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

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

8 de diciembre de 2014

Profesor Guía
Marcela Munizaga

Profesor Co-guía
Benjamín Bustos

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

Índice

Índice de tablas

Índice de figuras

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 paraderos[1] que permitan registrar el correcto descuento del valor asociado al pasaje, concepto que llamaremos validación.

Transantiago[2] 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 laboral[3], 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 (2012) [7] que logra una identificación 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

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

[2] fue implementado a partir del año 2007.

[3] Lunes, martes, miércoles, jueves o viernes.

estructuras espaciales urbanas [4]:

- 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 de manera sintética dado que comprende una gran cantidad de datos y además abarcará otros 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, las herramientas de procesamiento actuales no logran obtener toda la información que los datos proveen.

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.

Esta memoria creará las estructuras espaciales mencionadas anteriormente a partir de los datos de transporte público de Santiago de Chile y diseñará una aplicación para que puedan ser visualizadas, de manera de comunicar información existente en los millones de datos provistos y que las herramientas actuales de procesamiento no abarcan.

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* [4], 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 (2012)[7], por lo que los datos corresponden a una tabla de una base de datos postgreSQL llamada *tabla_de_etapas* donde cada fila representa una etapa de un viaje[4]. 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 AFC[5].

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 los datos.

   Se realizará un estudio de los datos existentes para comprender concretamente las bases de datos 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 item anterior. Dentro de esta etapa se llevará a cabo la normalización y selección de los datos 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 los estudiados.

[4] un viaje puede tener una o más etapas.

[5] Automatic 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 los datos
    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 antecendentes

5.1. Análisis de bibliografía

Esta memoria se basa principalmente en la metodología propuesta por Cheng Zhong et al. (2014) [4], 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. El termino "complejo"se anexa al término de red espacial cuando el grafo de esta red se caracteriza por tener propiedades estructurales que se relacionan con la accesibilidad de un nodo y su aporte dentro de la red. Estas propiedades tienen la particularidad de estar presentes en muchos problemas reales (Doursat 2005) [5].

Hace algunos años atrás los análisis espaciales urbanos se limitaban a utilizar el diseño de las calles en términos de su topología urbana (Cardillo et al. 2006)[3], 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 estás redes (Soh et al. 2010) [14] 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 (Rinzivillo et al. 2012)[11] o conjuntos de datos telefónicos (Ratti et al. 2010)[10]. 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 de viajes realizadas. N denota los paraderos existentes, a éste se asocia 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 (2012) [7] (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 Santiago 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 topologicas 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 mínmo posible de un área a otra.
- **"clustering centrality"**es un índice que mide cuán cohesionados o cercanos están los nodos a otro en términos de su accesibilidad para compartir vecinos.
- **Centralidad de cercanía** es un índice que evalúa cuan rápido viaja la información en un grafo.

Buscar traducción a español

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 será útil 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 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.

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 puede ser vista como

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 (Brin y Page 1998)[13], 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)[12], muy similar al *ranking de páginas de Google*. Según lo anterior de 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 el estudio de Chen Zhong et al. (2014) fue 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 una *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 estos 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 deben ser 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 utiliza el framework *map equation* basado en un procedimiento llamado *infomap* desarrollado por Rosvall y Bergstrom el 2008 [12]. Lo anterior se justifica por Lancichinetti y Fortunato (2009)[6], donde concluyen que 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 probabilisticos creados a partir de generaciones aleatoreas 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 (Rosvall y Bergstrom, 2008)[12]. Esta entropía total del sistema está dividida en la entropía de moverse entre los módulos más en la entropía de moverse dentro de un módulo, las proporciones son relacionados a la probabilidad de ocurrencia de cada uno. Por lo anterior, Rosvall y Bergstrom (2008) 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 ser visitada la comunidad i, p es la probabilidad que un caminante cualquiera de cambiarse de modulo y P_k es la probabilidad de visitar el nodo k. Además M_i es la cantidad de nodos que contiene la comunidad i y la división por P_i que afecta a los P_k cumple la función de normalizar.

La forma en que trabaja es primeramente asociando cada nodo al módulo que pertenece, donde en cada paso, se identifica 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 (Rosvall y Bergstrom, 2008). Luego, estas comunidades son mapeadas a sus respectivas ubicaciones geográficas.

Con lo anterior ya hecho, lo que resta por hacer es transformar los conjuntos de puntos discretos obtenidos por *PageRank* en regiones que particionan el espacio geográfico, para esto se realiza una interpolación espacial que considera el siguiente supuesto:

"Cada persona escoge el paradero de bus/metro más cercano a la posición en la que se encuentra"

Según lo anterior, Chen Zhong et al. (2014) propone aplicar una 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

$$W_i(x, y) = 1/d_{ij}(x, y)^{\lambda}$$

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 lo logran 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 limites de las áreas que definen las comunidades y se da principalmente porque las personas que viven en esas áreas tienen diferentes preferencias de viajes.

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

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 practica lo que se ocurre es un desplazamiento de los límites entre comunidades, de esta forma se obtiene una partición geográfica que abarca todo el espacio.

5.1.4. Georeferenciación espacial

La georeferenciación espacial se refiere a un sistema de referencia que se utiliza para representar las ubicaciones de objetos o fenómenos dentro de un marco geográfico común. Existen varios y cada uno queda definido por: punto de referencia, unidad de medición, entre otras [**?**].

Existen dos tipos de sistemas de coordenadas que se utilizan ampliamente:

latitud-longitud son los denominados sistemas de coordenadas geográficas

proyectado

Los datos geoespaciales a diferencia de otros tipos de datos describen objetos o fenómenos con una ubicación específica en el mundo real.

En esta sección se debe explicar que es la georeferenciación, para que sirve, su definición formal y que es lo que se puede lograr con ella.

Falta completar.

5.1.5. 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 es la visualización? ¿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* [8].

Falta documentar toda esta sección.

¿Que es la visualización? El concepto de visualización se define como la *comunicación de información usando representaciones gráficas*. El beneficio de utilizar imágenes está dado por la riqueza de información que puede contener y el tiempo que requiere para ser procesado en vez de información en páginas con texto, esto ocurre debido a que la interpretación de imágenes es realizada en paralelo dentro del sistema perceptual mientras que la lectura

es un proceso secuencial. Otra de las ventajas que tiene la visualización es su independencia del lenguaje, siendo más eficaz que un texto para un grupo de personas que hablan diferentes idiomas.

La importancia de esta forma de comunicación radica en que los seres humanos son seres visuales que usan la visión como uno de los sentidos claves para entender la información, según esto es que se justifica el hecho de que sea ocupada en muchos lugares en la toma de decisiones, y dado el incremento de datos presentes hoy en día en cada área del conocimiento, es que hay una creciente necesidad por herramientas y técnicas que ayuden a convertirlos en información útil.

La visualización no es efectiva *per se* sino que depende de varios factores su impacto. En 1999 Linda Elting demostró por medio de un experimento que la visualización afecta el proceso de decisión y que está muy relacionada con las preferencias y nivel de entrenamiento de los usuarios involucrados.

Visualización hoy en día La visualización a menudo ofrece distintos niveles de vistas de información, éstas pueden ser vistas **cualitativas** o **cuantitativas** dependiendo de que parte de la información existente es la relevante de ser mostrada. Un ejemplo de esto se observa en la figura 5.1.5, que corresponde a la red de metro subterráneo de la ciudad de Tokio, en el se produce una distorsión de los puntos en donde se interceptan varias líneas para facilitar su interpretación.

Las imagenes distorsionadas tienden a ser imprecisas y dependen de quien las observa. La mayoria de los mapas bidimensionales exhibe algún grado de distorsión debido a la proyección que se debe hacer desde un volúmen a un plano (3D a 2D)

Técnicas para datos geoespaciales asdas

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 (2012) [7].

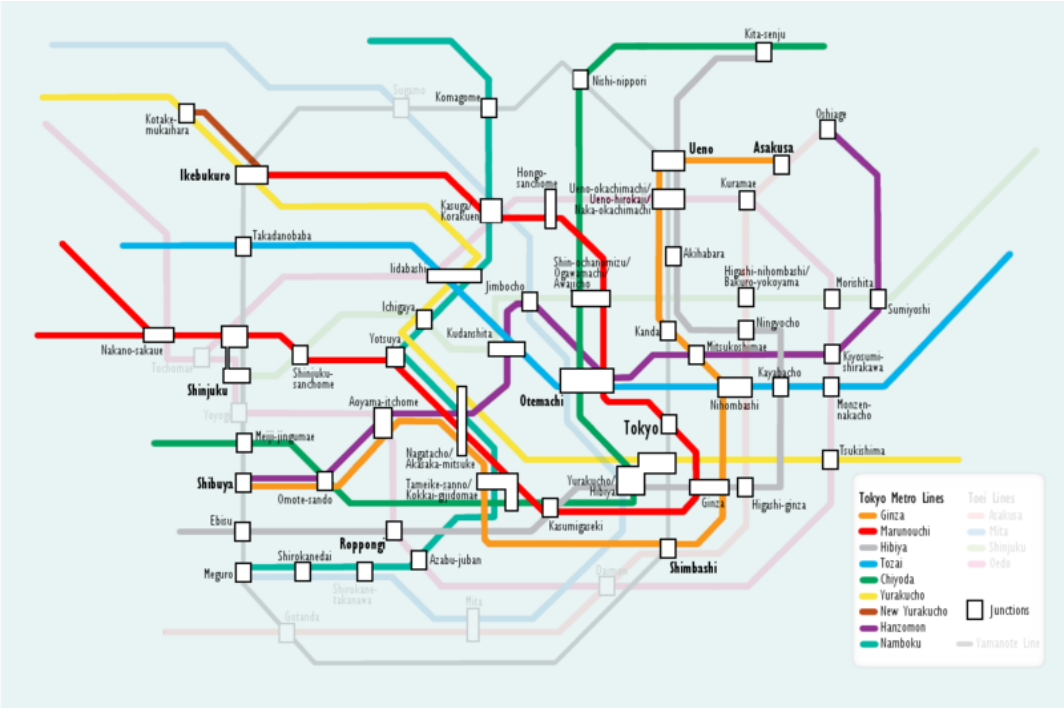

Figura 1: Mapa de la red de metro subterráneo de la ciudad de tokio. Representación cualitativa de las estaciones y las líneas que pasan por ella. (obtenida desde *Wikimedia Commons*)

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 operadores[7], aproximadamente 10.000 paraderos y una cifra que bordea los 150 kilómetros de rieles para el metro.

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 un usuario. 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, zona paga o metro. La información recolectada por cada transacción incluye: **id de la tarjeta**, **tipo**[8], **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 son 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/zona paga 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 en el 98,5 % y 99,9 % de los casos respectivamente.[7].

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 Munizaga y

[7]Un operador es una empresa que se encarga de prestar servicio a una zona de santiago.
[8]puede ser comercial o estudiante.

Palma (2012)[7] utilizan una serie de supuestos para entender el comportamiento general de los usuarios dentro del sistema. Para lo anterior utilizan la definición de viaje dada por *Ortúzar y Willumsen, 2011*[9]:

"Un viaje se define como un movimiento desde un punto de origen a un punto de destino donde se realiza una actividad"

Esta definición origina la consideración de etapas dentro de un viaje, una etapa es la utilización de un servicio en particular (bus o metro), entonces la combinación de estos hecha hasta que el usuario llegue al lugar donde debe llevar a cabo su actividad es lo que entendemos como un viaje. 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. [7]
- Se asume que una persona camina hasta la siguiente parada un máximo de 1.000 metros [7]
- Si el tiempo de transbordo[9] es inferior a 30 minutos, entonces este último servicio forma parte del mismo viaje, de lo contrario se considera como uno nuevo.[7]

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

Según todo lo anterior Munizaga y Palma (2012)[7] son capaces de estimar cerca del 80 % de los datos utilizados. Dado que es necesario conocer esta información, se restringe los datos al porcentaje de datos que forman parte del resultado exitoso.

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 omite la mayoría. En la tabla 1 se listan los campos a ser utilizados para el procesamiento de los datos.

Nombre campo	Descripción
id	identificador de la tarjeta que realizó la validación
nviaje	Indica el número de viaje asociado al id de la tarteja 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

:

No se descarta que una vez realizada la etapa de visualización 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 **redpadaras**(ver tabla 2) en conjunto con la tabla **estaciones_metro** (ver tabla 3) que contiene los datos de las estaciones de metro, éstos serán tratados de manera indistinta ya que tanto los paraderos de bus como las estaciones de metro funcionan como puntos de origen o destino en una etapa.

Nombre campo	Descripción
codigousuario	identificador único de paradero, ej:PJ156
x	posición en eje horizontal dentro del sistema de coordenadas cartesiana utilizado para la georeferenciación
y	posición en eje vertical dentro del sistema de coordenadas cartesiana utilizado para la georeferenciación

Tabla 2: Atributos a utilizar de la tabla REDPARADAS

:

explicar tabla paradero

Nombre campo	Descripción
codigosinlinea	Nombre de la estación de metro
x	posición en eje horizontal dentro del sistema de coordenadas cartesiana utilizado para la georeferenciación
y	posición en eje vertical dentro del sistema de coordenadas cartesiana utilizado para la georeferenciación

Tabla 3: Atributos a utilizar de la tabla ESTACION_METRO

:

explicar tabla est. metro

5.3. Conclusiones

Es importante destacar que la particularidad del estudio radica en que el grafo formado para construir las estructuras espaciales no representa viajes sino actividades urbanas que realizan las personas, construyendo así una red social formada por actividades urbanas.

Tanto las definiciones de propiedades globales como las formulas definidas aquí serán ocupadas tal cual aparecen de manera de más adelante poder comparar los resultados obtenidos con los mostrados en el artículo, que está basado en el sistema de transporte público de Singapur.

El *factor de amortiguación* será ajustado al propuesto por Chen Zhong et al. (2014)[4], no es campo de esta memoria el considerar variaciones de este factor, por lo tanto será fijado en 0,85.

Se utilizará el framework *map equation* para llevar a cabo la generación de comunidades dado que plantea el mejor desempeño según Lancichinetti y Fortunato (2009) [6].

La variante de interpolación escogida, al igual que en Chen Zhong et.al (2014) corresponde a la *Inverse Distance Weighting*(IDW) con un $\lambda = 2$ dado que no se hacen supuestos adicionales a los propuestos.

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 corresponden al sistema de transporte de Santiago de Chile, con estos se construirá la red espacial para la posterior construcción de las estructuras descritas (centros, centros de flujo y bordes).

explicar que se usará georeferenciación, que estándar, etc..

Como los datos de Santiago de Chile presentan las mismas características luego de aplicar el método de estimación de Munizaga y Palma (2012), se observa que estos son aptos para replicar el análisis hecho en Singapur. Es importante decir que este procesamiento fue realizado con anterioridad a este trabajo y aquí se documenta con fines explicativos.

La efectividad de la visualización depende de las preferencias y nivel de entrenamiento de los usuarios por lo que se debe considerar la elaboración del perfil de los usuarios que usarán la aplicación.

A partir de todo lo expuesto, el aporte de este trabajo a lo ya existente se resume en:

Hablar que es lo que se hará de la sección de la visualización.

- Validación del modelo propuesto por Cheng Zhong et al. (2014) [4] por

medio de la aplicación de su metodología a otra fuente de datos real (sistema de transporte público de Santiago de Chile).

- Diseño de una herramienta de visualización que permite observar los resultados obtenidos por la metodología de Cheng Zhong et al.
- Aumentar la cantidad de información provista por los datos pasivos del transporte público de Santiago de Chile.

6. Estrategia de procesamiento

En esta sección se describen las acciones a llevar a cabo en el procesamiento de los datos, 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 interfaz gráfica (*GUI*) llamada *PgAdmin III* cuya versión es 1.18.1 con la cuál se manipulan los distintos requerimientos al SGBD.

6.2. Filtrado de datos

En este paso se deben quitar aquellos datos que no cumplen con el requisito de tener un paradero de bajada asociado, para esto se extraen los datos que cumplen con la siguiente consulta en lenguaje *SQL* (*Structure Query Language*)

```
SELECT id, nviaje, netapa, par_subida,
    par_bajada
FROM etapa
WHERE par_bajada IS NOT NULL;
```

Figura 2: Consulta para quitar etapas sin paradero de bajada.

Los resultados de la consulta anterior son almacenados en una nueva tabla que contiene solo los campos a ser utilizados (ver figura 2).

Falta describir todo esto.

6.3. Cálculo de centralidades

6.3.1. Centralidad de intermediación

6.3.2. PageRank

6.4. Cálculo de Bordes

6.5. Generación de comunidades

6.6. Interpolación espacial

6.7. Corrección de bordes

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 systen data. *Transportation research record*, 1817:183–187, 2002.

[3] Cardillo, A., et al. Structural properties of planar graphs of urban street patterns. *Physical Review*, 73, 2006.

[4] 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.

[5] Doursat, R. Topology and dynamics of complex networks. *In: CS 790R Seminar modeling & simulation. Department of Computer Science & Engineering University of Nevada, Reno, Spring*, 2005.

[6] Lancichinetti, A. and Fortunato, S. Community detection algorithms: a comparative analysis. *Physical review E*, 2009.

[7] Marcela A. Munizaga, Carolina Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18, 2012.

[8] Matthew O. Ward, Georges Grinstein, Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, 2010.

[9] Ortuzar, J. de D., Willumsen, L.G. *Modelling transport*. Wiley, Chichester, 2011.

[10] Ratti, C., et al. Redrawing the map of great britain from a network of human interactions. *PloS One*, 5, 2010.

[11] Rinzivillo, S., et al. Discovering the geographical borders of human mobility. *KI-Künstliche Intelligenz*, 26:253–260, 2012.

[12] Rosvall, M. y Bergstrom, C.T. Maps of random walks on complex networks. *PloS One*, 5, 2008.

[13] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[14] Soh, H., et al. Weighted complex network analysis of travel routes on the singapore public transportation system. *Physica A: Statistical Mechanics and its Applications*, 389:5852–5863, 2010.