Estimación de Origen-Destino de usos en colectivo en base a datos registrados por el sistema SUBE

Sidoni Guido[1]

[1] Universidad de Buenos Aires
sidoni.g@husky.neu.edu

Abstract. Los Sistemas de Recolección Automática de Datos (ADCS) se han vuelto muy populares para sistemas de transporte de todo el mundo. Aunque generalmente los ADCS fueron diseñados con el objetivo de ser funcionales en el cobro de tarifas, la información recolectada tiene un amplio rango de utilización. Esta tesis ilustra el potencial del ADCS de SUBE para proveer información novedosa a las agencias de transporte de Argentina, a bajos costos marginales, y con bajo tiempo de respuesta en comparación con métodos más convencionales como las encuestas. Para esto fue necesario el procesamiento de los datos del ADCS de SUBE , la utilización de métodos de Minería de Datos junto con soporte de tecnologías como el manejo de bases de datos relacionales, el uso de sistemas de información geográfico y el uso de técnicas de programación. Esta tesis presenta como resultado una matriz de origen destino zonificada, en base en usos realizados en colectivo en el área del AMBA (Área Metropolitana de Buenos Aires) para la primera semana de Mayo de 2015. La tesis además aborda distintos métodos para dar solución a problemas relacionados con el pre procesamiento de la información de SUBE, para que esta sea adecuada como entrada para el algoritmo de estimación de destino.

1 Introducción

El Sistema Automático de Recolección de Datos (ADCS) es un sistema destinado a recolectar información concerniente a diferentes aspectos del transporte. Está integrado por un conjunto de subsistemas independientes, los cuales pueden o no estar implementados en un sistema de transporte. Estos subsistemas son: el Sistema de Cobro Automático (Automatic Fare Collection (AFC)), el Sistema de Localización de

Vehículos (Automatic Vehicle Location (AVL)) y el Sistema de Conteo Automático de Pasajeros (Automatic Passenger Count (APC)). Si bien el sistema APC fue diseñado para recolectar datos para el soporte en la toma de decisiones, los otros tienen funciones bien específico, recolectar información sobre el pago en el transporte público y recolectar información de la localización de la flota de vehículos. Sin importar el fin último por el cual diseñaron dichos sistemas, estos pueden ser utilizados para proveer información importante a las autoridades de transporte, como ser la Matriz de Origen y Destino de viajes.

La matriz Origen Destino es una matriz cuadrada que muestra el número total de viajes realizados por cada par Origen Destino en la red de transporte. Provee información sobre donde los pasajeros inician y donde terminan sus viajes. Por lo que la matriz de origen destino es de fundamental importancia tanto para la planificación de transporte a largo plazo como para la operatoria diaria del sistema de transporte. La generación de esta matriz puede realizarse mediante dos métodos diferentes, una mediante encuestas y otra mediante el procesamiento de los datos recolectados por el ADCS del sistema de transporte en caso de que este esté instalado.

Para el primer método, los datos sobre de origen y de destino de los viajes son recolectada mediante encuestas específicamente diseñadas para tal fin. Sin embargo estos estudios son costosos, están sujetos a sesgos y no son fáciles de actualizar con frecuencia. Un antecedente de este tipo de estudios en Argentina es la realizada por la Secretaría de Transporte de la Nación y publicada en el año 2007 bajo el nombre de Intrupuba [1]. Por otro lado los Sistemas Automáticos de Recolección de Datos (ADCS) posibilitó obtener información sobre el origen y el destino de los viajes de manera automática, aunque no siempre en forma directa, a un costo marginalmente bajo. La complejidad en la utilización de los datos obtenidos mediante ADCS para generar un análisis de origen y destino de los viajes depende de cómo esté implementado el sistema ADCS, en particular el sistema de controles de ingreso. Por ejemplo la inferencia no es sencilla en sistemas como los implementados por Chicago Transit Authority CTA, New York City Transit Authority, e incluso en el sistema de cobro en colectivos con SUBE, en los cuales solo se implementa un control del usuario al ingreso de las unidades de transporte. Para estos casos, el método más ampliamente difundido de estimación del destino es el denominado ***Destination Inference***, fue abordado inicialmente por Berry ar al. (2002) [3] en el año 2002 y seguido por Zhao 2007 [4] y Munizaga 2012 [2]. A partir de estos estudios definimos los siguientes criterios a seguir para la inferencia del destino para ser aplicados a los datos de ADCS de SUBE:

- cada tarjeta corresponde a un pasajero, por lo que se utilizará indistintamente el concepto de tarjetas o pasajero.
- el pasajero no se trasladó con ningún medio privado entre el descenso de un viaje y comienzo del siguiente para un mismo día.
- Los pasajeros no caminan largas distancias para comenzar un nuevo viaje. Para esta tesis se toma el supuesto de Zhao 2004 [4] en el cual se definió que los pasajeros no caminan por un periodo mayor a 5 minutos o 406 metros [4].

- Un gran porcentaje de usuarios terminan su último viaje del día en la estación de la cual partió al comienzo del día.
- Para este estudio se utilizaron todos los viajes completos de SUBE para la primera semana de mayo, (aproximadamente 57 millones*),* que utilizaron solamente colectivo como medio de transporte.

2 Pre procesamiento: Identificación de Paradas de colectivo

Para lograr implementar una versión del algoritmo de *Destination Inference* fue necesario pre procesar los datos de SUBE de manera tal de poder dar solución a dos aspectos claves necesario:

1. La identificación de todas las paradas de colectivo y su rango de acción (look up table de paradas).
2. La creación de la tabla de distancias entre paradas de colectivo (look up table de distancias).

En esta sección proponemos métodos para lograr generar ambas tablas.

2.1 Asignación de Rutas

El primer paso para determinar el conjunto de paradas de colectivo consiste en asociar a cada línea / ramal SUBE una ruta geográfica teórica tomadas de un archivo shape de rutas. El problema a resolver es que no hay coincidencia entre los nombres de las rutas SUBE y las geográficas. Por lo que se propone el siguiente método: realizar una votación por la cual los puntos de control[1] de 25 turnos de colectivo tomados al azar de cada línea / ramal voten por la ruta que mejor se ajusta mediante la medida RMSE. La figura 2 muestra las rutas teóricas posibles para la línea X y la figura 3 los puntos de control SUBE para la misma línea para los 25 turnos seleccionados.

Para determinar el valor de ajuste se calculó el RMSE (Root Mean Square Error, (1)) entre, la ruta, variable estimada $\hat{Y}$, y los puntos de control, variable dependiente Y. Su diferencia ($\hat{Y}$ - Y) consiste en medir la distancia mínima entre ellos, calculada con técnicas de GIS mediante fórmula ”Haversine”. Finalmente cada turno vota a aquella ruta que tiene menor RMSE y por lo tanto mejor ajuste.

$$RMSE = \sqrt{\frac{\sum_n^{t=1}(\hat{y}_t - y)^2}{n}} \qquad \textbf{(1)}$$

[1] Un Punto de control es una ubicación GPS que las unidades de colectivo guardan cada una cierta cantidad de minutos.

Fig 1. Rutas posibles para la línea X

Fig 2. puntos de control SUBE para la línea X

2.2 Asignación de Sentido

Siguiendo con el objetivo de armar la tabla de paradas de colectivo, tenemos que considerar que las mismas se ubican en distintos ubicaciones cuando el recorrido es de ida o vuelta. Por lo que fue necesario determinar si un determinado punto de control se mueve en dirección de ida o vuelta sobre por ruta. Esto nos posibilitará luego, agrupar los puntos de control de un mismo sentido a los efectos de determinar las paradas de colectivo. El sistema SUBE no registra si el colectivo se dirige en sentido ida o vuelta por lo que debemos obtenerla en base a las posiciones sucesivas de los puntos de control de un turno (puntos GPS sucesivos) sobre la ruta teórica asociada. Para ello se desarrollará un algoritmo ad-hoc en python. El algoritmo completo puede verse en tesis de Sidoni 2016 [6].

2.3 Identificación de paradas de colectivo

En esta sección se propone un método para identificar los puntos de ascenso (paradas de colectivo) de cada línea /ramal / sentido mediante el agrupamiento de los puntos de control. La metodología para la identificación de las paradas de colectivo se basa en la hecho de que los usuarios acercan su tarjeta apenas ascienden a la unidad de colectivo, grabándose una transacción de uso con su fecha y hora. Se espera que, al utilizar un gran número ascensos ya georreferenciados, los correspondientes a una misma parada permanezcan cercanos entre sí y lejanos con respecto de los ascensos pertenecientes a otras paradas. Además, para aumentar la precisión de las ubicaciones de las paradas se aplica el siguiente criterio de filtro:**sólo se tomarán en cuenta aquellos usos de colectivo georreferenciados que se distancian en un tiempo menor de 10 segundos con respecto a su punto de control más cercano.** El resto de los puntos serán descartados En la figura 5 se muestra en rojo los puntos de control muestreados para la línea Y luego de aplicado el filtro. Puede notarse que los mismos presentan cierta agrupación natural.

Fig 3 Puntos de control luego de aplicado el criterio filtros para la línea Y

Al tener identificados los grupos naturales, estamos en condiciones de definir los grupos de manera automática utilizando el algoritmo *K-Means (*con parametros k de 10 a 120) y utilizando *Silhoutte* como medida de la calidad del agrupamiento. A manera de ejemplo podemos ver en la Fig 6 la evaluación del *Silhoutte* para la línea Y ramal 325 para su sentido de ida y de vuelta. En el mismo puede observarse que a medida que crece el valor del parámetro *k* (número de grupos) el valor del *Silhoutte* se incrementa hasta que llega a su máximo valor en *k*=43 para ida y *k*=38 para la vuelta con valores de *silhouette* promedios de 0.87 y 0.90 respectivamente.

2.4 Generación de Look up tables

En base al pre procesamiento anterior se genera una **look up table de paradas** que contiene información de cada parada detectada por el método de agrupamiento: el identificador formado por línea / ramal / sentido y un número secuencial de parada, el *milepost* de la parada en sí mismo, los *milepost* de los límites inferior y superior, la latitud y la longitud de la parada. Esta tabla contiene información sobre las 56.991 paradas detectadas. Además se genera la **look up table de distancias**. Se trata de un tabla que contiene para cada parada, la información de la parada más cercana de las otras línea / ramal / sentido. Dicha tabla contiene 72 millones de registros aproximadamente.

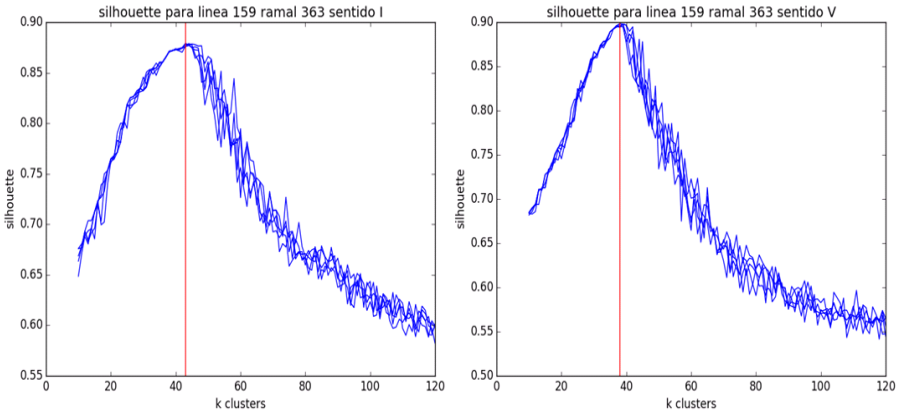

Fig 4, evaluación de *silhuette* para determinar número óptimo de clusters.

3. Estimación de Origen y Destino

Para la estimación de la parada de colectivo origen que dió inicio al viaje, se abordó mediante dos pasos:

1. Establecer el *milepost*[2] del pasajero al momento del ascenso.
2. Utilizar el *milepost* del ascenso para determinar la parada de colectivo de ascenso mediante la lookup table de paradas.

Para la estimación de la parada de colectivo de destino se realizó una implementación en python del algoritmo *Destination Inference*, el método de inferencia se basa en la cadena de viajes de un pasajero y en las distancias entre las paradas de colectivo (look up table de distancias). El método puede explicarse de la siguiente manera (figura 5): La cadena de viajes comienza con la etiqueta "1er viaje del día" donde el pasajero ingresa al colectivo de la línea 60 en la parada 3. Seguidamente su segundo viaje del día lo inició en la parada 4 de la línea 93. Dadas estas dos paradas de ingreso, se infiere que la parada de destino del primer viaje del día fue en la parada 11 (de la líneas 60) ya que esta es la que se encuentra a menor distancia de la parada de inicio del segundo viaje del día. El mismo procedimiento puede seguirse para estimar el destino del segundo viaje del día.

[2] Milepost: es la distancia recorrida sobre la ruta teórica medida desde el oriden de esta

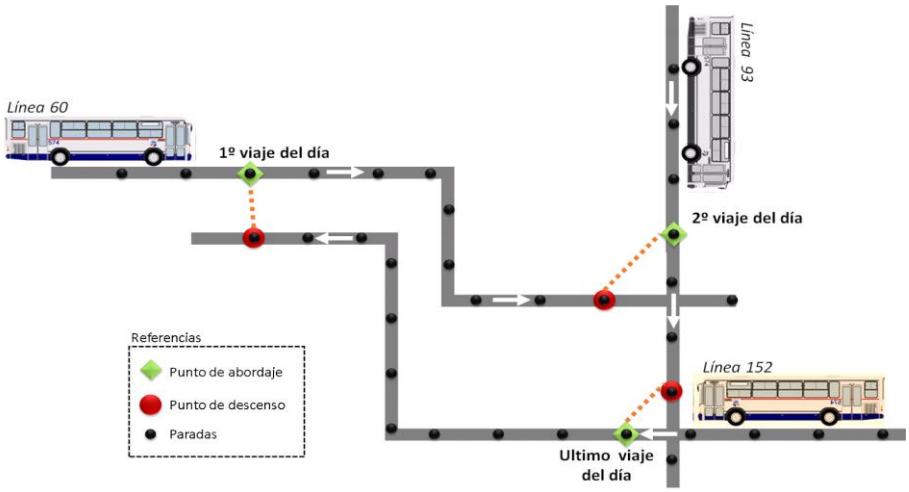

Fig. 5. Esquema de explicación de método de inferencia de destino por medio de la cadena de viajes del pasajero

Como puede observarse en la figura 6 el algoritmo fue implementado en 3 ciclos anidados (nested loops): el primero itera por cada día, el segundo por cada tarjeta, el tercero para cada uno de los viajes de la tarjeta ordenados cronológicamente. Para cada uno de los viajes recuperados se corre el subproceso de Inferencia de destino.

En la figura 7 podemos ver el diagrama de flujo del subproceso Inferencia de Destino. En él se pregunta si el próximo viaje es el último viaje del día para la tarjeta, si lo es se toma el origen del primer viaje del día para estimar el destino, caso contrario se toma el origen del viaje siguiente. Una vez definido el origen que se utilizara para la estimación se llama al subproceso Próximo Viaje

En el subproceso "próximo viaje" (figura 8) se verifica que el origen del primer viaje no sea igual al origen del viaje siguiente, esto se verifica comparando la igualdad entre los "Id de lote" y el número de parada de ascenso de ambos viajes. En caso de que sean iguales se verifica que el tiempo entre ambos viajes sea menor a tres minutos. Si es menor, estamos ante un caso de multiviaje, esto quiere decir que un mismo usuario pagó varios viajes con la misma tarjeta ya que evidentemente viajó acompañado. En este caso a todos los multiviajes se le asigna el destino del viaje original. En caso que el tiempo transcurrido entre los dos viajes sea mayor a tres minutos, el algoritmo no encuentra solución de destino. Por otro lado en caso de que los orígenes no sean iguales, el destino del viaje origen es asignado en base a la matriz de distancias entre paradas calculada en el apartado anterior.

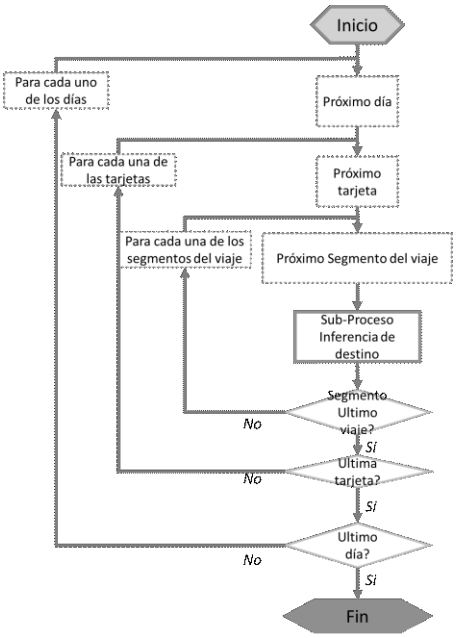

Fig. 6. Esquema general del algoritmo de inferencia del destino

Fig. 7. Diagrama de flujo del subproceso Inferencia de destino

Luego de corrido el algoritmo de inferencia de origen y destino sobre los 57 millones de viajes circulares pertenecientes a los días de 1 al 8 de Mayo de 2015, se tuvo éxito en la estimación para el 48% de ellos.

En este estudio logramos estimar las paradas de colectivo de cada viaje individual, por otro lado el estudio Intrupuba [1] muestra una Matriz Origen Destino de colectivos agregada a nivel zona en función de los principales patrones de circulación de las líneas de colectivo y de la orientación de las principales vías de comunicación. De este modo, el área de estudio quedó clasificada en seis áreas[3].

[3] Centro: Parque patricios, retiro, recoleta, san cristobal, san nicolas, constitución, monserrat, balvanera, san telmo Norte:Belgrano, Nuñez, Palermo, Coghlan, Vicente Lopez, San Isidro, San Fernando, Tigre, Escobar. Noroeste: Agronomía, Villa Crespo, Chacarita, Paternal, Villa del parque, Saavedra, Villa Pueyrredon, Villa Devoto, Villa Urquiza, Colegiales, Villa Real, Parque Chas, Villa Ortuzar San Martín, Tres de Febrero, Hurlingham, San Miguel, Jose C. Paz, Malvinas Argentinas, Pilar. Oeste: Flores, Villa Santa Rita, Caballito, Almagro, Mataderos, Parque Chacabuco, Parque Avellaneda, Floresta, Villa Gral Mitre, Boedo, Velez Sarfield, Villa Luro, Versalles, Monte Castro, Liniers, Moron, Ituzaingó, Moreno, Merlo. Sudoeste: Villa Lugano, Villa Riachuelo, Villa Soldati, La Matanza, Est Echeverria, Ezeiza. Sur:

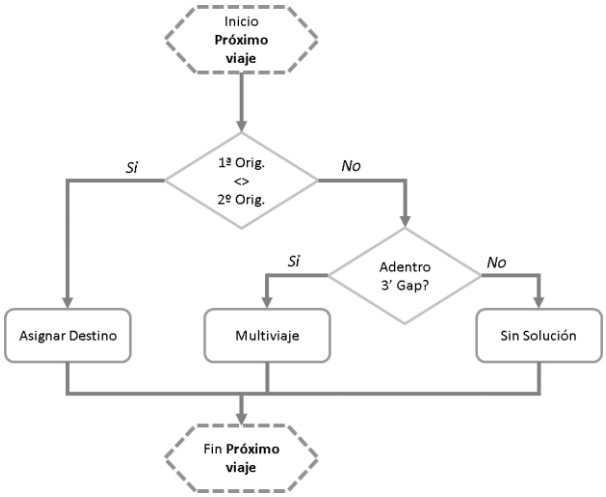

Fig 8. Diagrama de flujo del subproceso Próximo viaje

En la figura 9 se presenta la matriz zonificada de Origen-Destino para los días de hábiles 4, 5, 6, 7 y 8 de Mayo de 2015.

Origen \Destino	centro	noroeste	norte	oeste	sudeste	sudoeste	sur	% Destino
centro	5.02%	0.58%	1.55%	1.33%	1.30%	0.34%	1.25%	11.37%
noroeste	0.49%	11.75%	2.36%	1.88%	0.02%	0.44%	0.12%	17.06%
norte	1.52%	2.16%	11.54%	1.12%	0.06%	0.14%	0.23%	16.77%
oeste	1.29%	2.10%	1.05%	7.95%	0.10%	1.68%	0.58%	14.74%
sudeste	1.16%	0.02%	0.07%	0.12%	10.10%	0.02%	1.28%	12.78%
Sudoeste	0.37%	0.42%	0.14%	1.76%	0.02%	10.79%	0.49%	13.99%
sur	1.23%	0.18%	0.29%	0.66%	1.26%	0.55%	9.11%	13.29%
%Origen	11.08%	17.20%	17.01%	14.81%	12.87%	13.97%	13.06%	100.00%

Fig. 9. Matriz origen destino para días hábiles 4, 5, 6, 7 y 8 de mayo 2015

La matriz de la figura 9, muestra que la mayor cantidad de viajes en colectivo se realizan dentro de cada zona, alcanzando un total del 66,2% del total de viajes analizados. También puede observarse que la mayor cantidad de viajes en colectivo se origina en la zona Noroeste con 17,2% y le siguen las zonas Norte y Oeste, con alrededor del 17% y 14% respectivamente. Siendo la zona centro la que da origen a la menor cantidad de viajes (5%).

Comparado con la matriz de origen destino publicada por el estudio intrupuba 2007, ambas matrices presentan mayor porcentaje de viajes interzona (diagonal de la matriz)

Nueva Pompeya, Barracas, Lanús, Lomas de Zamora, Alte Brown, Presidente Perón, Sudeste, La Boca, Puerto Madero, Avellaneda, Quilmes, Florencio Varela, Berazategui, La Plata

que los porcentajes intrazona, y los valores totales de origen y destino por zona rondan en ambos matrices entre el 9% y el 17%.

4. Conclusiones

El presenta estudio demuestra que mediante el pre procesamiento de los datos de ADCS de SUBE, es posible generar información valiosa para las autoridades de Transporte de la Nación, sin tener que incurrir en los gastos derivados en el armado de encuestas.

Por otro lado, si bien los resultados obtenidos son similares a los ya informados por Intrupuba [1], generar una matriz de origen destino mediante datos de ADCS poder desagregar los resultados a nivel de parada de colectivo y con cortes horarios, detalles que son imposibles de concebir mediante encuestas.

Finalmente, hay que considerar que los resultados aquí obtenidos no corresponde a una matriz de origen destino completa, ya que no se consideraron viajes en trenes ni en subterráneos como así tampoco se hizo un estudio de trasbordo que encadene usos consecutivos. Futuros estudios deberán abordar estos aspectos de manera de enriquecer los resultados de este estudio.

Referencias

1. Intrupuba (2007). Investigación de transporte urbano público de Buenos Aires. Ministerio de Planificación Federal. Secretaría de Transporte 2007
2. Munizaga (2012), Estimation of a dissagregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile.
3. Berry et al. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data
4. Zhao et al. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix
5. Robert Levine (1999).THE PACE OF LIFE IN 31 COUNTRIES.California State University, Fresno.
6. Sidoni (2016). Estimación de Origen y Destino de viajes en base a datos registrados por el sistema SUBE. Universidad de Buenos Aires. Buenos Aires, Argentina.