

UNIVERSIDAD DE CHILE
Facultad de Ciencias Físicas y Matemáticas
Departamento de ciencias de la computación
CC6908 Introducción al trabajo de título



*Visualización de estructuras
espaciales de desplazamiento a
partir de datos de transporte público*

24 de noviembre de 2014

Profesor Guía
Marcela Munizaga

Profesor Co-guía
Benjamín Bustos

Memorista
Felipe A. Hernández G.
fhernand@dcc.uchile.cl
90977379

Índice

1. Introducción	1
2. Motivación	3
3. Objetivos	3
3.1. Objetivo General	3
3.2. Objetivos específicos	3
4. Metodología	4
5. Revisión de antecedentes	7
5.1. Análisis de bibliografía	7
5.1.1. Redes espaciales	7
5.1.2. Construcción y representación de la red	8
5.1.3. Análisis complejo de la red	8
5.1.4. Visualización de información	13
5.2. Análisis de datos	13
5.2.1. Sistema tarifario	13
5.2.2. Estimación de bajadas	14
5.2.3. Descripción de los datos	16
5.2.4. Conclusiones	17
6. Estrategia de procesamiento	18
6.1. Herramientas a utilizar	18
6.2. Filtrar la data	18
7. Bibliografía	19

Índice de tablas

1. Atributos a utilizar de la tabla ETAPAS	16
--	----

1. Introducción

En los últimos años la utilización de tecnología en el transporte público ha ido en aumento debido a varios factores, mayor regulación, usuarios más exigentes, aumento en seguridad, etc. Lo anterior ha llevado al sistema público de transporte a implantar diversos dispositivos que permiten controlar los aspectos mas relevantes al momento de transportar una persona de un punto a otro.

Dentro de las tecnologías más usadas podemos nombrar AVL (Automatic Vehicle Location), que permite conocer la posición geográfica de un vehículo en todo momento con un margen de error bajo y los sistemas AFC (Automated Fare Collection) que automatizan el proceso de pago, en particular, nos interesa el basado en tarjetas de pago, que albergan un chip que permite mantener un saldo para que sea utilizado al abordar un bus, esto implica la existencia (paralelamente) de dispositivos asociados a los buses o paraderos¹ que permitan registrar el correcto descuento del valor asociado al pasaje, concepto que llamaremos validación.

Transantiago² es el sistema de transporte público de Santiago de Chile que implementa las tecnologías nombradas anteriormente, por lo que hoy en día se sabe que se realizan aproximadamente 6.000.000 de validaciones durante un día laboral³, lo que genera una cifra cercana a los 35.000.000 de transacciones a la semana (incluyendo sábado y domingo) con aproximadamente 3.000.000 de tarjetas de pago. Por otro lado, hay 80.000.000 de emisiones proveniente de la tecnología AVL del sistema. Al procesar estos datos en conjunto es posible identificar el paradero de origen, recorrido utilizado para desplazarse y el paradero de destino, este último requiere un procesamiento adicional basado en una metodología desarrollada por Munizaga y Palma [4] que logra una identificación acertada en el 80 % de las validaciones.

La estructura espacial moderna de las ciudades ha sido formada, en gran medida, por avances en transporte y comunicaciones [1]. La forma en la cual se mueven los habitantes de una ciudad ha ido modificando la estructura de esta, motivados por la transferencia de recursos como materiales, dinero, personas e información. Considerando una persona como un transportador de recursos de un área urbana a otra es que se identifican las siguientes

¹Lugar físico donde un bus de transporte público se detiene para que personas ingresen y/o desciendan a el.

²fue implementado a partir del año 2007.

³Lunes, martes, miércoles, jueves o viernes.

estructuras espaciales urbanas [3]:

- Centros de flujo: Se refiere a las áreas que sirven para conectar otro par de áreas para transferencia de personas. Funcionan como puentes espaciales entre distintas áreas.
- Centros: Se refiere a áreas que concentran personas. Pueden diferir de los Centros de flujo, pero a menudo, son lo mismo.
- Bordes: Se refiere a límites socioeconómicos generados a partir de la agrupación de paraderos que divide la ciudad en pequeños barrios que llamamos comunidades.

Lo anterior se enmarca en la necesidad de comunicar esta información a personas sin una formación ingenieril dado que es un proyecto que abarcará muchos campos de investigación, como lo es la arquitectura o planificación urbana, por lo que es necesario transmitir datos de forma clara y concisa.

De todo lo relatado podemos ver que hoy en día el sistema de transporte público de Santiago cuenta con una gran cantidad de datos pasivos por lo que existe una gran base de datos que mantiene un potencial de información que puede mejorar la planificación y operación del sistema, además de tener la potencialidad de detectar otras necesidades. Sin embargo, con las capacidades de procesamiento actuales no es posible obtener información que complemente los resultados de los indicadores usados actualmente.

Según lo anterior, el problema que se busca resolver en esta memoria es interesante de abordar debido a que ayudará a entender la estructura espacial de los viajes realizados por la población y permitirá diseñar servicios pensando en las necesidades observadas de los usuarios. Esto ayudará a:

- Mejorar las posibilidades de realizar actividades en el entorno de la zona de residencia.
- Disminución de la demanda en los Centros de flujo.
- Disminución en el tiempo requerido para trasladarse hasta el punto de interés para una comunidad determinada.
- Mejorar las condiciones de viaje de grupos vulnerables.

2. Motivación

Como hemos dado a conocer, existe una gran fuente de datos con mucha información pero que actualmente encuentra sus dificultades en el procesamiento y la forma en que puede ser comunicada. Por lo que una solución a este problema puede abrir las puertas a nuevas preguntas y según esto, nuevas investigaciones.

También es interesante académicamente debido a la masividad de los datos, ya que se deberá implementar una estrategia de procesamiento que permita manejar millones de registros y además realizar análisis sobre estos que permitan comunicar información por medio de la visualización.

3. Objetivos

3.1. Objetivo General

“Diseñar una herramienta que permita identificar y visualizar estructuras espaciales de movimiento en la ciudad de Santiago utilizando datos pasivos y masivos de transporte público.”

3.2. Objetivos específicos

1. Construir modelo de red para la ciudad de Santiago.
2. Identificar patrones de viaje, centros y puntos de alto flujo de pasada.
3. Desarrollar una herramienta que permita visualizar las estructuras espaciales.

4. Metodología

Esta metodología está basada en una investigación publicada en la *International Journal of Geographical Information Science* [3], por lo que los procedimientos ya han sido probados en otro contexto, reduciendo de esta forma posibles inconvenientes que puedan ocurrir a lo largo del desarrollo de esta memoria.

Los datos a utilizar se han definido como los producidos en una semana de calendario (lunes a domingo). Estos ya se encuentran procesados según la metodología diseñada por Munizaga y Palma [4], por lo que la data corresponde a una tabla de una base de datos PostgreSQL llamada *tabla_de_etapas* donde cada fila representa una etapa de un viaje⁴. Según lo anterior la cantidad de datos a utilizar es de aproximadamente 35.000.000, que corresponde a la cantidad de etapas realizadas por el 80 % de las transacciones del sistema AFC⁵.

Dado lo anterior, el desarrollo de esta memoria considera la siguiente metodología de trabajo:

1. Investigación bibliográfica

Se está realizando una recopilación y redacción de las ideas y estrategias más relevantes que aporten y justifiquen la base teórica de esta memoria.

2. Estudio de la data.

Se realizará un estudio de los datos existentes para comprender concretamente las bases de datos requeridos.

3. Definición de estrategia de pre-procesamiento de datos.

Se investigará sobre las estrategias de pre-procesamiento y elegirá la que mejor se adapte en base al estudio realizado en el ítem anterior. Dentro de esta etapa se llevará a cabo la normalización y selección de la data para realizar los análisis.

4. Construcción de la red de nodos.

En esta etapa se realizará la construcción de un grafo dirigido con nodos a partir de la data estudiada.

⁴un viaje puede tener una o más etapas.

⁵Automatic Fare Collection

5. Análisis de la red

a) Definición de propiedades básicas.

Aquí asociaremos un atributo de las estructuras urbanas a cada propiedad matemática de un grado a partir de las ideas obtenidas de la investigación bibliográfica.

b) Definición de centralidades.

1) Centro de flujo.

Se define el concepto de Centro de flujo en un grafo (*betweenness centrality*) y se propone una fórmula para medirlo.

2) Centro

Aquí estudiaremos y definiremos la estrategia para detectar centros de la ciudad ocupando el algoritmo *PageRank*.

c) Estructura de comunidad.

Para la detección de estructuras de comunidad se utilizará el software *infomap*.

6. Análisis espacial

a) Interpolación espacial.

Lo relevante de esta etapa es relacionar una zona geográfica a un paradero de bus de manera de poder particionar la ciudad.

b) Cálculo estadístico.

En esta etapa se realizará la asociación de las comunidades detectadas a las áreas geográficas establecidas en la interpolación espacial.

7. Análisis de los resultados.

Se estudiarán los resultados obtenidos.

8. Definir visualizaciones y nivel de interactividad de cada una.

A partir del punto anterior se definirán las visualizaciones a realizar y las posibles interacciones que puedan haber en cada una de ellas.

9. Diseño de aplicación de visualización.

Se desarrollará una aplicación que permita ver cada una de las implementaciones definidas en el punto anterior.

Este trabajo será realizado a lo largo de 2 semestres (2014-2 y 2015-1) por lo que se dividirá de la siguiente forma:

- Semestre 2014-2
 1. Investigación bibliográfica
 2. Estudio de la data
 3. Definición de estrategia de pre-procesamiento de datos
- Semestre 2015-1
 4. Construcción de la red de nodos
 5. Análisis de la red
 6. Análisis espacial
 7. Análisis de los resultados
 8. Definir visualizaciones y nivel de interactividad de cada una
 9. Diseño de aplicación de visualización

Es importante decir que el punto 3 se espera abordarlo de manera parcial, realizando un acercamiento durante este período para luego finalizarlo previo inicio del segundo y así poder lograr el desarrollo del resto.

5. Revisión de antecedentes

5.1. Análisis de bibliografía

Esta memoria se basa principalmente en la metodología propuesta por Cheng Zhong et al. (2014) [3], la cual provee un método cuantitativo para la detección de **Centros de flujo**, **Centros** y **Bordes** pudiendo identificar estructuras urbanas a partir de datos pasivos de transporte público. Dentro de este mismo documento se establece una vinculación entre los centros de flujo, centros y bordes con fenómenos urbanos reales dado que utiliza un grafo generado a partir de datos obtenidos del comportamiento de la gente. Además permite aplicar nuevas técnicas para detección de bordes basadas en otras metodologías que puedan aparecer en el futuro, permitiendo la comparación entre ellas. Por lo tanto, la metodología allí expuesta es utilizada en esta memoria para la generación de las mismas estructuras pero para ser analizada con los datos locales de transporte público.

5.1.1. Redes espaciales

Una red espacial se entiende como un grafo cuyos vértices y arcos representan objetos geométricos del mundo real. Los nodos tienen una posición relativa a un sistema de referencia específico y los arcos expresan la forma física en que interactúan entre ellos, entendiéndose esto último como la forma en que se puede llegar físicamente de uno a otro.

Hace algunos años atrás los análisis espaciales urbanos se limitaban a utilizar el diseño de las calles en términos de su topología urbana, lo que tiene la limitante de no considerar la accesibilidad asociada a una calle como una característica dependiente de los movimientos humanos existentes. Además, este tipo de análisis tiende a ignorar los flujos urbanos y a justificar espacios y su forma en función de las propiedades de la red.

En los últimos años, los estudios sobre estas redes comenzaron a incorporar medidas de peso que reflejan los datos de movimientos urbanos como flujos sobre la red pero concentrado en el sistema de tránsito, no sobre los espacios urbanos asociados a estos. Además de los datos obtenidos a partir del transporte público han existido investigaciones basadas en otras fuentes de datos, como lo son los AVL basado en GPS (*Global Positioning System*) o conjuntos de datos telefónicos. En particular, el artículo de Cheng Zhong et al. (2014) utiliza los datos generados a partir del uso de tarjetas inteligentes

del transporte público de singapur.

5.1.2. Construcción y representación de la red

Formalmente definimos un grafo dirigido con pesos como $G = (N, L, W)$ que representa todas las etapas realizadas sobre cada para de paraderos durante una semana. N denota los paraderos o nodos que representa el área o sector donde está ubicado, el conjunto L denota los traslados entre dos paraderos o áreas, por lo que L corresponde a un conjunto de pares ordenados de N , y el conjunto W denota el volumen (cantidad) de traslados entre dos paraderos. Según lo anterior, N son los nodos del grafo, L representa los arcos y W denota los pesos de cada arco en L .

Por otro lado, Munizaga y Palma (2011) [4] (explicado en la sección 5.2) proponen una metodología para crear una *matriz Origen-Destino*, generando la oportunidad de modelar la construcción teórica descrita en el párrafo anterior en sistemas donde solo se valida en un sentido, como lo es en el caso de Chile.

5.1.3. Análisis complejo de la red

Este análisis se abarca desde 3 perspectivas: propiedades globales, información local asociada a *centros* y *centros de flujo*, y en la detección de comunidades.

Propiedades globales Las propiedades topológicas de un grafo puede proveer importante información sobre las interacciones espaciales que se producen en el modelo real, según esto, se define lo siguiente

- El número N de nodos indica cuantos paraderos o áreas son accesibles, y el número de arcos J indica cuantos paraderos o áreas son directamente conectadas.
- El **grado** de cada nodo en la red indica cuantos paraderos o áreas son directamente conectadas desde una en particular, aquí podemos diferenciar entre **grado de salida** (cantidad de nodos que tienen traslados cuyo origen es ese paradero) y **grado de entrada** (cantidad de nodos que tienen como destino ese paradero).

- El **peso** de cada arco indica la intensidad (volumen) de traslados desde un paradero a otro.
- La **ruta más corta** se refiere al camino mínimo posible de un área a otra.
- clustering centrality es un índice que mide cuán cohesionado están los nodos a uno nodo en términos de su accesibilidad.
- closeness centrality es un índice que evalúa cuan rápido se expande la información en el grafo.

the clustering centrality is an index that measures how ‘close’/‘cohesive’ the areas are to one another in terms of their accessibility to shared neighbors; and

the closeness centrality is an index that evaluates how fast information spreads in the whole area.

Mejorar
des-
crip-
ción

Mejorar
des-
crip-
ción

Estas propiedades permiten descubrir los niveles de actividad que mantiene cada área de una ciudad y su participación en los flujos que se realizan.

Centralidades A las propiedades generales agregamos 2 tipos de centralidad adicionales, una es la *centralidad de intermediación*, la que se usará para definir los **centros de flujo** y el segundo es el *PageRank* que mide la accesibilidad en la red tomando en cuenta todos los vínculos, directos e indirectos, sus pesos y dirección. Este último indicador para medir el grado en que cada nodo es un **centro**.

La centralidad de intermediación es un indicador que mide cuán bien conectada está un área o paradero, formalmente se define para un nodo k como el número el número de caminos más cortos que conecta dos áreas i y j en el grafo que pasan a través del nodo k , y se define como

$$C_{intermediacion}(k) = \sum_{ij} \frac{\delta_{ij}(k)}{\delta_{ij}}$$

donde $\delta_{ij}(k)$ es el número de caminos más cortos entre i y j que pasan por k , mientras que δ_{ij} es el número total de caminos entre i y j . En ciertos casos esta medida se normaliza dividiéndola por N (cantidad de nodos) lo cual ellos no realizaron, utilizando la forma tal cual aparece en la definición.

Por otro lado, PageRank mide el rol de un nodo o área local en atraer flujos desde todos los nodos de la red. Esta medida es una representación genérica de las probabilidades de un peatón cualquiera de visitar un nodo cualquiera, en este sentido está relacionado con procesos de Markov de primer orden, que es la base de muchos procesos de interacción social, en este contexto fue originalmente usado para extraer información acerca de las estructuras de los vínculos de Internet (Rosvall y Bergstrom 2008)[7], muy similar al *ranking de páginas de Google*. Según lo anterior se define la probabilidad r_j de visitar el nodo j como:

$$r_j = [(1 - \rho)/N] + \rho \sum_i r_i p_{ij}$$

Donde $1 - \rho$ puede ser visto como la probabilidad de que un caminante decida quedarse en el nodo j , y p_{ij} como la probabilidad de escoger ir al nodo j dado que estoy en el nodo i , este valor es proporcional al peso del arco de i a j , en resumen

$$p_{ij} = w_{ij} / \sum_k w_{ik}, \text{ y } \sum_j p_{ij} = 1$$

El parámetro ρ es conocido como *factor de amortiguación* y toma valores entre 0 y 1, en esta oportunidad será fijado en 0,85. Si $\rho = 1$ entonces todos los nodos tienen una probabilidad positiva y luego, la matriz $\{p_{ij}\}$ tiene que estar fuertemente conectada.

Estructura de comunidad Los bordes a identificar sobre la superficie a analizar sirven para particionar la estructura espacial y así crear pequeños vecindarios a partir de ésta que denominamos comunidades. Estos son obtenidos a partir de la detección de *estructura de comunidad*, que se refiere a una **propiedad de un grafo que permite agrupar nodos de éste que están densamente conectados entre ellos en comparación con el resto de nodos del grafo**. Según lo anterior, los bordes son generados a partir de un descriptor de bordes que particiona la red en dos niveles donde los nodos forman módulos que llamamos comunidades y la división entre los módulos que llamamos bordes. Existen varias formas de generar comunidades pero una condición necesaria para este trabajo es la consideración de las variables de **densidad** y **flujo de interacciones** al momento de crear éstas. Lo anterior basado en que estas dos variables sean mas fuertes dentro de una

comunidad y que el volumen que está dentro de cada comunidad es mayor en comparación con el resto de la red.

Para la generación de las comunidades se utilizará el framework *map equation* basado en un procedimiento llamado *infomap* desarrollado por Rosvall y Bergstrom el 2008 [?]. Es uno de los algoritmos que ha mostrado mejor rendimiento para la generación de comunidades y uno de los pocos adecuados para redes con peso y dirección. Otra característica relevante del algoritmo *infomap* es que no solo considera la relación entre pares de nodos sino que también toma en cuenta los flujos presentes entre estos. Para llevar a cabo lo anterior se utilizan flujos probabilísticos creados a partir de generaciones aleatorias que simulan recorridos sobre el grafo y que asignan también probabilidades a cada nodo de ser visitado aleatoriamente (utilizando el algoritmo *PageRank*), con el objetivo de modelar los comportamientos de flujo de un sistema real.

En resumen, el algoritmo divide los nodos del grafo en módulos que son altamente estructurados, lo que implica que la entropía del grafo particionado es mínima. Esta entropía es una subdivisión de la entropía total del sistema distribuida entre los módulos con un peso entropico entre los módulos, esos pesos son relacionados a la probabilidad de ocurrencia de cada módulo. Por lo anterior, Rosvall y Bergstrom define esta entropía como:

$$\left. \begin{aligned} Lg(M) &= H(P) + \sum_{i=1}^m P_i H(p)_i \\ &= -p \sum_{i=1}^m P_i \log P_i - \sum_{i=1}^m P_i \sum_{k=1}^{M_i} \frac{P_k}{P_i} \log \frac{P_k}{P_i} \end{aligned} \right\}, P_i = \sum_k p_k$$

donde P_i es la probabilidad de se visitada la comunidad i y P_k/P_i es la probabilidad de visitar el nodo k dado que la comunidad i es visitada. Por último M_i es la cantidad de nodos que contiene la comunidad i . Por último, La división por P_i que afecta a los P_k cumple la función de normalizar.

La forma en que trabaja es primeramente configurando cada nodo al módulo que pertenece para luego, en cada paso, identificar que nodo se puede agregar a que módulo tal que la entropía general decrezca. Este proceso continua hasta que no se pueda reducir más la entropía, asegurando que con esa configuración se obtiene la partición más estructurada.

Es importante decir que M_i es un módulo que contiene un conjunto de $k \in M_i$ nodos y que llega a ser estable (sin alteraciones) una vez que se obtiene la entropía mínima. Luego, estas comunidades son mapeadas a sus respectivas ubicaciones geográficas.

Revisar párrafo siguiente. Falta explicar porque tiene entropía mínima.

Lo que resta por hacer en este punto es transformar un conjunto de puntos discretos en regiones que particionan el espacio. Para lo anterior se realiza una interpolación espacial, para esto se considera el siguiente supuesto:

“Cada persona escoge el paradero de bus/metro más cercano”

Según lo anterior, aplicamos la interpolación a cada zona geográfica cercana a un paradero de bus/metro. La variante de interpolación escogida corresponde a la *Inverse Distance Weighting (IDW)* lo que hace disminuir la influencia de un paradero de bus en función de la distancia. Estos pesos son definidos de la siguiente forma

$$W_i(x, y) = d_{ij}(x, y)^{\lambda}$$

Donde $W_i(x, y)$ es el peso de la ubicación del paradero i en las coordenadas x, y que son los puntos vecinos más cercanos a j y $d_{ij}(x, y)$ es la distancia a la coordenada x, y desde el paradero i hacia el paradero vecino más cercano j .

Una observación que se desprende de la fórmula es que los pesos están normalizados tal que su suma sea 1, es decir, $\sum_{\forall x, y} W_i(x, y) = 1$ y λ es un parámetro arbitrario que en este caso es 2, lo que implica que sigue la ley del inverso al cuadrado.

Por último cada coordenada espacial debe ser asignada a una única comunidad, esto se hace utilizando un *resumen estadísticos*⁶. El principal problema aquí es tratar aquellas coordenadas que pertenecen a una comunidad en la red pero que no están geográficamente adyacentes al grupo principal que define la comunidad, esto se origina porque el algoritmo de detección de comunidades no está restringido a lograr áreas geográficamente contiguas.

Aunque la investigación reporta que esto no ocurre muy a menudo, y que cuando sucede, es en los límites de las áreas que definen las comunidades y se da principalmente porque las personas que viven en esas áreas tienen diferentes preferencias de viajes.

Para solucionar el problema planteado se cuenta el número de puntos en las comunidades en conflicto y se computa el algoritmo *PageRank*, esto provoca que las coordenadas en disputa sean asignadas a su comunidad más cercana geográficamente. En la práctica lo que se ocurre es un desplazamiento

⁶Se entiende como la información dada por una rápida y simple descripción de los datos como la media, mediana, moda, rango y desviación estándar.

de los límites entre comunidades, de esta forma se obtiene una partición geográfica que abarca todo el espacio.

5.1.4. Visualización de información

En esta sección serán descritos algunos aspectos relevantes de la visualización de información con el fin de justificar teóricamente preguntas como: ¿que colores utilizar? ¿que tipo de visualización mostrar? ¿cuál es el objetivo de la visualización que estoy creando? entre otras.

Toda la información necesaria será obtenida del libro *Interactive Data Visualization: Foundations, Techniques, and Applications* [5].

Falta documentar toda esta sección.

5.2. Análisis de datos

A continuación se detalla la forma en que funciona el sistema de transporte público que origina los datos, como se estima el proceso de bajada y por último los detalles de las tablas obtenidas posterior a la estimación. Tanto la metodología como los supuestos aquí explicados son expuestos y propuestos por Munizaga y Palma (2011) [4].

5.2.1. Sistema tarifario

En Santiago de Chile, el sistema AFC utilizado corresponde a las tarjetas de pago, donde en buses es el único método disponible y en metro es el más utilizado.

El sistema de pago en Transantiago es tal que cada pasajero paga una tarifa cuando accede al sistema, que permite a él o ella hacer tres transbordos dentro de las dos horas siguientes al pago. La estructura de pago es diferente entre el metro y buses. En buses, el único sistema de pago es mediante la tarjeta de pago (llamada comercialmente tarjeta bip!), mientras que en metro, es posible comprar un ticket o usar la tarjeta bip!, sin embargo el porcentaje de usuarios que compra el ticket es de aproximadamente 3 %.

El sistema se caracteriza por tener cerca de 300 rutas de buses, 6000 buses disponibles agrupados en 6 operadores⁷, aproximadamente 10.000 paraderos y una cifra que bordea los 150 kilómetros de rieles para el metro.

⁷Un operador es una empresa que se encarga de prestar servicio a una zona de santiago.

actualizar con datos más recientes el siguiente párrafo.

Dada la alta demanda que ha experimentado, se crearon 150 zonas físicas llamadas “zona paga” que están equipadas con sistema de pago (validadores) de vehículo donde el pasajero paga cuando entra a la estación, lo cual incrementa la eficiencia de las subidas a los buses pero genera una dificultad para determinar cual bus de todos los que allí se detienen tomó. Es importante decir que estas estaciones de buses operan durante los horarios de alta demanda en puntos de congestión identificados previamente.

Todas las transacciones bip! Son guardadas en una base de datos que contiene información sobre los operadores y el instante en que la transacción fue hecha. Lo anterior se lleva a cabo por cada pasajero acercando su tarjeta al validador, cuando ingresa al bus, zona paga o metro. Cada validador adjunta a cada transacción que realiza un id asociado a el y que está a su vez, asociado con un bus, estación de bus o metro. La información recolectada por cada transacción incluye: **id de la tarjeta, tipo⁸, código de bus o sitio donde se realizó la transacción, fecha y hora, monto de pago**. La posición espacial de la transacción puede ser conocida directamente para las zonas paga y las estaciones de metro dado que son conocidas con anterioridad, para las transacciones hechas en buses es posible pero no está disponible en la base de datos de transacciones.

Otra base de datos contiene información sobre la localización de todos los buses, como la **latitud, longitud, tiempo, fecha y velocidad instantánea**. Estos datos obtenidos a intervalos de 30 segundos y son asociados a cada bus a través de un número de placa y código de operador.

Cruzando la información de transacción y posición de las bases de datos por cada placa de bus o código de metro/estación de bus y tiempo, es posible identificar la localización espacial donde la transacción es realizada. Es así como en datos analizados del año 2009 y 2010 se logra una estimación exitosa en el 98,5 % y 99,9 % de los casos respectivamente.[4].

5.2.2. Estimación de bajadas

Como en el sistema de tarifa solo se validan las subidas, es necesario estimar los puntos de bajadas de las transacciones. Es aquí donde se utilizan una serie de supuestos para entender el comportamiento general de los usuarios dentro del sistema. Para lo anterior debemos definir lo que se entiende como un viaje, y *Ortúzar and willumsen, 2011*[6] lo definen como:

⁸puede ser comercial o estudiante.

“Un viaje se define como un movimiento desde un punto de origen a un punto de destino”

Esta definición origina la consideración de etapas dentro de un viaje, una etapa es la utilización de un servicio en particular (bus o metro). Es importante notar que no se consideran los cambios entre líneas de metro.

Básicamente la idea es seguir una cadena de viajes de una tarjeta e identificar la posición de bajada (de bus o metro) mirando la posición y el tiempo de la próxima subida de esta tarjeta. Esto es solamente posible cuando la actual y siguiente transacción tiene información de posición, la cual es tomada de la base de datos de localización automática de vehículos. En el caso de la última transacción del día, se asume que el destino es cercano al punto donde el primer viaje del día comienza, encontrando así un viaje cíclico diario para los usuarios particulares. Si hay solo un viaje por tarjeta, no es posible inferir con solo un día de información.

Los supuestos para llevar a cabo esta estimación son:

- Después de un viaje, el origen del siguiente determina el destino del primero. [2]
- al final del día, los usuarios van a volver a la estación donde abordaron en el primer viaje del mismo día. [2]
- Cada tarjeta corresponde a un usuario. [4]
- Se asume que una persona camina hasta la siguiente parada un máximo de 1.000 metros [4]
- Si el tiempo de transbordo⁹ es inferior a 30 minutos, entonces este último servicio forma parte del mismo viaje, de lo contrario se considera como uno nuevo.[4]

Según todo lo anterior Munizaga y Palma (2011)[4] son capaces de estimar correctamente cerca del 80 % de los datos utilizados. Dado que es necesario conocer esta información, se restringe la data al porcentaje de datos que forman parte del resultado exitoso.

⁹entendido como el tiempo entre que se bajó de un servicio y se subió al siguiente.

Nombre campo	Descripción
id	identificador de la tarjeta que realizó la validación
nviaje	Indica el número de viaje asociado al id de la tarjeta de pago
netapa	lugar que ocupa la etapa dentro de un viaje.
par_subida	Indica el paradero en donde abordo el servicio.
par_bajada	Señala el paradero de donde descendió del servicio.

Tabla 1: Atributos a utilizar de la tabla ETAPAS

:

5.2.3. Descripción de los datos

El procedimiento anterior ya se encuentra realizado para los datos comprendidos entre el 14 de abril de 2013 al 20 de abril de 2013 por lo que será este tramo el que será utilizado para realizar el análisis espacial y dado que es requerido conocer cada etapa que realiza una persona, se opta por utilizar la tabla de etapas en donde una fila representa una etapa de un viaje determinado. Esta tabla actualmente contiene 42 columnas producto de diversos análisis que se han realizado con ellas pero que no son todos útiles para el desarrollo de esta memoria por lo que se omiten algunos. A continuación se listan los campos a ser utilizados para el procesamiento:

No se descarta que una vez realizada la etapa de visualizar los datos puedan requerirse más datos con respecto a cada etapa, como puede ser la fecha y hora, tipo de transporte, entre otros.

Además de la tabla de etapas es necesario conocer todos los paraderos disponibles con el fin de poder realizar la interpolación, por lo que también se utiliza la tabla en conjunto con la tabla que contiene los datos de las estaciones de metro, que para efectos de este análisis son considerados paraderos adicionales.

nombre
de
tabla
de
para-
deros

nombre
de la
tabla
de
esta-
ciones
de
metro

5.2.4. Conclusiones

Este trabajo al igual que el artículo en el que está basado, utiliza los datos generados a partir del uso de tarjetas inteligentes del transporte público, con la salvedad de que aquí son de Santiago de Chile para llevar la creación de la red espacial con el objetivo de analizar estructuras de movimiento urbano.

logramos obtener el destino. Con lo anterior se construirá una donde se registran los volúmenes de etapas realizadas. Esta matriz será convertida posteriormente en un grafo dirigido con pesos. Es importante destacar que el grafo formado no es una matriz de viajes sino que una red social formada por actividades urbanas.

Nuestro caso tampoco normalizará la definición de centralidad intermedia pero para este caso se utilizará como aparece en la definición.

para el PageRank se decide esto, la medida que se ocupará en esta memoria es la misma que se ellos ocuparon para determinar la importancia de un nodo en una red

Hablar de que se ha mejorado el porcentaje de estimación exitosa.

6. Estrategia de procesamiento

En esta sección se describirán las acciones a llevar a cabo en el procesamiento de la data, esto consiste en describir la forma en la que se llevará a cabo el procedimiento para elaborar el grafo y calcular los distintos indicadores descritos en la sección 5.1. Para lo anterior es necesario primero detallar las herramientas que se utilizarán para luego pasar a describir

6.1. Herramientas a utilizar

La base de datos que contiene las etapas se encuentra en un Sistema Gestor de Base de Datos (*SGBD*) **Postgres** versión 9.1 y para gestionar los datos se posee una Graphic Unit Interface (*GUI*) llamada *PgAdmin III* cuya versión es 1.18.1 con la cuál se manipulan los distintos requerimientos al SGBD.

6.2. Filtrar la data

7. Bibliografía

- [1] Alex Anas, Richard Arnott, Kenneth A. Small. Urban spatial structure. *Journal of Economic Literature*, 36:1426–1464, 1998.
- [2] Barry, J.J., Newhouser, R. Rahbee, A., Sayeda, S. Origin and destination estimation in new york city with automated fare system data. *Transportation research record*, 1817:183–187, 2002.
- [3] Chen Zhong, Stefan Müller Arisona, Xianfeng Huang, Michael Batty & Gerhard Schmitt. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 50:1–21, 2014.
- [4] Marcela A. Munizaga, Carolina Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smart-card data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18, 2012.
- [5] Matthew O. Ward, Georges Grinstein, Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, 2010.
- [6] Ortuzar, J. de D., Willumsen, L.G. *Modelling transport*. Wiley, Chichester, 2011.
- [7] Rosvall, M. y Bergstrom, C.T. Maps of random walks on complex networks. *PloS One*, 5, 2008.