Secretaría de Transporte y Obras Públicas
Subsecretaría de Planificación de la Movilidad
Dirección General de Planificación, Uso y Evaluación
Observatorio de Movilidad y Seguridad Vial de la Ciudad de Buenos Aires

Dataset de Viajes y Etapas en Transporte Público del Área Metropolitana de Buenos Aires

Notas para su uso

2019-2022

Octubre 2023

Buenos Aires Ciudad | Observatorio de Movilidad y Seguridad Vial

/ Introducción

Con el objetivo de mejorar la información disponible para estudiar la movilidad del AMBA la Subsecretaría de Planificación de la Movilidad de la Secretaría de Transporte y Obras del Gobierno de la Ciudad Autónoma de Buenos Aires desarrolló y aplicó, en conjunto con especialistas, una metodología para estimar orígenes y destinos de los viajes con base en los datos SUBE, cuyo resultado se encuentra publicado en BADATA.

El sistema SUBE (Sistema Único de Boleto Electrónico) recolecta la información de todos los usos de las tarjetas de transporte en todos los modos de transporte del AMBA. Por las características del sistema, con excepción del modo ferroviario, solo se encuentran registrados los ascensos a las unidades de transporte colectivo o el acceso a andenes pero no los descensos o salidas.

Por otro lado, la unidad de análisis de la movilidad son los viajes, que pueden estar compuestos por una o más etapas realizadas en distintas unidades de transporte.

El objetivo de este documento es la descripción del contenido del dataset de orígenes y destinos de los viajes y sus etapas en transporte público del AMBA de fuente SUBE, los procesos intervinientes en su creación, expansión y anonimización y el diccionario de datos necesario para su utilización.

Definiciones

Etapa: se trata de un trayecto realizado dentro de un mismo vehículo de transporte, sea un colectivo, un subte (incluye Premetro) o un tren. Comienza con el ascenso a la unidad o acceso al andén y finaliza cuando se desciende de dicha unidad o se egresa de la estación. Se corresponde con una transacción SUBE (del tipo "uso"), es decir, cada validación de ingreso a una unidad en una terminal SUBE (sea en una unidad de colectivo o en molinete de ingreso a los modos guiados).

Viaje: Un viaje es un trayecto realizado entre un origen y un destino con un motivo determinado. Puede estar conformado por una o más etapas. Por ejemplo, si una persona utiliza un colectivo y un tren para llegar de su hogar a su trabajo, esto es un viaje con dos etapas.

Características del dataset

El dataset se encuentra compuesto por tres tablas: etapas, viajes y departamentos, una para cada año de 2019 a 2022.

Los datos corresponden a un día típico de octubre o noviembre los años 2019 a 2022.

Se circunscribe a los viajes realizados en el Área Metropolitana de Buenos Aires en los modos colectivo, tren y subte (incluye premetro).

/ Metodología de obtención del dataset

Los datos utilizados se obtienen a partir de los registros del sistema SUBE para un día tipo del mes de octubre o noviembre de cada año que no presenta ninguna eventualidad que afecte los patrones de movilidad (feriados, grandes manifestaciones, eventos meteorológicos extremos, etc.). A partir de los datos crudos registrados por el SUBE se realiza un proceso de depuración que consiste en eliminar primero aquellos registros que no corresponden a transacciones válidas de ascenso a las unidades del sistema de transporte público y luego aquellas que presentaran alguna falta de información (registros con id de tarjetas nulos, problemas en la geolocalización, entre otros). Sobre este subconjunto de datos se procedió a imputar los destinos.

Imputación de destinos

Para la imputación de destinos se tuvieron en cuenta dos supuestos:

Parada más cercana: los usuarios que utilizan el transporte público parten de un lugar cercano a la estación de destino de su etapa anterior para comenzar el siguiente.

Regla de la simetría diaria: los usuarios inician el primer viaje del día y concluyen el último viaje del día en el hogar.

De esta forma, utilizando el supuesto de la parada más cercana, para cada usuario se imputó como destino de cada etapa la parada/estación de la línea utilizada en dicha etapa que minimice la distancia con el origen de la etapa siguiente. Si dicha distancia es superior a un margen de tolerancia establecido en 2,2km. Para el caso de la última etapa del día, dado que no existe una etapa posterior, se utilizó el supuesto de la simetría diaria y se consideró a la primera etapa del día como si fuese el destino de la última. Para el caso de los trenes que presenten una transacción de tipo "*check out*" (registro de salidas de estaciones ferroviarias) se utilizó la localización de esta transacción.

En el proceso de imputación de destinos se considera a las distintas líneas del subte (A, B, C, D, E, H, Premetro) como si fueran ramales de una única línea, es decir que todas las estaciones del sistema son posibles orígenes y destinos entre sí. Esta situación se replica al interior de cada línea ferroviaria.

/ Metodología de obtención del dataset

Para concatenar las diferentes etapas dentro de un mismo viaje, se utilizó el criterio de RED SUBE, según el cual toda etapa hecha en un modo o línea diferente dentro de una ventana de 2 horas con respecto a la primera transacción conforman un único viaje.

El código utilizado para la construcción de este dataset es una adaptación del utilizado por el BID en el trabajo "Construcción de una Matriz OD del Transporte Público para el Área Metropolitana de Buenos Aires en base a datos SUBE". Para mayores detalles puede leerse el documento metodológico disponible en: https://github.com/EL-BID/Matriz-Origen-Destino-Transporte-Publico

Agregación Espacial

Los datos fueron trabajados seccionando el área de análisis en hexágonos H3 de resolución 10. Las coordenadas presentadas son los centroides de dichos hexágonos. Para mayor información sobre la utilización de hexágonos puede consultarse: https://h3geo.org/

Anonimización

Con el fin de ofrecer un dataset abierto plausible de ser utilizado del mejor modo por la sociedad civil y al mismo tiempo preservar los datos privados de las personas usuarias del transporte público, se han utilizado algunas técnicas de anonimización de datos, siguiendo la guía "Consejos y recomendaciones para la anonimización de datos personales" de la Subsecretaría de Políticas Públicas Basadas en Evidencia del Gobierno de la Ciudad de Buenos Aires. En primer lugar, se enmascaran los datos de los id tarjeta de modo que no referencien un atributo real de la tarjeta. En segundo lugar, no se brinda información de la línea utilizada, solo el modo. En tercer lugar, se procede a agregar datos tanto temporal (informando solo la hora entera de la transacción) como espacialmente, ofreciendo la coordenada del centroide de una unidad de análisis espacial más agregada.

/ Metodología de obtención del dataset

Factores de expansión

Durante el proceso de estimación de destinos, tal como se comentó, se van eliminando transacciones realizadas cuando se trata de datos incompletos o erróneos. Estos pueden ser, por ejemplo: transacciones sin identificador único para la tarjeta, transacciones únicas en el día, errores cartográficos o de GPS, entre otros.

Como resultado de esto, la cantidad de etapas/transacciones al final del proceso es menor que el total de transacciones iniciales. Para poder trabajar con una aproximación a la dimensión completa inicial y llegar así a una estimación de patrones de viaje que sea consistente con el total de transacciones diarias registradas en el SUBE para el AMBA se crea un factor de expansión, que se elabora tomando como base de referencia las transacciones totales por línea, procurando que la expansión de las etapas finales por línea iguale ese total. Como la línea es un atributo de la etapa, para contar con un factor de expansión para viajes, se ha hecho un promedio simple de los factores de expansión de las etapas que integran cada viaje. En caso de que la perdida de casos en una línea particular haya sido excesiva, y por ende el factor de expansión para esa línea es muy alto, se fuerza un valor tope 3 para que no perturbe el resultado.

Se resumen en la siguiente tabla los principales motivos de pérdida de datos en % con respecto a las transacciones originales, para el año 2022.

Motivo	Transacciones (%)
Problemas geolocalización	6,6%
Transacción única en la tarjeta	7,6%
Problemas imputación destino	6,2%
Transacciones simultaneas	6,4%
Total	26,8%

/ Diccionario de datos – Tabla Etapas

Identifica las etapas con el modo y hora de la transacción. La misma tabla tiene también las coordenadas agregadas del origen y destino y los departamentos censales correspondientes. Cada etapa cuenta también con el id tarjeta (enmascarado) y el id de viaje que permite vincularla con la tabla viajes. Las etapas con destinos sin imputar fueron eliminadas. Esto puede causar que un viaje con más de una etapa tenga al menos una de ellas sin destino imputado y ausente en la tabla “etapas”. Esta situación se indica en la columna “viaje_incompleto”. Esto se debe a que algunos resultados se calculan sobre etapas (por ej. la demanda total de una linea) y se buscaba perder la menor cantidad de datos posibles. Eliminar etapas con destinos imputados correctamente porque otra a etapa del mismo viaje no se pudo imputar un destino conspiraba contra esta posibilidad. Al aclarar en ambas tablas esta situación deja en manos del usuario el modo en que prefiere trabajar.

Dato	Tipo de dato
id_tarjeta	Identificador único enmascarado de la tarjeta
id_viaje	Identificador del viaje para cada tarjeta
id_etapa	Identificador de la etapa para cada viaje y tarjeta
hora	Hora de la transacción de inicio
modo	Identificador único del modo
linea	Línea de la etapa (solo para Subte y tren)
lon_o	Longitud de origen de la etapa
lat_o	Latitud de origen de la etapa
lon_d	Longitud de destino de la etapa
lat_d	Latitud de destino de la etapa
departamento_o	Departamento censal de origen de la etapa
departamento_d	Departamento censal de destino de la etapa
factor_expansion	Factor de expansión de la etapa
viaje_incompleto	Indica si la etapa pertenece a un viaje con etapas sin imputar destino

/ Diccionario de datos – Tabla Viajes

Identifica los viajes, sus orígenes y destinos, la cantidad de etapas y los modos utilizados. La misma tabla tiene también las coordenadas agregadas del origen y destino y los departamentos censales correspondientes. Cada viaje cuenta también con el id tarjeta (enmascarado) y el id de viaje que permite vincularla con la tabla etapas. El atributo “etapas_incompletas” indica si ese viaje tiene alguna etapa cuyo destino no pudo ser imputado para que el usuario decida si corresponde utilizarse en su análisis.

Dato	Tipo de dato
id_tarjeta	Identificador único de la tarjeta
id_viaje	Identificador único del viaje
hora	Hora del viaje
cantidad_etapas	Número de etapas validas completas realizadas
etapas_subte	Número de etapas realizadas en subte
etapas_tren	Número de etapas realizadas en tren
etapas_colectivo	Número de etapas realizadas en colectivo
lon_o	Longitud de origen del viaje
lat_o	Latitud de origen del viaje
lon_d	Longitud de destino del viaje
lat_d	Latitud de destino del viaje
departamento_o	Departamento censal de origen del viaje
departamento_d	Departamento censal de destino del viaje
factor_expansion	Factor de expansión del viaje
etapas_incompletas	Indica si el viaje tiene alguna etapa sin imputar destino

/ Diccionario de datos – Tabla Departamentos

Para facilitar el análisis a nivel jurisdiccional y el *matcheo* con otras fuentes de datos, se incluyeron además de las coordenadas geográficas de los centroides de los hexágonos h3 de resolución 10, la identificación con los departamentos y sus códigos censales.

Dato	Tipo de dato
link_departamento	Identificador de departamento del censo 2010
codpcia	Identificador de provincia del censo 2010
departamento	Nombre del departamento
provincia	Nombre de la provincia