a su vez, se observa una duración de actividad promedio de 15,3 horas, con una moda de 13 a 14 horas.

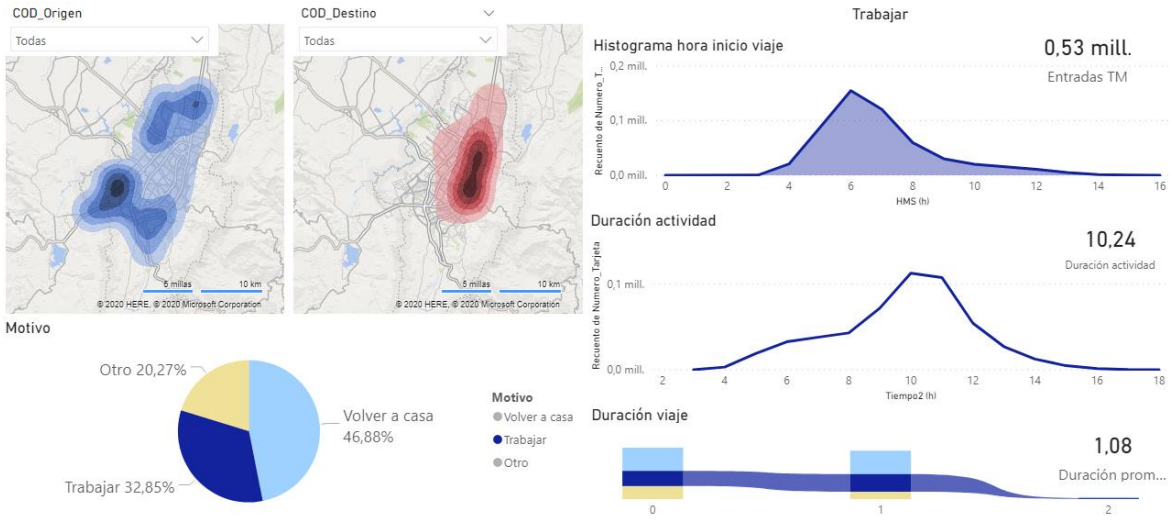
Figura 5-3 Motivo volver a casa – Inferencia matrices por motivo de viaje día hábil



Fuente: elaboración propia.

La Figura 5-4 muestra los comportamientos de viaje para el caso del motivo trabajar. Al igual que ocurre con lo hallado a partir de la EODH, los orígenes se concentran principalmente en las periferias de la ciudad, con destinos concentrados en el centro y el centro extendido de la ciudad, con una hora pico concentrada a las 6 horas (6 a.m.) y una duración promedio de actividad de 10,2 horas con moda entre las 10 y las 11 horas.

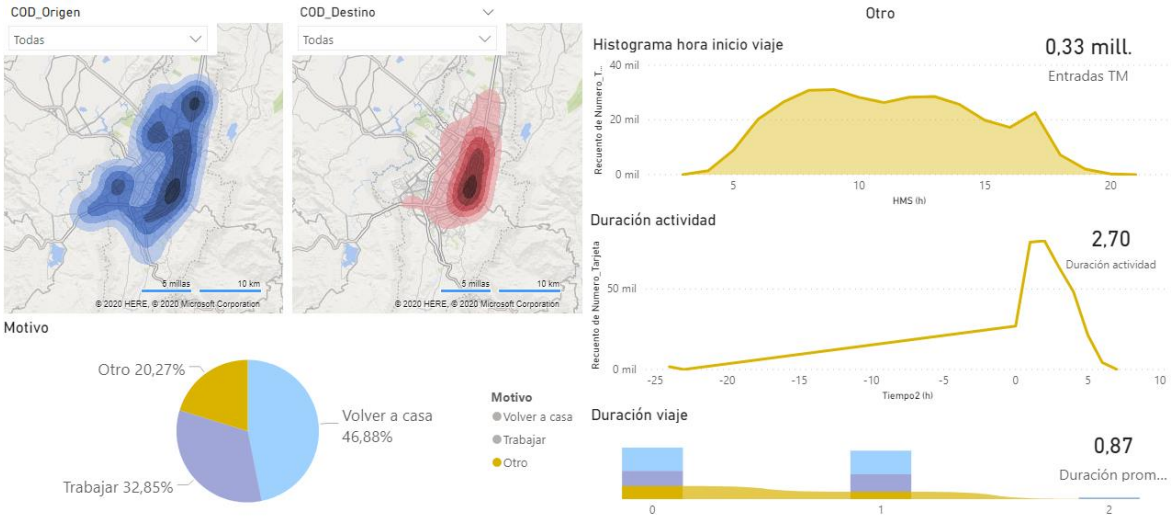
Figura 5-4 Motivo trabajar – Inferencia matrices por motivo de viaje día hábil



Fuente: elaboración propia.

Por otra parte, para el caso de motivo de viaje “otro”, se obtienen los resultados que se presentan en la Figura 5-5, en donde los orígenes de los viajes no presentan una concentración geográfica en una sola parte de la ciudad, sino que se distribuye en diferentes zonas, a diferencia de lo que ocurre con los destinos que se concentran en la zona céntrica de la ciudad. Los viajes con motivo otro, representan el 20,3% de la demanda, una quinta parte del total de los viajes. Con respecto a la concentración temporal, no existe una concentración marcada de la demanda en una hora, sino una distribución más o menos uniforme a lo largo del día. Finalmente, se reporta que la duración de la actividad es en promedio de 2,8 horas con una moda entre 1 a 2 horas.

Figura 5-5 Motivo otro – Inferencia matrices por motivo de viaje día hábil

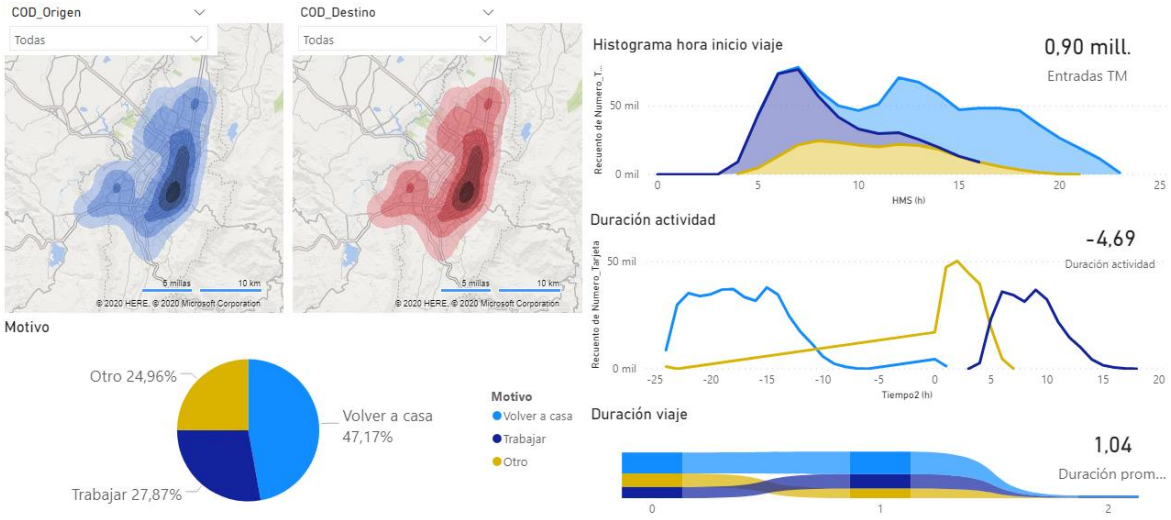


Fuente: elaboración propia.

Adicionalmente, se obtuvo información para el comportamiento de la demanda en el sábado y domingo, asociado a los motivos del viaje, lo cual no pudo ser contrastado con la EODH debido a que esta no tiene información para el fin de semana.

La Figura 5-6, muestra para el sábado que el 25% de los viajes tienen motivo otro, una mayor participación que en los días hábiles. También se reporta una menor concentración de la hora de inicio del viaje en un solo pico para el motivo volver a casa, y una menor duración de la actividad trabajar.

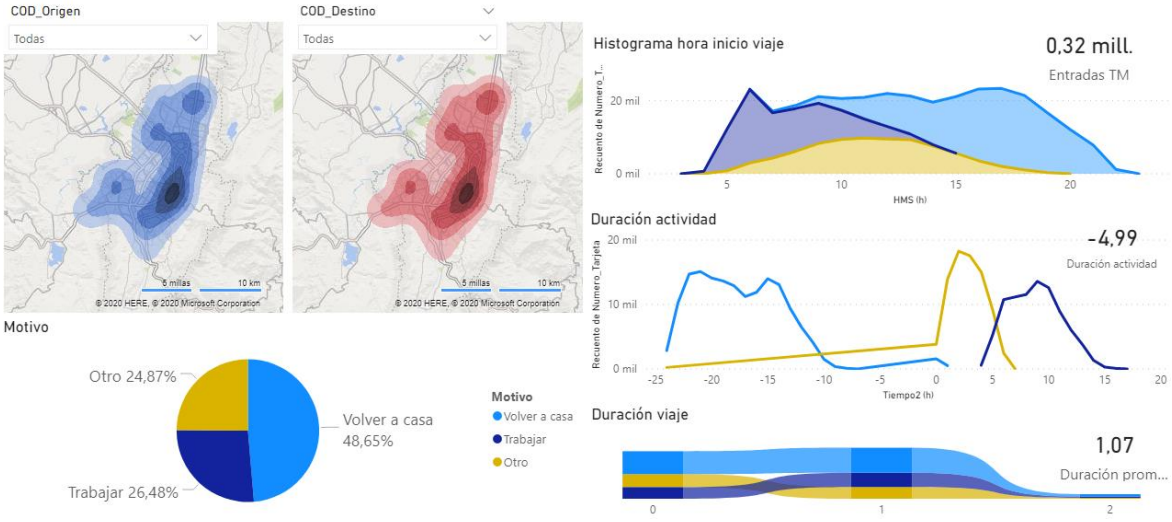
Figura 5-6 Inferencia matrices por motivo de viaje del sábado



Fuente: elaboración propia.

Las estimaciones para el domingo se presentan en la Figura 5-7. Se observa que el motivo trabajar pierde aún más participación en el total de viajes del día, y a su vez, los picos horarios son menos marcados a lo largo del día.

Figura 5-7 Inferencia matrices por motivo de viaje del domingo



Fuente: elaboración propia.

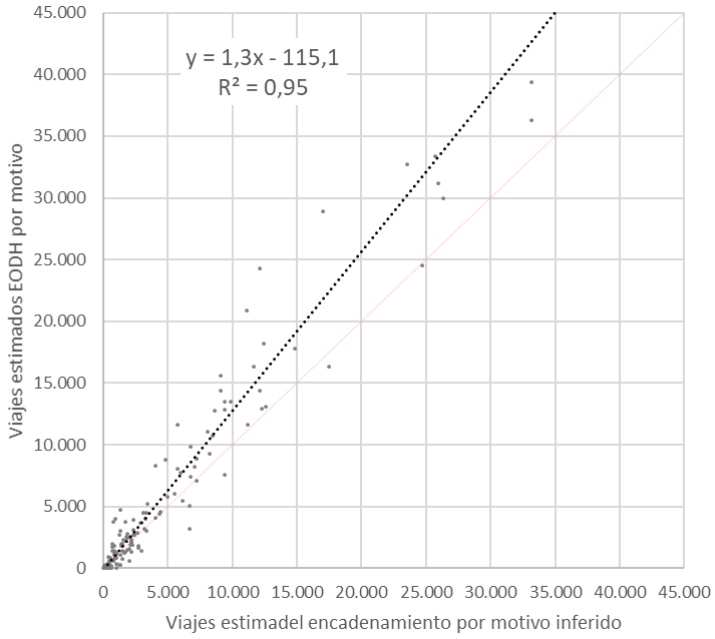
5.3 Comparación de matrices por motivo de viaje

Considerando los resultados obtenidos de la comparación de las matrices obtenidas del encadenamiento de viajes, donde se evidenció que la mejor comparación desagregada se

consigue a nivel de matrices de línea cuando se compara con la EODH, se aplica la comparación para cada uno de los motivos de viaje de igual forma como se presentó en la sección 3.4.

Al realizar el cálculo de la recta de regresión entre los viajes de la EODH por motivo volver a casa y los viajes estimados por el método de encadenamiento con motivo volver a casa, se evidencia una tendencia lineal, como se presenta en la Figura 5-8.

Figura 5-8 Comparación de matrices de viajes con motivo de viaje volver a casa



Fuente: elaboración propia.

Para probar si existe una diferencia estadística en la media de las estimaciones realizadas por los dos métodos, se realizó una prueba de hipótesis de medias, cuyos resultados se presentan junto con los de la regresión lineal en la Tabla 5-2. Para un nivel de confianza del 95%, se contrasta con un valor de $Z_{\alpha/2} = 1,96$, lo que implica que se acepta la hipótesis nula, considerando que Z* no es muy mayor Z, lo que indica que las medias de los valores de viajes obtenidos por los dos métodos no son estadísticamente diferentes. A su vez se observa un buen ajuste de la recta de regresión entre las estimaciones de las matrices de

los dos métodos, con un coeficiente de correlación de 0,95 y una pendiente del 1,3 y un intercepto de -115,1.

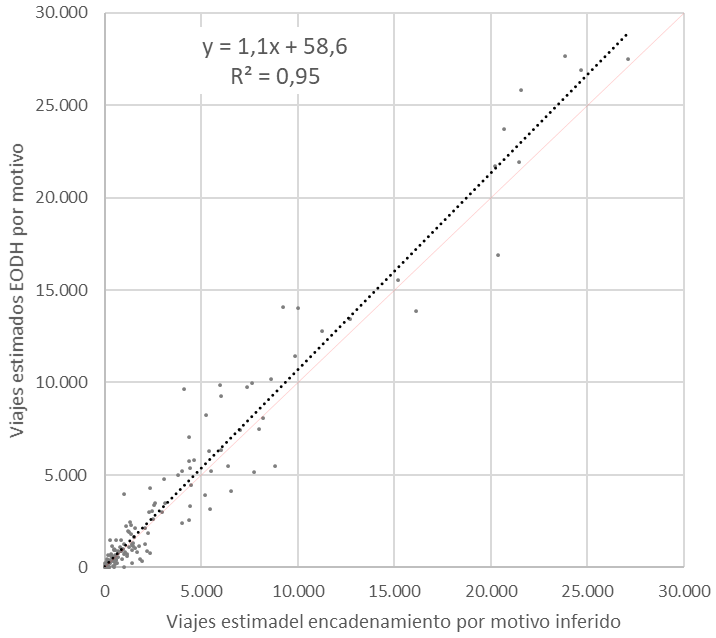
Tabla 5-2 Comparación de pares OD a nivel de línea por motivo volver a casa

Parámetro	EODH/TM
Z*	-1,30
R²	0,95
Pendiente	1,3
Intercepto	-115,1

Fuente: elaboración propia.

La Figura 5-9 muestra los resultados de correlación entre los viajes de la EODH con motivo trabajar, con respecto a los viajes de la matriz estimada por el método de encadenamiento con motivo de viaje trabajar.

Figura 5-9 Ajuste motivo de viaje trabajar



Fuente: elaboración propia.

De igual forma que para el proceso de comparación anterior, se aplica la prueba de hipótesis de diferencia de medias, donde se encuentra un valor de Z* menor al 1,96 por lo que se acepta la hipótesis nula. Lo que implica que no existe diferencia entre las medias de los viajes de la EODH por motivo trabajar y los viajes de la matriz estimada por el método de encadenamiento con motivo de viaje trabajar a nivel de línea; por otra parte,

existen valores altos de correlación con un valor de 0,95, similares a los obtenidos en la sección 3.4.3 Comparación OD a nivel de líneas del sistema, con una pendiente cercada a uno de 1,1 y un intercepto pequeño de 58,6.

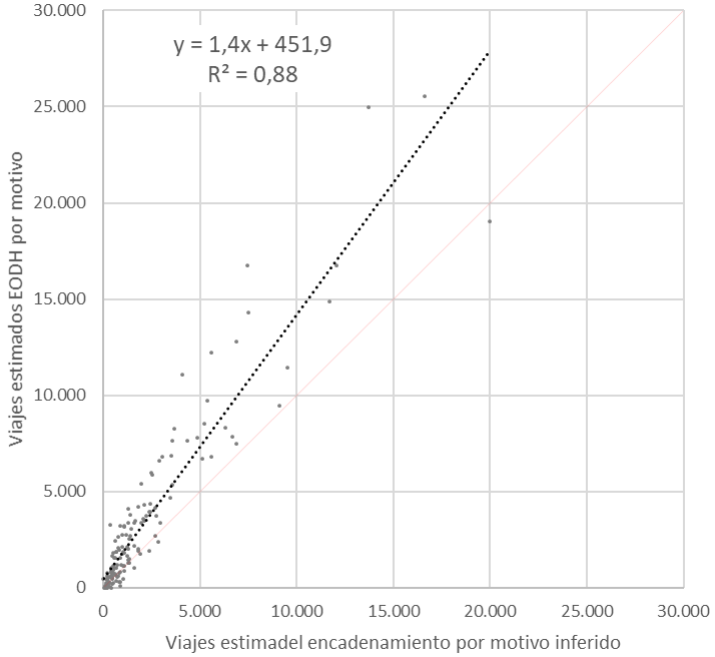
Tabla 5-3 Comparación de pares OD a nivel de línea por motivo trabajar

Parámetro	EODH/TM
Z*	-0,43
R²	0,95
Pendiente	1,1
Intercepto	58,6

Fuente: elaboración propia.

Por otra parte, al construir la recta de regresión lineal para evaluar la similaridad de las matrices de la EODH con motivo de viaje otro, con respecto a la matriz estimada por el método de encadenamiento con motivo de viaje otro, se observa en la Figura 5-10 que existe una relación lineal.

Figura 5-10 Ajuste motivo de viaje otros



Fuente: elaboración propia.

Para la comparación de los viajes estimados por motivo de viaje otros, por los dos métodos, se reporta un nivel de correlación del 0,88, siendo menor que para los dos casos anteriores, la pendiente es de 1,4 cercana a uno y el intercepto es de 451,9. Como se observa en la Tabla 5-4, al evaluar la diferencia estadística de medias se encuentra que el valor de Z* es mayor a $Z_{\alpha/2}$ para un nivel de confianza del 95%, lo que implica que se rechaza la hipótesis nula, es decir que existe diferencia estadística entre las medias de las estimaciones.

Tabla 5-4 Comparación de pares OD a nivel de línea por motivo otros

Parámetro	EODH/TM
Z*	-2,75
R²	0,88
Pendiente	1,4
Intercepto	451,9

Fuente: elaboración propia.

5.4 Discusión y conclusiones

Sobre los más de 9 millones de pares OD que pudieron ser encadenados con el método de encadenamiento de viajes de la sección 3.2 Heurísticas del modelo de encadenamiento, se completa la información de las variables hora de inicio del viaje, duración del viaje y duración de la actividad, para estimar la inferencia de motivo de viaje para los días hábil, sábado y domingo de la semana del primero (1) al siete (7) de octubre de 2018 y obtener las matrices por motivo de viaje, que son contrastadas a nivel de líneas con las matrices por motivo de viaje de un día hábil de la EODH.

A pesar de que solamente el 30% de los pares origen destino pueden ser completados haciendo uso de solo una ruta, como se presentó en la sección 2.2.4 Matriz de duración del viaje de TransMilenio, este 30% de las rutas brinda cobertura al 70% de los viajes, como resultado de que el diseño de rutas del sistema busca atender los pares OD con mayor demanda.

Al comparar de forma gráfica el comportamiento horario y la distribución espacial de la demanda de las matrices de la EODH por motivo de viaje, frente al comportamiento de las matrices estimadas por el método de encadenamiento con inferencia del motivo de viaje, se observan similitudes que, si bien no obedecen a un contraste estadístico, muestran tendencias en los datos muy similares, por ejemplo, concentración de viajes en portales

para el caso de los orígenes del motivo trabajar y los destinos del motivo volver a casa, y con la concentración de la hora de inicio del viaje en horas de la mañana para el motivo trabajar y en horas de la tarde para el motivo volver a casa; con el aporte adicional de tener información por motivo de viaje para los días sábado y domingo.

Por otra parte, al comparar por medio de una prueba de hipótesis de diferencia de medias, los pares OD a nivel de línea de las matrices de la EODH, frente a las matrices estimadas por el método de encadenamiento con inferencia del motivo de viaje, se acepta la hipótesis nula para el caso de el motivo volver a casa y trabajar, lo que significa que las medias son iguales, diferente a lo que ocurrió cuando se aplicó la prueba de hipótesis al total de datos por línea de la sección 3.4.3 Comparación OD a nivel de líneas del sistema, donde se rechazó la hipótesis nula, lo que sugiere que la segmentación mejora la comparación entre las medias.

Para el caso de la comparación de los datos, haciendo uso de la recta de regresión lineal, se observa un buen ajuste para los motivos volver a casa, trabajar y otro, ya que se obtiene un valor de R^2 superiores a 0,85, con pendientes cercanas a uno e interceptos pequeños; si bien se observa un buen ajuste para estos motivos, el modelo de inferencia de motivo de viaje seleccionado (Tabla 4-21), no muestra clasificación para el motivo estudiar, a pesar de presentar significancia en los parámetros del modelo, lo cual sugiere la necesidad de incluir otras variables, con el fin de obtener respuesta del motivo de viaje estudiar.

6. Conclusiones y recomendaciones

6.1 Conclusiones

Sobre los más de 13 millones de datos de entradas al sistema troncal del SITP (TransMilenio) que se dieron en la semana típica del primero (1) al siete (7) de octubre de 2018, fue posible obtener por medio de la aplicación de heurísticas y del método de encadenamiento de viajes, pares OD para más de 9 millones de viajes, los cuales fueron complementados aplicando los métodos disponibles para la inferencia del motivo del viaje, que resultó de la aplicación de un modelo logit multinomial, de la selección de más de 194 modelos evaluados por medio de matrices de confusión, que involucraron las variables temporales y espaciales resultantes del procesamiento y minería de datos, que fueron contrastadas con las matrices obtenidas por motivo de viaje de la EODH.

Del procesamiento y minería de datos, se obtuvo que las variables temporales, como la duración de la actividad, presentan comportamientos asociados a cada motivo del viaje, donde los viajes con motivo trabajar tienen una media de duraciones de 8,8 horas y el motivo estudiar duraciones de 6,7 horas; a su vez, las variables espaciales asociadas al tipo de uso del suelo, presentan mayor concentración según la ubicación de las estaciones, como ocurre para el uso de suelo de educación que se concentra en estaciones como la Universidad Nacional, Las Aguas y la Calle 45.

Por otra parte, existe una distribución de la hora promedio de entrada a TransMilenio, según la localización de las estaciones y portales con respecto al centro de la ciudad, donde las más alejadas tienen una hora promedio de inicio del viaje menor. Es así como se concluye, que el procesamiento y minería de datos, permitió evidenciar que los datos de diferentes fuentes de información mantienen una consistencia, y que el comportamiento espacial y temporal, se puede asociar a diferencias en los motivos del viaje.

La aplicación del método de encadenamiento de viajes para la obtención de matrices OD a partir de los datos del uso de tarjetas inteligentes en los sistema AFC, presentan una ventaje sobre el método de la EODH, ya que permite la reconstrucción de cerca de 1,6 millones de pares OD con una representación para todas las estaciones y portales del sistema, con casi 100 veces más de la cantidad de viajes que fueron encuestados en la EODH, donde se reportaron 16.589 viajes del medio principal TransMilenio.

La comparación entre las matrices obtenidas de la EODH, frente a la obtenida por el método de encadenamiento de viaje, presentaron buenas correlaciones a nivel de pares OD a nivel de líneas, con valores de R^2 del 95%, pendiente cercana a 1 e intercepto cercano a cero; no obstante, la correlación a nivel pares OD de estación y/o portales fue baja debido a que los datos de la EODH obedecen a resultados de diseño muestral a nivel de ZAT, resultando en que el número de pares contenidos en la matriz estimada de la EODH sea inferior a la matriz estimada por el método de encadenamiento de viajes, lo que impide realizar una comparación a nivel de pares OD.

A partir de los datos de las encuestas en hogares y otras fuentes, se obtuvieron variables temporales y espaciales de los viajes, asociadas al motivo de viaje del componente troncal del SITP (TransMilenio) de Bogotá, D.C., que permitieron estimar 194 modelos de tipo logit multinomial, utilizando diferentes combinaciones de variables. Se determinó por medio de una matriz de confusión, que el mejor modelo es capaz de estimar aciertos en la clasificación de más de 77% de los datos de evaluación, para cuatro (4) motivos de viaje (volver a casa, trabajar, estudiar y otro), empleando tres (3) variables independientes de tipo temporal (inicio del viaje, duración del viaje y duración de la actividad); demostrándose que la variable duración de la actividad tiene el mayor potencial en la clasificación de motivo del viaje.

Al comparar los pares OD a nivel de línea de las matrices de la EODH por motivo de viaje, frente a las matrices estimadas por el método de encadenamiento de viajes con inferencia del motivo de viaje, mediante una prueba de hipótesis de diferencia de medias para los motivos volver a casa y trabajar, se acepta la hipótesis nula, sugiriendo que la segmentación mejora la comparación entre las medias para los pares OD a nivel de líneas.

La comparación entre las MOD estimadas por los dos métodos reporta porcentajes de correlación superiores al 85% con pendientes cercanas a uno e interceptos pequeños; no obstante, no se obtiene clasificación para el motivo estudiar, a pesar de presentar significancia en los parámetros del modelo logit multinomial seleccionado.

6.2 Recomendaciones y futuras investigaciones

El presente trabajo ha mostrado que los modelos para la inferencia del motivo de viaje requieren del procesamiento y minería de datos, en este sentido, futuras investigaciones podrían probar la inclusión de otras variables que mejoren la estimación de los motivos de viaje, especialmente el motivo estudio. Estas variables adicionales, podrían considerar cupos educacionales, empleos ofertados, tasas de motorización alrededor de estaciones, entre otras.

A su vez, se podría ampliar el alcance en la estimación de matrices por motivo de viaje para el resto del sistema de transporte público de pasajeros, considerando que la ciudad de Bogotá D.C., cuenta con el Sistema Integrado de Transporte Público, que usa un único medio de pago. Adicionalmente, se podría aplicar los métodos usados en los sistemas tipo BRT de otras ciudades en Colombia e incluso en sistemas de transporte público que empleen tarjetas inteligentes para ingreso en estaciones como los sistemas tipo metro.

Por otra parte, se podría hacer un análisis de sensibilidad sobre el tamaño de las áreas de influencia alrededor de estaciones, sobre las cuales se construyeron las variables de uso del suelo, aplicando distancias tipo Manhattan, y variando el área de cobertura según el tipo de estación, involucrando modos adicionales de acceso.

También se podrían explorar otros tipos de modelos que puedan permitir obtener el motivo del viaje, como modelos de tipo árbol de decisión o modelos de redes neuronales.

6.3 Principal contribución a nuevo conocimiento

La principal contribución a nuevo conocimiento se da con la posibilidad de estimar información adicional sobre el uso de tarjetas inteligentes del sistema TransMilenio, asociada al motivo del viaje aplicando modelos de tipo logit multinomial, que permiten

obtener matrices origen destino por motivo de viaje a partir de los datos de uso de las tarjetas inteligentes.

6.4 Limitaciones de la investigación

Una de las limitaciones del presente trabajo, es que compara dos métodos de estimación de MOD que usan diferentes fuentes de datos, los dos métodos son estimaciones aproximadas. Por otra parte, la temporalidad de la información de las encuestas origen destino en hogares, y los datos del uso de tarjetas inteligentes, al no ser tomados al mismo tiempo, puede generar algunas distorsiones en las magnitudes de los datos cuando son comparadas; no obstante, esto no ha afectado el alcance del presente trabajo ya que se han podido obtener matrices de viaje por motivo, a partir de las tarjetas inteligentes del sistema automático de recaudo.

7. Bibliografía

Alcaldia Mayor de Bogotá (2013) *Capitulo de la infraestructura para los escenarios del sistema integrado de transporte público - SITP*. Bogotá, D.C.

Alsger, A. *et al.* (2016) 'Validating and improving public transport origin-destination estimation algorithm using smart card fare data', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 68, pp. 490–506. doi: 10.1016/j.trc.2016.05.004.

Alsger, A. A. M. (2017) 'Estimation of transit origin destination matrices using smart card fare data', *Ph. D. University of Queensland*.

Banco Interamericano de Desarrollo (2018) *Disrupción exponencial en la economía digital*. Washington, D.C. doi: 10.18235/0001068.

Banco Mundial (2009) *Una nueva geografía económica*, *Informe sobre el Desarrollo*. Available at: http://siteresources.worldbank.org/INTWDR2009/Resources/WDR_OVERVIEW_ES_Web.pdf.

Barry, J., Freimer, R. and Slavin, H. (2009) 'Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City', *Transportation Research Record: Journal of the Transportation Research Board*, 2112, pp. 53–61. doi: 10.3141/2112-07.

Chen, Z. and Fan, W. (2018) 'Extracting bus transit boarding stop information using smart card transaction data', *Journal of Modern Transportation*. Springer Berlin Heidelberg, 26(3), pp. 209–219. doi: 10.1007/s40534-018-0165-y.

CONPES - 3167 (2002) 'Política para mejorar el servicio de transporte público urbano de pasajeros'.

Dou, H.; Liu, H.; Yang, X. (2007) 'OD Matrix Estimation Method of Public Transportation Flow Based on Passenger Boarding and Alighting. Comput. Commun.', 25, 79–8.

Greene, W. (2003) *Econometric analysis*. 5th edn. Edited by New Jersey: Prentice Hall. 2003.

Guío Burgos, F. A. (2009) *Caracterización y modelación de flujos peatonales en infraestructuras continuas - caso estudio Tunja - Colombia*.

Gujarati, D. N. and Porter, D. C. (2010) *Econometría*. MCGRAW-HIL. Edited by MCGRAW-HILL. Available at: https://www.casadellibro.com/libro-econometria/9786071502940/1767935.

Jung, J. and Sohn, K. (2017) 'Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data', *IET Intelligent Transport Systems*, 11(6), pp. 334–339. doi: 10.1049/iet-its.2016.0276.

Kumar Molugaram Rao, G. S. R. (2017) *Statistical Techniques for Transportation Engineering*. doi: 10.1016/S0925-4005(03)00443-X.

Li, D. *et al.* (2011) 'Estimating a Transit Passenger Trip Origin-Destination Matrix Using Automatic Fare Collection System', in. Springer Berlin Heidelberg (Lecture Notes in Computer Science), pp. 502–513. doi: 10.1007/978-3-642-20244-5_48.

Li, T. *et al.* (2018) 'Smart card data mining of public transport destination: A literature review', *Information (Switzerland)*, 9(1), pp. 28–30. doi: 10.3390/info9010018.

Microsoft (2019) *Conceptos de minería de datos*. Available at: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017.

Munizaga, M. A. and 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*. Pergamon, 24, pp. 9–18. doi: 10.1016/J.TRC.2012.01.007.

Nassir, N. *et al.* (2011) 'Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System', *Transportation Research Record: Journal of the Transportation Research Board*, 2263, pp. 140–150. doi: 10.3141/2263-16.

Nunes, A. A., Galvao Dias, T. and Falcao e Cunha, J. (2016) 'Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation', *IEEE Transactions on Intelligent Transportation Systems*, 17(1), pp. 133–142. doi: 10.1109/TITS.2015.2464335.

Ortúzar, J. de D. (2015) *Modelos de demanda de transporte*. (c) Edicio, *Alfaomega, Ediciones UC, Pontificia Universidad Católica de Chile*. (c) Edicio. Available at: http://www.alfaomega.com.co/modelos-de-demanda-de-transporte-5714.html.

Rus, G. de, Campos, J. and Nombela, G. (2002) *Economía del transporte*. Edited by Antoni Bosch.

SAPORITI, E. M. (2016) *New Horizons for a Data-Driven Economy*. Springer O, *Revista de la Asociacion Medica Argentina*. Springer O. Edited by J. M. Cavanillas, E. Curry, and W. Wahlster. Cham: Springer International Publishing. doi: 10.1007/978-3-319-21569-3.

Secop II (2018) *Realizar la Encuesta de Movilidad, que comprende la Encuesta Origen-Destino de Hogares (EODH) y la Encuesta Origen-Destino de Interceptación (EODI) para Bogotá y los municipios vecinos de su área de influencia, y la actualización del modelo de transporte*.

TransMilenio S.A. (2018) *TransMilenio en cifras - Agosto 2018*, *Web TransMilenio*.

Available at: https://www.transmilenio.gov.co/publicaciones/151075/estadisticas-de-oferta-y-demanda-bimensual-del-sistema-integrado-de-transporte-publico---sitp---noviembre---diciembre---2018/.

TransMilenio S.A. (2019) *Respuesta radicado Transmilenio S.A., No 2018ER37643*.

Unidad Administrativa Especial de Catastro Distrital (2019) *Infraestructura de Datos Espaciales para el Distrito Capital*, *https://www.ideca.gov.co/*. Available at: https://www.ideca.gov.co/.

Wang, W., Attanucci, J. and Wilson, N. (2011) 'Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems', *Journal of Public Transportation*, 14(4), pp. 131–150. doi: 10.5038/2375-0901.14.4.7.

Willumsen, L. G. (2014) *Better traffic and revenue forecasting*. doi: 978-0992843304.

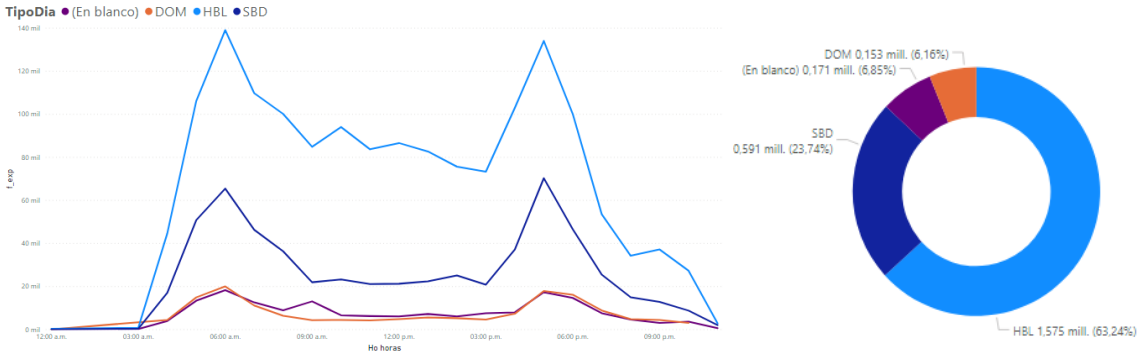
Yu, J. and Yang, X. (2006) 'Estimating a Transit Route OD Matrix from On-Off Data through an Artificial Neural Network Method', in *Applications of Advanced Technology in Transportation*. Reston, VA: American Society of Civil Engineers, pp. 467–472. doi: 10.1061/40799(213)74.

Zhang, M.; Guo, Y.; Ma, Y. A. (2014) 'Probability Model of Transit OD Distribution Based on the Allure of Bus Station.', *J. Transp. Inf. Saf.*, 32, 57–6.

Zhao, J., Rahbee, A. and Wilson, N. H. M. (2007) 'Estimating a rail passenger trip origin-destination matrix using automatic data collection systems', *Computer-Aided Civil and Infrastructure Engineering*, 22(5), pp. 376–387. doi: 10.1111/j.1467-8667.2007.00494.x.

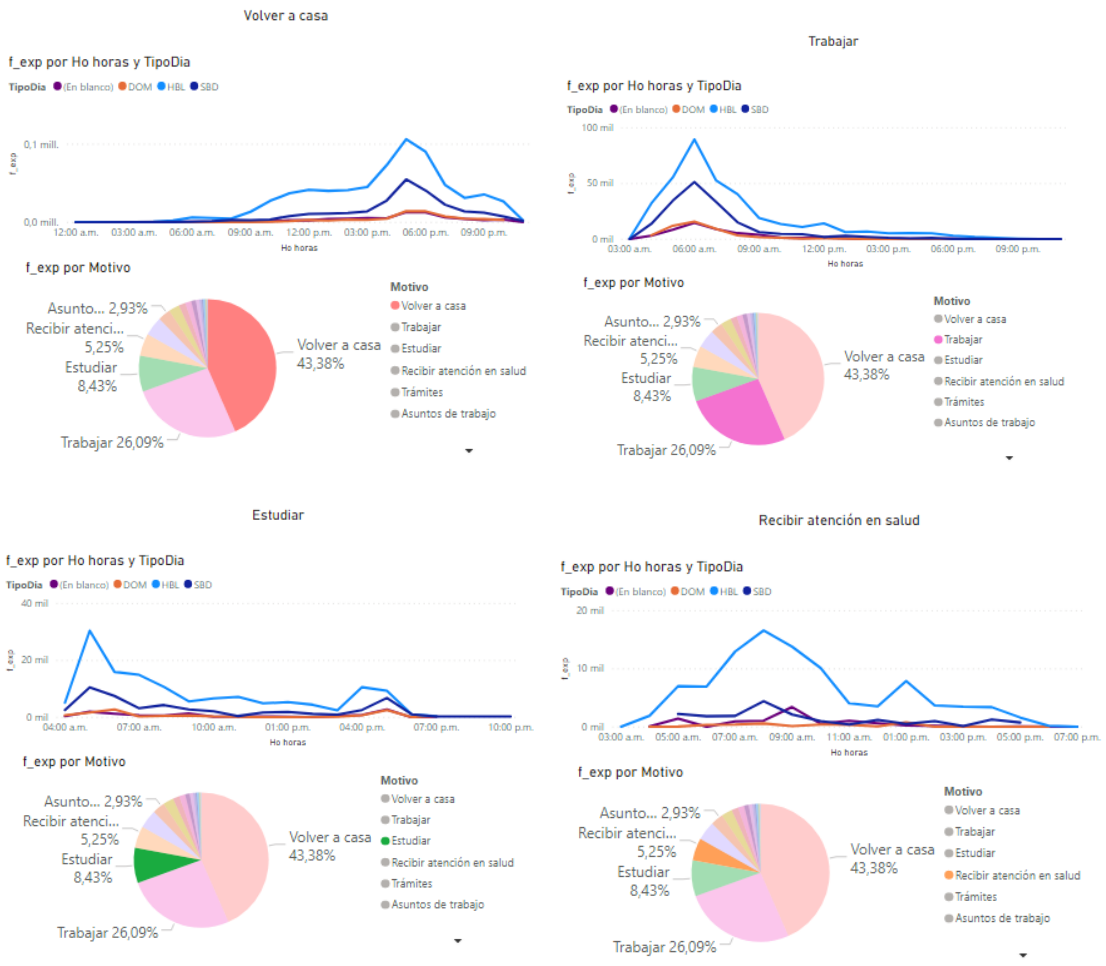
A. Anexo: tipo día viajes de la EODH

Al tomar la base de datos de viajes, se identifica que el atributo fecha incluye sábados y domingos, lo que permite obtener el siguiente gráfico diferenciando los viajes por tipo día. Al realizar esta diferenciación, se tienen 1,6 millones de viajes para los días hábiles, 0,6 millones de viajes para un sábado, 0,15 millones viajes para el domingo y 0,17 millones de viajes sin información de la fecha del viaje. La distribución horaria para los viajes se presenta en el siguiente histograma para los diferentes tipos día, donde el comportamiento es uniformes para los tipos día.



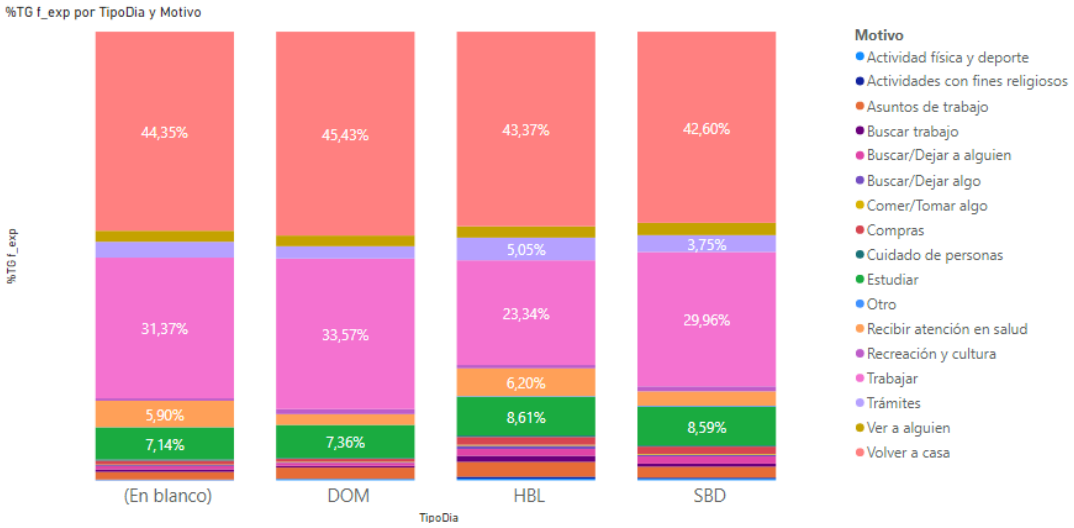
Fuente: elaboración propia a partir de la Encuesta Origen Destino en Hogares 2019

Por otra parte, al considerar el motivo de viaje y comparar los histogramas por tipo día, de las siguientes figuras no se ven cambios significativos en el comportamiento de los viajes.



Fuente: elaboración propia a partir de la Encuesta Origen Destino en Hogares 2019

Adicionalmente, se observa en la siguiente figura, que para los días hábiles el motivo trabajar representa el 23,34%, en proporción al total de viajes que se realizan en el día hábil, es inferior que para un día sábado (29,96%) o un día domingo (33,57%).



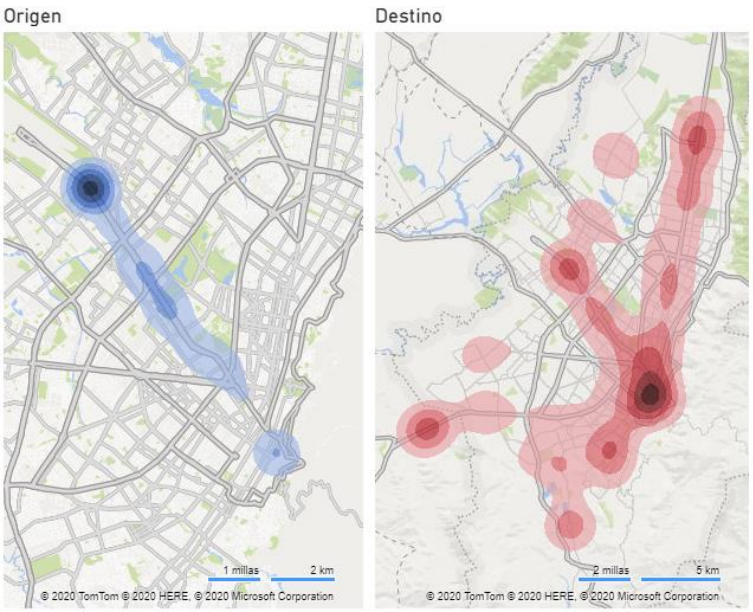
Fuente: elaboración propia a partir de la Encuesta Origen Destino en Hogares 2019

Lo anterior induce a creer que el atributo de fecha del viaje que se encuentra en los formularios de viaje no está dando información confiable sobre este atributo, esto sumado al hecho de que los reportes de los informes generales hablan de la demanda típica de un día hábil, lo que permite concluir que toda la información de la EODH se asociará a la demanda típica de un día hábil.

B. Anexo: Patrones de viajes encadenados por línea

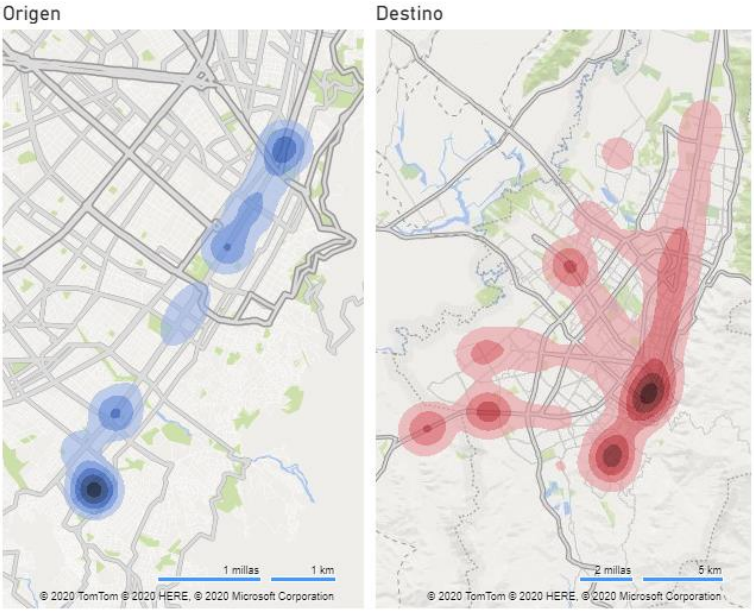
La siguiente tabla presenta la matriz de viajes a nivel de línea de la calle 26 obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	Matriz encadenada
(11) Zona K Calle 26	(12) Zona L Carrera 10	9,702
(11) Zona K Calle 26	(39) Zona F Calle 13	943
(11) Zona K Calle 26	(37) Zona J Eje Ambiental	3,546
(11) Zona K Calle 26	(33) Zona B AutoNorte	19,648
(11) Zona K Calle 26	(32) Zona C Av. Suba	5,250
(11) Zona K Calle 26	(35) Zona D Calle 80	4,317
(11) Zona K Calle 26	(31) Zona F Av. AmÃ©ricas	3,958
(11) Zona K Calle 26	(34) Zona H Caracas Sur	14,844
(11) Zona K Calle 26	(11) Zona K Calle 26	19,531
(11) Zona K Calle 26	(30) Zona G NQS Sur	14,660
(11) Zona K Calle 26	(38) Zona E NQS Central	3,902
(11) Zona K Calle 26	(36) Zona A Caracas	12,221



La siguiente tabla presenta la matriz de viajes a nivel de línea de la carrera 10 obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

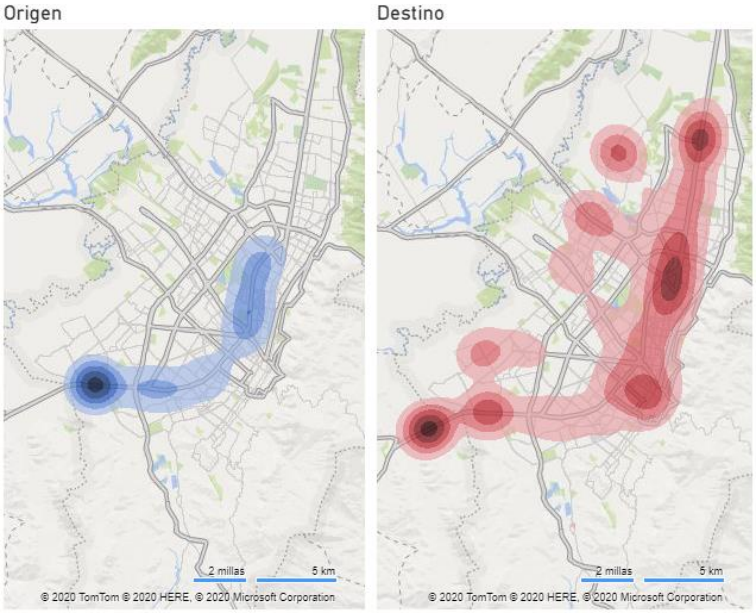
Lin_Origen	Lin_Destino	TM
(12) Zona L Carrera 10	(12) Zona L Carrera 10	16,120
(12) Zona L Carrera 10	(39) Zona F Calle 13	1,138
(12) Zona L Carrera 10	(37) Zona J Eje Ambiental	385
(12) Zona L Carrera 10	(33) Zona B AutoNorte	11,197
(12) Zona L Carrera 10	(32) Zona C Av. Suba	2,996
(12) Zona L Carrera 10	(35) Zona D Calle 80	3,411
(12) Zona L Carrera 10	(31) Zona F Av. AmÃ©ricas	7,204
(12) Zona L Carrera 10	(11) Zona K Calle 26	9,591
(12) Zona L Carrera 10	(30) Zona G NQS Sur	12,343
(12) Zona L Carrera 10	(38) Zona E NQS Central	1,508
(12) Zona L Carrera 10	(34) Zona H Caracas Sur	3,125
(12) Zona L Carrera 10	(36) Zona A Caracas	8,456



La siguiente tabla presenta la matriz de viajes a nivel de línea de la NQS Sur y NQS Central, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

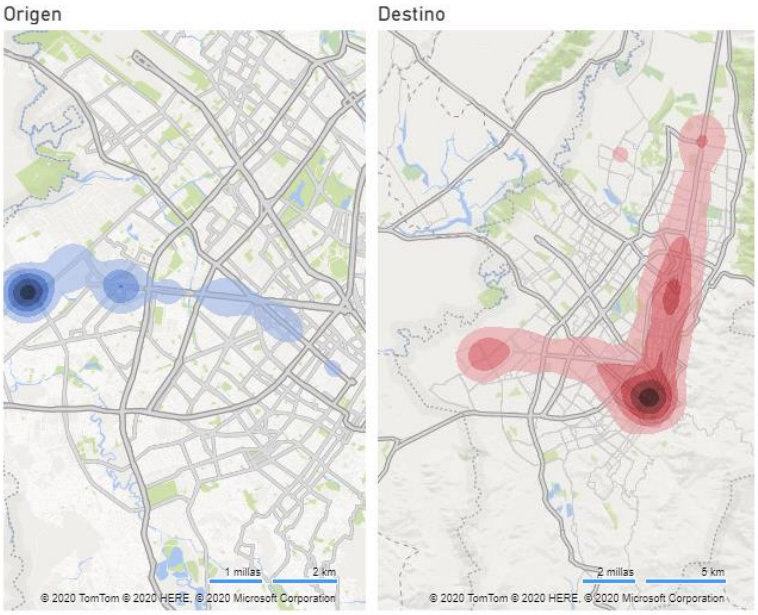
Lin_Origen	**Lin_Destino**	**TM**
(30) Zona G NQS Sur	(12) Zona L Carrera 10	11,753
(30) Zona G NQS Sur	(39) Zona F Calle 13	3,364
(30) Zona G NQS Sur	(37) Zona J Eje Ambiental	2,111
(30) Zona G NQS Sur	(33) Zona B AutoNorte	39,013
(30) Zona G NQS Sur	(32) Zona C Av. Suba	12,768
(30) Zona G NQS Sur	(35) Zona D Calle 80	9,974
(30) Zona G NQS Sur	(31) Zona F Av. AmÃ©ricas	2,678
(30) Zona G NQS Sur	(11) Zona K Calle 26	14,911
(30) Zona G NQS Sur	(30) Zona G NQS Sur	31,662
(30) Zona G NQS Sur	(38) Zona E NQS Central	27,672

(30) Zona G NQS Sur	(34) Zona H Caracas Sur	1,821
(30) Zona G NQS Sur	(36) Zona A Caracas	18,617
(38) Zona E NQS Central	(12) Zona L Carrera 10	1,411
(38) Zona E NQS Central	(39) Zona F Calle 13	680
(38) Zona E NQS Central	(37) Zona J Eje Ambiental	443
(38) Zona E NQS Central	(33) Zona B AutoNorte	18,096
(38) Zona E NQS Central	(32) Zona C Av. Suba	10,526
(38) Zona E NQS Central	(35) Zona D Calle 80	8,664
(38) Zona E NQS Central	(31) Zona F Av. AmÃ©ricas	10,625
(38) Zona E NQS Central	(11) Zona K Calle 26	3,698
(38) Zona E NQS Central	(30) Zona G NQS Sur	28,456
(38) Zona E NQS Central	(38) Zona E NQS Central	3,083
(38) Zona E NQS Central	(34) Zona H Caracas Sur	5,530
(38) Zona E NQS Central	(36) Zona A Caracas	2,614



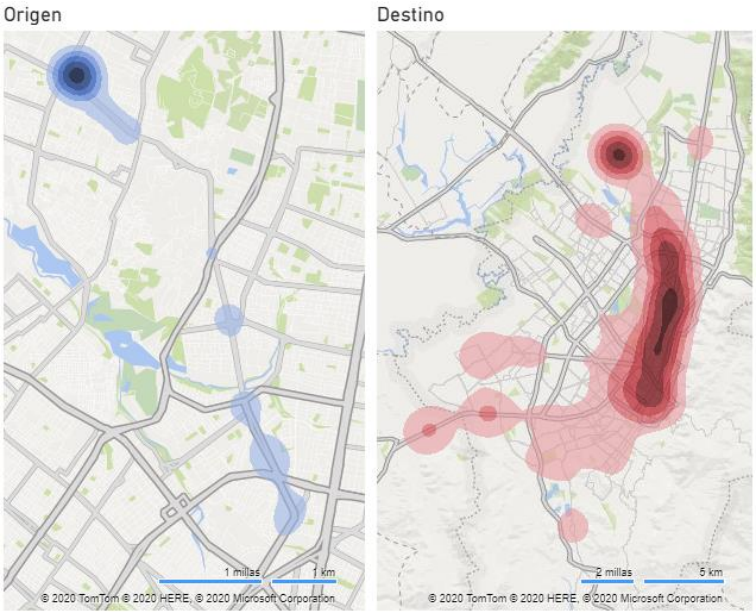
La siguiente tabla presenta la matriz de viajes a nivel de línea de la Américas, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	TM
(31) Zona F Av. AmÃ©ricas	(12) Zona L Carrera 10	7,241
(31) Zona F Av. AmÃ©ricas	(39) Zona F Calle 13	12,140
(31) Zona F Av. AmÃ©ricas	(37) Zona J Eje Ambiental	9,335
(31) Zona F Av. AmÃ©ricas	(33) Zona B AutoNorte	28,722
(31) Zona F Av. AmÃ©ricas	(32) Zona C Av. Suba	6,573
(31) Zona F Av. AmÃ©ricas	(35) Zona D Calle 80	4,163
(31) Zona F Av. AmÃ©ricas	(31) Zona F Av. AmÃ©ricas	18,623
(31) Zona F Av. AmÃ©ricas	(34) Zona H Caracas Sur	3,576
(31) Zona F Av. AmÃ©ricas	(11) Zona K Calle 26	4,191
(31) Zona F Av. AmÃ©ricas	(30) Zona G NQS Sur	2,578
(31) Zona F Av. AmÃ©ricas	(38) Zona E NQS Central	10,674
(31) Zona F Av. AmÃ©ricas	(36) Zona A Caracas	34,851



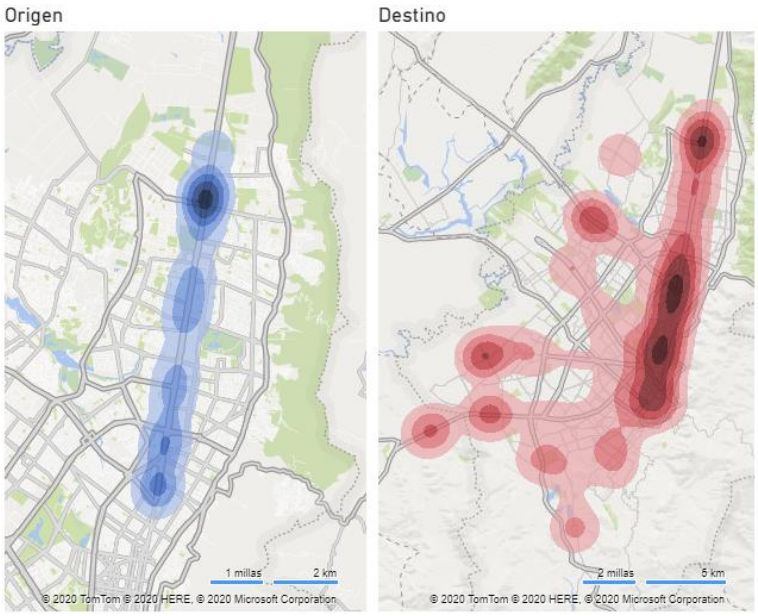
La siguiente tabla presenta la matriz de viajes a nivel de línea de la Avenidas Suba, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	TM
(32) Zona C Av. Suba	(12) Zona L Carrera 10	3,189
(32) Zona C Av. Suba	(39) Zona F Calle 13	1,654
(32) Zona C Av. Suba	(37) Zona J Eje Ambiental	2,938
(32) Zona C Av. Suba	(33) Zona B AutoNorte	8,447
(32) Zona C Av. Suba	(32) Zona C Av. Suba	25,270
(32) Zona C Av. Suba	(35) Zona D Calle 80	3,959
(32) Zona C Av. Suba	(31) Zona F Av. AmÃ©ricas	6,597
(32) Zona C Av. Suba	(34) Zona H Caracas Sur	7,828
(32) Zona C Av. Suba	(11) Zona K Calle 26	5,301
(32) Zona C Av. Suba	(30) Zona G NQS Sur	12,751
(32) Zona C Av. Suba	(38) Zona E NQS Central	10,389
(32) Zona C Av. Suba	(36) Zona A Caracas	26,785



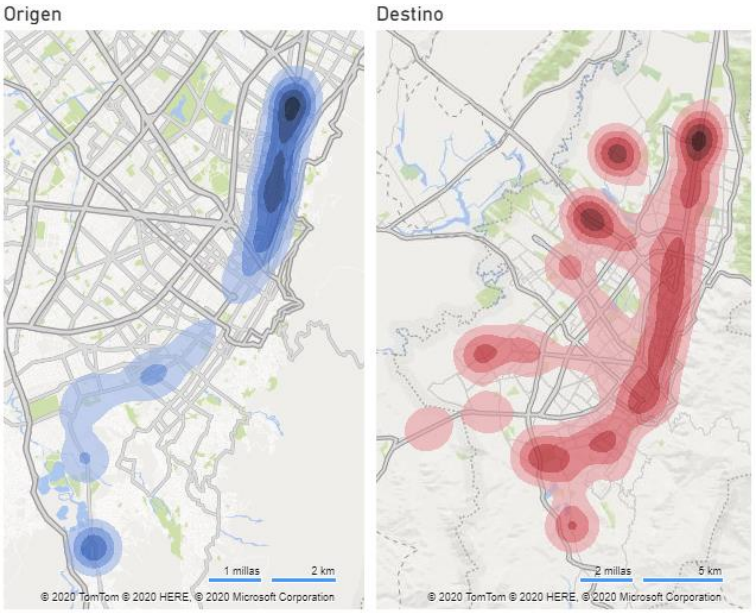
La siguiente tabla presenta la matriz de viajes a nivel de línea de la Autopista Norte, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	TM
(33) Zona B AutoNorte	(12) Zona L Carrera 10	11,750
(33) Zona B AutoNorte	(39) Zona F Calle 13	2,722
(33) Zona B AutoNorte	(37) Zona J Eje Ambiental	7,237
(33) Zona B AutoNorte	(33) Zona B AutoNorte	70,463
(33) Zona B AutoNorte	(32) Zona C Av. Suba	8,273
(33) Zona B AutoNorte	(35) Zona D Calle 80	23,097
(33) Zona B AutoNorte	(31) Zona F Av. AmÃ©ricas	28,525
(33) Zona B AutoNorte	(34) Zona H Caracas Sur	30,037
(33) Zona B AutoNorte	(11) Zona K Calle 26	20,193
(33) Zona B AutoNorte	(30) Zona G NQS Sur	38,358
(33) Zona B AutoNorte	(38) Zona E NQS Central	17,967
(33) Zona B AutoNorte	(36) Zona A Caracas	51,865



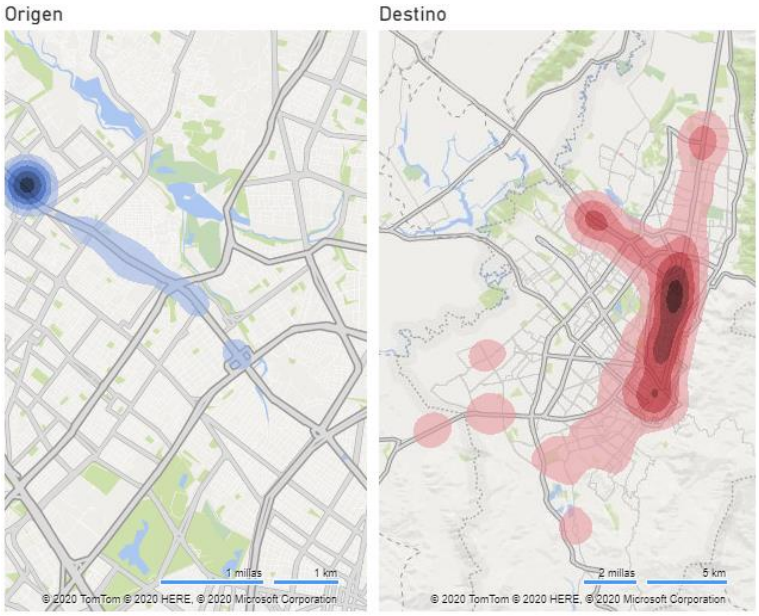
La siguiente tabla presenta la matriz de viajes a nivel de línea de la Caracas Sur y centro, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	TM
(34) Zona H Caracas Sur	(12) Zona L Carrera 10	3,048
(34) Zona H Caracas Sur	(39) Zona F Calle 13	2,443
(34) Zona H Caracas Sur	(37) Zona J Eje Ambiental	1,414
(34) Zona H Caracas Sur	(33) Zona B AutoNorte	30,379
(34) Zona H Caracas Sur	(32) Zona C Av. Suba	7,851
(34) Zona H Caracas Sur	(35) Zona D Calle 80	9,369
(34) Zona H Caracas Sur	(31) Zona F Av. AmÃ©ricas	3,618
(34) Zona H Caracas Sur	(34) Zona H Caracas Sur	33,379
(34) Zona H Caracas Sur	(11) Zona K Calle 26	14,955
(34) Zona H Caracas Sur	(30) Zona G NQS Sur	1,716
(34) Zona H Caracas Sur	(38) Zona E NQS Central	5,647
(34) Zona H Caracas Sur	(36) Zona A Caracas	39,965
(36) Zona A Caracas	(12) Zona L Carrera 10	8,182
(36) Zona A Caracas	(39) Zona F Calle 13	2,616
(36) Zona A Caracas	(37) Zona J Eje Ambiental	1,807
(36) Zona A Caracas	(33) Zona B AutoNorte	51,462
(36) Zona A Caracas	(32) Zona C Av. Suba	27,180
(36) Zona A Caracas	(35) Zona D Calle 80	30,176
(36) Zona A Caracas	(31) Zona F Av. AmÃ©ricas	34,870
(36) Zona A Caracas	(34) Zona H Caracas Sur	40,416
(36) Zona A Caracas	(11) Zona K Calle 26	11,572
(36) Zona A Caracas	(30) Zona G NQS Sur	18,419
(36) Zona A Caracas	(38) Zona E NQS Central	3,190
(36) Zona A Caracas	(36) Zona A Caracas	16,169



La siguiente tabla presenta la matriz de viajes a nivel de línea de la Calle 80, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

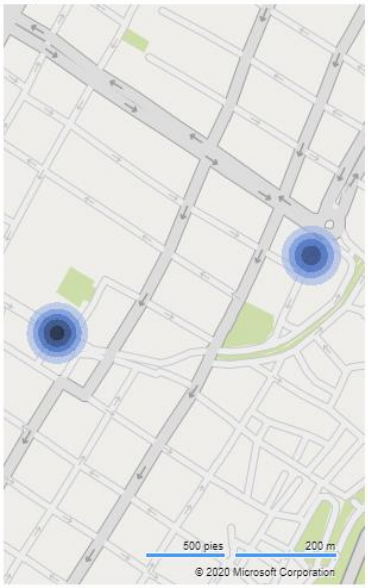
Lin_Origen	Lin_Destino	TM
(35) Zona D Calle 80	(12) Zona L Carrera 10	3,576
(35) Zona D Calle 80	(39) Zona F Calle 13	1,234
(35) Zona D Calle 80	(37) Zona J Eje Ambiental	3,148
(35) Zona D Calle 80	(33) Zona B AutoNorte	23,424
(35) Zona D Calle 80	(32) Zona C Av. Suba	3,900
(35) Zona D Calle 80	(35) Zona D Calle 80	19,829
(35) Zona D Calle 80	(31) Zona F Av. AmÃ©ricas	4,147
(35) Zona D Calle 80	(34) Zona H Caracas Sur	9,359
(35) Zona D Calle 80	(11) Zona K Calle 26	4,306
(35) Zona D Calle 80	(30) Zona G NQS Sur	9,913
(35) Zona D Calle 80	(38) Zona E NQS Central	8,527
(35) Zona D Calle 80	(36) Zona A Caracas	30,020



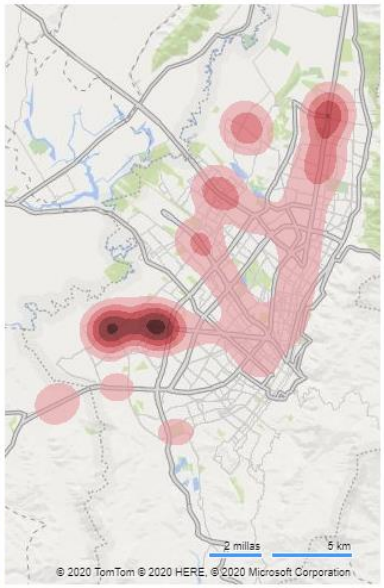
La siguiente tabla presenta la matriz de viajes a nivel de línea del Eje Ambiental, obtenida del proceso de encadenamiento de viajes, junto con su distribución espacial.

Lin_Origen	Lin_Destino	TM
(37) Zona J Eje Ambiental	(12) Zona L Carrera 10	265
(37) Zona J Eje Ambiental	(39) Zona F Calle 13	513
(37) Zona J Eje Ambiental	(37) Zona J Eje Ambiental	69
(37) Zona J Eje Ambiental	(33) Zona B AutoNorte	7,103
(37) Zona J Eje Ambiental	(32) Zona C Av. Suba	2,968
(37) Zona J Eje Ambiental	(35) Zona D Calle 80	3,210
(37) Zona J Eje Ambiental	(31) Zona F Av. AmÃ©ricas	9,604
(37) Zona J Eje Ambiental	(11) Zona K Calle 26	3,351
(37) Zona J Eje Ambiental	(30) Zona G NQS Sur	2,166
(37) Zona J Eje Ambiental	(38) Zona E NQS Central	547
(37) Zona J Eje Ambiental	(34) Zona H Caracas Sur	1,459
(37) Zona J Eje Ambiental	(36) Zona A Caracas	1,728

Origen



Destino



C. Anexo: Modelos construidos con variables espaciales

Se presentan los resultados obtenidos para los 52 modelos que pueden ser construidos con las seis variables espaciales.

Casa	Trabajo	Estudio	Comercio	Buscar trabajo	Otros	Núm.. Variables	Entrenamiento	Evaluación
Casa	Trabajo	Estudio	Comercio	BuscarTrabajo	Otros	6	45.28%	44.90%
Casa		Estudio	Comercio	BuscarTrabajo	Otros	5	45.25%	44.97%
Casa		Estudio		BuscarTrabajo	Otros	4	45.43%	44.86%
Casa		Estudio	Comercio	BuscarTrabajo		4	45.63%	44.95%
Casa			Comercio	BuscarTrabajo	Otros	4	45.50%	45.13%
		Estudio		BuscarTrabajo	Otros	3	45.97%	44.72%
Casa	Trabajo	Estudio	Comercio	BuscarTrabajo		5	45.70%	44.99%
	Trabajo	Estudio		BuscarTrabajo	Otros	4	46.00%	44.88%
Casa		Estudio		BuscarTrabajo		3	45.78%	45.11%
Casa				BuscarTrabajo	Otros	3	45.95%	44.97%
Casa	Trabajo	Estudio		BuscarTrabajo		4	45.84%	45.21%
Casa			Comercio	BuscarTrabajo		3	46.01%	45.05%
Casa				BuscarTrabajo		2	45.99%	45.19%
Casa	Trabajo			BuscarTrabajo		3	46.02%	45.19%
			Comercio	BuscarTrabajo	Otros	3	46.66%	45.38%
				BuscarTrabajo	Otros	2	46.69%	45.38%
	Trabajo	Estudio	Comercio	BuscarTrabajo		4	46.60%	45.71%
		Estudio	Comercio	BuscarTrabajo		3	46.70%	45.73%
	Trabajo	Estudio		BuscarTrabajo		3	46.97%	45.85%
		Estudio		BuscarTrabajo		2	46.97%	46.02%
			Comercio	BuscarTrabajo		2	47.32%	45.85%
	Trabajo			BuscarTrabajo		2	47.28%	45.94%
				BuscarTrabajo		1	47.26%	46.08%
Casa		Estudio	Comercio		Otros	4	48.61%	47.15%
Casa	Trabajo	Estudio			Otros	4	48.67%	47.19%
Casa	Trabajo	Estudio	Comercio		Otros	5	48.67%	47.19%
Casa			Comercio		Otros	3	48.79%	47.24%
	Trabajo	Estudio	Comercio		Otros	4	49.40%	47.75%
		Estudio	Comercio		Otros	3	49.39%	47.81%

Casa	Trabajo	Estudio	Comercio	Buscar trabajo	Otros	Núm.. Variables	Entrenamiento	Evaluación
			Comercio		Otros	2	49.51%	48.00%
Casa		Estudio	Comercio			3	49.26%	48.43%
Casa	Trabajo	Estudio	Comercio			4	49.22%	48.52%
Casa		Estudio			Otros	3	49.87%	48.70%
Casa	Trabajo		Comercio			3	49.45%	49.24%
Casa	Trabajo				Otros	3	49.73%	49.09%
Casa			Comercio			2	49.55%	49.34%
Casa					Otros	2	49.83%	49.18%
		Estudio			Otros	2	50.48%	49.18%
	Trabajo	Estudio	Comercio			3	50.79%	49.05%
		Estudio	Comercio			2	50.76%	49.11%
	Trabajo	Estudio			Otros	3	50.66%	49.24%
	Trabajo				Otros	2	50.85%	50.04%
					Otros	1	50.85%	50.10%
Casa	Trabajo	Estudio				3	50.91%	50.41%
Casa		Estudio				2	50.95%	50.52%
	Trabajo		Comercio			2	51.38%	50.43%
			Comercio			1	51.39%	50.45%
Casa	Trabajo					2	51.54%	51.79%
Casa						1	51.55%	51.92%
		Estudio				1	52.87%	52.15%
	Trabajo	Estudio				2	52.92%	52.14%
	Trabajo					1	55.37%	55.32%

D. Anexo: Modelos construidos con variables temporales y espaciales

Para las variables temporales y espaciales, se estiman los modelos para las posibles combinaciones de las 7 variables independientes, con el fin de identificar aquella combinación que reportan los menores errores en la estimación, para los cuatro motivos de viaje. Se estiman y evalúan un total de 127 modelos, los cuales se han organizado de forma ascendente según su error en evaluación y entrenamiento.

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
Ho	DuracionActividad_h	Duracion_h					Temporal	3	21,41%	22,11%
Ho	DuracionActividad_h	Duracion_h	Casa_DO				Espaciotemporal	4	21,94%	22,25%
	DuracionActividad_h	Duracion_h					Temporal	2	21,60%	22,75%
Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO			Espaciotemporal	5	21,98%	22,41%
Ho	DuracionActividad_h	Duracion_h		Estudio_DO			Espaciotemporal	4	21,96%	22,51%
Ho		Duracion_h				Otros_DO	Espaciotemporal	3	21,89%	22,65%
Ho	DuracionActividad_h	Duracion_h				Otros_DO	Espaciotemporal	4	21,89%	22,65%
Ho		Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	4	21,97%	22,65%
Ho	DuracionActividad_h	Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	5	21,97%	22,65%
Ho	DuracionActividad_h						Temporal	2	21,98%	22,65%
Ho		Duracion_h	Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	5	22,18%	22,55%
Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	6	22,18%	22,55%
Ho		Duracion_h	Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,02%	22,73%
Ho	DuracionActividad_h	Duracion_h	Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	6	22,02%	22,73%
	DuracionActividad_h	Duracion_h	Casa_DO				Espaciotemporal	3	22,01%	22,75%
Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	7	22,02%	22,77%
Ho	DuracionActividad_h	Duracion_h	Casa_DO		BuscarTrabajo_DO		Espaciotemporal	5	22,06%	22,81%
Ho		Duracion_h		Estudio_DO		Otros_DO	Espaciotemporal	4	22,26%	22,67%
Ho	DuracionActividad_h	Duracion_h		Estudio_DO		Otros_DO	Espaciotemporal	5	22,26%	22,67%

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
Ho	DuracionActividad_h	Duracion_h			BuscarTrabajo_DO		Espaciotemporal	4	22,06%	22,97%
Ho		Duracion_h			BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,20%	22,87%
Ho	DuracionActividad_h	Duracion_h			BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,20%	22,87%
Ho		Duracion_h		Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,22%	22,89%
Ho	DuracionActividad_h	Duracion_h		Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	6	22,22%	22,89%
Ho	DuracionActividad_h	Duracion_h		Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	5	22,17%	22,99%
Ho	DuracionActividad_h		Casa_DO				Espaciotemporal	3	22,55%	22,69%
	DuracionActividad_h	Duracion_h		Estudio_DO			Espaciotemporal	3	21,94%	23,33%
	DuracionActividad_h						Temporal	1	22,15%	23,15%
	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	6	22,10%	23,29%
	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	5	22,17%	23,25%
	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	5	22,36%	23,09%
	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO			Espaciotemporal	4	22,34%	23,13%
	DuracionActividad_h	Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	4	22,30%	23,17%
	DuracionActividad_h	Duracion_h	Casa_DO		BuscarTrabajo_DO		Espaciotemporal	4	22,30%	23,19%
	DuracionActividad_h	Duracion_h		Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	22,20%	23,33%
	DuracionActividad_h		Casa_DO				Espaciotemporal	2	22,44%	23,13%
	DuracionActividad_h	Duracion_h			BuscarTrabajo_DO		Espaciotemporal	3	22,34%	23,29%
	DuracionActividad_h	Duracion_h				Otros_DO	Espaciotemporal	3	22,22%	23,41%
Ho	DuracionActividad_h		Casa_DO	Estudio_DO			Espaciotemporal	4	22,54%	23,09%
Ho			Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,47%	23,17%
Ho	DuracionActividad_h		Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,47%	23,17%
Ho	DuracionActividad_h			Estudio_DO			Espaciotemporal	3	22,60%	23,07%
	DuracionActividad_h	Duracion_h	Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,31%	23,39%
Ho						Otros_DO	Espaciotemporal	2	22,45%	23,25%
Ho	DuracionActividad_h					Otros_DO	Espaciotemporal	3	22,45%	23,25%
Ho			Casa_DO			Otros_DO	Espaciotemporal	3	22,55%	23,15%
Ho	DuracionActividad_h		Casa_DO			Otros_DO	Espaciotemporal	4	22,55%	23,15%
Ho	DuracionActividad_h		Casa_DO		BuscarTrabajo_DO		Espaciotemporal	4	22,52%	23,31%
	DuracionActividad_h	Duracion_h		Estudio_DO		Otros_DO	Espaciotemporal	4	22,32%	23,57%

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
	DuracionActividad_h	Duracion_h		Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,32%	23,57%
Ho			Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	4	22,68%	23,23%
Ho	DuracionActividad_h		Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	5	22,68%	23,23%
Ho			Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,63%	23,29%
Ho	DuracionActividad_h		Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	6	22,63%	23,29%
	DuracionActividad_h			Estudio_DO			Espaciotemporal	2	22,43%	23,55%
Ho					BuscarTrabajo_DO	Otros_DO	Espaciotemporal	3	22,70%	23,29%
Ho	DuracionActividad_h				BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,70%	23,29%
	DuracionActividad_h	Duracion_h			BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,45%	23,59%
Ho	DuracionActividad_h				BuscarTrabajo_DO		Espaciotemporal	3	22,61%	23,49%
Ho				Estudio_DO		Otros_DO	Espaciotemporal	3	22,80%	23,33%
Ho	DuracionActividad_h			Estudio_DO		Otros_DO	Espaciotemporal	4	22,80%	23,33%
	DuracionActividad_h		Casa_DO	Estudio_DO			Espaciotemporal	3	22,78%	23,41%
Ho	DuracionActividad_h		Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	5	22,72%	23,55%
Ho				Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,86%	23,43%
Ho	DuracionActividad_h			Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,86%	23,43%
	DuracionActividad_h		Casa_DO			Otros_DO	Espaciotemporal	3	22,75%	23,55%
	DuracionActividad_h		Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	22,65%	23,69%
	DuracionActividad_h		Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	22,60%	23,76%
	DuracionActividad_h		Casa_DO		BuscarTrabajo_DO		Espaciotemporal	3	22,68%	23,67%
	DuracionActividad_h		Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	4	22,80%	23,55%
	DuracionActividad_h					Otros_DO	Espaciotemporal	2	22,68%	23,80%
	DuracionActividad_h		Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,73%	23,84%
	DuracionActividad_h				BuscarTrabajo_DO		Espaciotemporal	2	22,72%	23,86%
	DuracionActividad_h			Estudio_DO		Otros_DO	Espaciotemporal	3	22,77%	23,88%
	DuracionActividad_h			Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	3	22,83%	23,84%
Ho	DuracionActividad_h			Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	22,83%	23,84%
	DuracionActividad_h			Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	22,81%	23,90%
	DuracionActividad_h				BuscarTrabajo_DO	Otros_DO	Espaciotemporal	3	22,87%	24,06%
Ho		Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	6	33,74%	34,30%
Ho			Casa_DO		BuscarTrabajo_DO		Espaciotemporal	3	33,77%	34,32%

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
Ho		Duracion_h	Casa_DO		BuscarTrabajo_DO		Espaciotemporal	4	33,74%	34,48%
Ho		Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	5	33,88%	34,44%
Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	6	33,88%	34,44%
Ho			Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	33,83%	34,54%
Ho		Duracion_h		Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	33,78%	34,96%
Ho					BuscarTrabajo_DO		Espaciotemporal	2	34,03%	34,86%
Ho				Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	3	33,90%	35,00%
Ho		Duracion_h			BuscarTrabajo_DO		Espaciotemporal	3	34,03%	34,96%
Ho			Casa_DO	Estudio_DO			Espaciotemporal	3	34,75%	35,74%
Ho		Duracion_h	Casa_DO	Estudio_DO			Espaciotemporal	4	34,92%	35,68%
Ho		Duracion_h		Estudio_DO			Espaciotemporal	3	35,11%	36,27%
Ho				Estudio_DO			Espaciotemporal	2	35,23%	36,20%
Ho		Duracion_h	Casa_DO				Espaciotemporal	3	35,40%	36,08%
Ho			Casa_DO				Espaciotemporal	2	35,66%	36,22%
Ho		Duracion_h					Temporal	2	36,38%	36,77%
Ho							Temporal	1	36,63%	37,13%
		Duracion_h		Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	45,96%	46,95%
			Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espacial	4	45,83%	47,09%
		Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espaciotemporal	5	45,86%	47,07%
				Estudio_DO	BuscarTrabajo_DO	Otros_DO	Espacial	3	46,12%	47,07%
			Casa_DO		BuscarTrabajo_DO	Otros_DO	Espacial	3	45,95%	47,45%
		Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	4	45,99%	47,59%
			Casa_DO		BuscarTrabajo_DO		Espacial	2	46,12%	47,49%
		Duracion_h	Casa_DO		BuscarTrabajo_DO	Otros_DO	Espaciotemporal	4	46,06%	47,55%
		Duracion_h	Casa_DO		BuscarTrabajo_DO		Espaciotemporal	3	46,14%	47,49%
			Casa_DO	Estudio_DO	BuscarTrabajo_DO		Espacial	3	46,00%	47,63%
				Estudio_DO	BuscarTrabajo_DO		Espacial	2	46,68%	47,83%
		Duracion_h			BuscarTrabajo_DO	Otros_DO	Espaciotemporal	3	46,85%	47,77%
		Duracion_h		Estudio_DO	BuscarTrabajo_DO		Espaciotemporal	3	46,80%	47,85%
					BuscarTrabajo_DO	Otros_DO	Espacial	2	47,03%	47,71%

Ho	DuracionActividad_h	Duracion_h	Casa_DO	Estudio_DO	BuscarTrabajo_DO	Otros_DO	Tipo	Núm. Variables	Entrenamiento	Evaluación
					BuscarTrabajo_DO		Espacial	1	47,51%	48,43%
		Duracion_h			BuscarTrabajo_DO		Espaciotemporal	2	47,46%	48,53%
		Duracion_h	Casa_DO	Estudio_DO		Otros_DO	Espaciotemporal	4	49,97%	51,71%
			Casa_DO	Estudio_DO		Otros_DO	Espacial	3	49,99%	51,81%
			Casa_DO			Otros_DO	Espacial	2	50,08%	51,73%
		Duracion_h	Casa_DO			Otros_DO	Espaciotemporal	3	50,12%	51,81%
		Duracion_h		Estudio_DO		Otros_DO	Espaciotemporal	3	50,97%	52,23%
				Estudio_DO		Otros_DO	Espacial	2	51,13%	52,19%
						Otros_DO	Espacial	1	51,55%	52,25%
		Duracion_h				Otros_DO	Espaciotemporal	2	51,57%	52,31%
		Duracion_h	Casa_DO	Estudio_DO			Espaciotemporal	3	51,63%	53,17%
			Casa_DO	Estudio_DO			Espacial	2	51,66%	53,29%
			Casa_DO				Espacial	1	52,42%	54,32%
		Duracion_h	Casa_DO				Espaciotemporal	2	52,47%	54,36%
		Duracion_h		Estudio_DO			Espaciotemporal	2	53,68%	53,94%
				Estudio_DO			Espacial	1	53,87%	54,04%
		Duracion_h					Temporal	1	56,51%	57,45%