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

23 de noviembre de 2014

Profesor Guía
Marcela Munizaga

Profesor Co-guía
Benjamín Bustos

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

1. Resumen ejecutivo

Aquí va el resumen ejecutivo

Índice

2. 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 [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

[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 [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.

3. 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.

4. Objetivos

4.1. 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."

4.2. 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.

5. 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 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 la data.

   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 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.

[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 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.

6. Revisión de antecendentes

6.1. Análisis de bibliografía

El documento principal en el cuál está basado esta memoria es el artículo *Detecting the dynamics of urban structure through spatial network analysis* [3] por lo que la metodología allí expuesta es utilizada en esta memoria para ser analizada con los datos locales de transporte público. El aporte de este documento al área radica en que provee un método cuantitativo para la detección de **Centro de flujo**, **Centros** y **Bordes** pudiendo detectar estructuras urbanas a partir de datos pasivos del transporte público, además de describir las técnicas que son aplicadas. También establece una vinculación entre parámetros medibles con fenómenos urbanos reales, la cuál es posible aplicar a nuevas técnicas para detección de bordes basadas en otras metodologías que puedan aparecer en el futuro, permitiendo la comparación entre ellas.

6.1.1. 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 interactuan entre ellos, entendiendose 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 estás 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. Este trabajo al igual que el artículo en el que está basado utiliza los datos de la tarjetas inteligentes del transporte público de santiago de Chile para llevar la creación de la red

espacial con el objetivo de analizar estructuras de movimiento urbano.

6.1.2. Construcción y representación

6.1.3. Análisis complejo de redes

-

6.1.4. 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 tal que permite agrupar nodos de éste tal 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 modulos que llamamos comunidades y la división entre los modulos son los bordes. La formación de comunidad se hace basado en que la medida de densidad y el flujo de interacciones es mas fuerte y en términos de 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 [7]. 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 de recprridos sobre el grafo (utilizando el algoritmo *PageRank*) y la probabilidad de visitar un nodo aletario, con el objetivo de modelar comportamientos de flujo de un sistema real.

En resumen, el algoritmo divide los nodos del grafo en modulos que son altamente estructurados, lo que implica la entropía del grafo particionado es mínima. Esta entropía es una subdivisión de la entropía total del sistema distribuida entre los modulos 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 la comunidad m de ser visitado y P_k/P_i es la probabilidad de que el nodo k, que es parte de la comunidad M_i de ser visitado

6.1.5. fenomeno de mundo pequeño

6.2. 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.

6.2.1. 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 %.

actualizar con datos más recientes el siguiente párrafo.

El sistema se caracteriza por tener cerca de 300 rutas de buses, 6000 buses disponibles agrupados en 6 operadores[6], 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. 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**, **tipo**[7], **código de bus o sitio donde**

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

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].

6.2.2. 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:

"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]

Según todo lo anterior sumado a un modelo ...Munizaga y Palma (2002)[4] son capaces de estimar 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.

¿cómo se llama el modelo?

6.2.3. Identificación de actividades

6.2.4. 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 utiliza 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:

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.

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 tambien 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 paraderos

nombre de la tabla de estaciones de metro

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] 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 smartcard 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.