UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERÍA

Minería de procesos para el análisis de movilidad urbana

Informe de Proyecto de Grado presentado por

Bruno Rodao, Nicolás Carignani, Santiago Ferreira

en cumplimiento parcial de los requerimientos para la graduación de la carrera de Ingeniería en Computación de Facultad de Ingeniería de la Universidad de la República

Supervisores

Dra. Ing. Andrea Delgado
Dr. Ing. Daniel Calegari

Montevideo, 23 de marzo de 2023

Resumen

La movilidad urbana plantea diversos desafíos de diseño y gestión para favorecer el desarrollo metropolitano y la sustentabilidad del sistema en toda su diversidad. El Sistema de Transporte Metropolitano (STM) es un cambio orientado a mejorar la movilidad de los ciudadanos, principalmente en el departamento de Montevideo, que entre otras cosas registra información sobre la red de transporte y viajes de sus usuarios. En este marco, existen diversos intereses de la Intendencia de Montevideo (IM) en cuanto a explotar la información disponible para la mejora de la movilidad de sus ciudadanos.

La minería de procesos es una disciplina de la Ciencia de Datos que se basa en técnicas de minería de datos para analizar los registros (logs) de eventos asociados a la ejecución de procesos en un dominio particular, definiendo proceso como un conjunto de actividades relacionadas para cumplir un objetivo específico. Esto permite descubrir modelos de procesos a partir de los eventos, chequear la conformidad de una traza de eventos en un modelo determinado, es decir, qué tanto respeta una ejecución el modelo predefinido, y obtener medidas de ejecución de los procesos como su duración, cuellos de botella, o la subutilización de recursos.

El objetivo principal de este proyecto es explorar la aplicación de técnicas de minería de procesos destinadas a la búsqueda de soluciones a problemas de interés de la IM en relación con la movilidad urbana. Este proyecto se enmarca en el proyecto de investigación "Minería de procesos para el análisis de movilidad urbana" del programa IM – Udelar "Ing. Oscar J. Maggiolo" de CSIC, en ejecución por el grupo COAL del INCO.

Para la aplicación de las técnicas de minería de procesos, se utilizó el software Disco. Este permite realizar descubrimiento de modelos y ofrece una interfaz para su posterior análisis. Como resultado se presentan distintos análisis de movilidad urbana utilizando el enfoque de minería de procesos y como fuente de datos los registros de ascensos de los usuarios del STM. En particular, los análisis de corredores y viajes multitramos de los pasajeros, muestran que la minería de procesos es de utilidad en el análisis de movilidad urbana, y enfatizan en los tipos de resultados que pueden ser obtenidos. Como complemento, se presenta un análisis de inferencia de destinos.

Palabras clave: Minería de procesos, movilidad urbana, GIS, STM, datos abiertos, inferencia destinos

Índice general

Capítulo 1

Introducción

La movilidad urbana es fundamental para el desarrollo y calidad de vida en una ciudad. Proporciona los medios para acceder a lugares de interés como nuestra vivienda y trabajo. El correcto modelado y planificación de la red de transporte exige obtener respuestas a interrogantes como: ¿cómo es el movimiento de personas a lo largo del día?, ¿qué zonas son más propensas a tener embotellamientos?, ¿cuáles son los destinos más frecuentes?, ¿cómo es la accesibilidad a la red de transporte?.

Los Sistemas Inteligentes de Transporte (ITS) (Sussman, 2005) permiten realizar análisis del transporte y del comportamiento de los usuarios, ya que hacen uso de dispositivos físicos y sistemas de información, que en conjunto permiten extraer datos que serán procesados para obtener nuevo conocimiento de la red.

El Sistema de Transporte Metropolitano (STM), fue definido en 2010 por la Intendencia de Montevideo (IM) en el Plan de Movilidad[1], y está conformado por corredores exclusivos para ómnibus, terminales e intercambiadores, sistema de control por GPS para los ómnibus y el uso de tarjetas inteligentes para el acceso a los ómnibus. El objetivo del STM es democratizar la movilidad de las personas y reducir los tiempos de viajes, aumentando la calidad del sistema de transporte.

Los desafíos que plantea el estudio de la movilidad urbana han sido atacados con técnicas de optimización y análisis de datos de redes de transporte. Estos estudios requieren de datos de encuestas de hogares, información del censo, y una representación de la red de transporte. Los modelos de demanda del transporte, resultado de la aplicación de estas técnicas, son utilizados en la planificación y operativa del sistema de transporte.

Minería de Procesos (van der Aalst, 2016) es una disciplina del área de Ciencia de Datos y Ciencia de Procesos, utilizada para obtener información valiosa de los procesos en forma de modelos (descubrimiento de procesos), haciendo uso de los eventos registrados en un sistema de información. Su estudio se centra en el

[1] Plan de Movilidad. https://montevideo.gub.uy/sites/default/files/plan_de_movilidad.pdf

proceso, permitiendo analizar la evolución del proceso (o un caso particular del mismo) a lo largo del tiempo. También permite obtener indicadores importantes del proceso, e.g tiempos entre actividades y actividades más frecuentes.

La minería de procesos ha sido utilizada en diferentes contextos como en la industria de la salud (Mans, Schonenberg, Song, van der Aalst, y Bakker, 2008) y en procesos de auditoría (Jans, van der Werf, Lybaert, y Vanhoof, 2011). Sin embargo, existe poco trabajo académico asociado a la aplicación de esta disciplina en movilidad urbana, destacándose (Diamantini, Genga, Marozzo, Potena, y Trunfio, 2017).

En el contexto de este proyecto, se pretende aplicar técnicas de minería de procesos para analizar la información existente sobre la red de transporte y los viajes de los usuarios del STM, con el objetivo de explorar y/o dar visibilidad a la utilidad de la aplicación de dichas técnicas en la búsqueda de un mejor entendimiento de la movilidad urbana, y eventualmente poder brindar soluciones a problemas de interés de la IM en el área. Particularmente, los objetivos específicos son:

1. Explorar la utilidad de las técnicas de minerías de procesos sobre los datos del STM.
2. Realizar un análisis de los principales corredores[2] utilizando datos del STM y minería de procesos.
3. Estudiar el uso de Sistemas de Información Geográfica (GIS, en inglés) en el contexto de minería de procesos y movilidad urbana.

Este proyecto se enmarca en el proyecto de investigación "Minería de procesos para el análisis de movilidad urbana" del programa IM – Udelar "Ing. Oscar J. Maggiolo" de la Comisión Sectorial de Investigación Científica (CSIC)[3], en ejecución por el grupo COAL del INCO.

Este informe y su organización se realizó siguiendo la metodología PM^2 (van Eck, Lu, Leemans, y van der Aalst, 2015), donde se destacan las secciones de extracción, procesamiento de datos, y análisis de datos. Teniendo en cuenta esto, la organización se presenta de la siguiente manera. En el capítulo 2 se presenta un análisis del estado del arte referente a la minería de procesos aplicada a la movilidad urbana, también se introduce un marco teórico con los conceptos mínimos necesarios para el entendimiento del informe. En el capítulo 3 se describen los datos crudos existentes, así como el proceso de extracción y armado de los logs. En el capítulo 4 se presenta un análisis estadístico de los datos contenidos en los logs así como el proceso de filtrado, enriquecimiento y agregación de eventos. En el capítulo 5 se presentan diferentes análisis realizados con los datos disponibles, así como sus resultados más relevantes. En el capítulo 6 se presentan las conclusiones obtenidas y el posible trabajo a futuro a realizar.

[2] Corredores. https://montevideo.gub.uy/areas-tematicas/movilidad/transito/corredores

[3] Programa IM - Udelar. https://www.csic.edu.uy/content/programa-im-udelar-ing-oscar-maggiolo

Capítulo 2

Estado del arte

En esta sección se presenta un relevamiento del trabajo existente al momento de la realización del proyecto, relativo al uso de datos de movilidad urbana. También incluye un marco teórico donde se introducen conceptos relativos a la minería de procesos, necesarios para una mejor comprensión del trabajo realizado.

2.1. Movilidad urbana

Como se menciona en (van der Aalst, 2016), el volumen de datos que se genera en la actualidad es enorme y ha surgido una nueva forma de procesarlos con la llegada del Big Data. Estos datos surgen de distintas fuentes: software tradicional (e.g compra y venta de artículos en una tienda online), redes sociales (comentarios, likes, creación de contenido, entre otros), IoT (dispositivos médicos, tarjetas inteligentes, relojes, entre muchos otros). Esta gran cantidad de datos brinda una posibilidad a los diferentes sectores de la industria de extraer información valiosa para el negocio. Sectores como el de salud, logística, transporte, y muchos otros aprovechan la información de los datos para crear estrategias a largo plazo, tácticas (e.g agregar nuevos recorridos de ómnibus en ciertas fechas), optimizar recursos, etc.

Con el avance de la tecnología, se ha facilitado la instalación de dispositivos que permiten generar información relacionada al transporte público, lo que se puede ver en aplicaciones que indican la posición actual de un determinado ómnibus, embotellamientos y lugares disponibles en un estacionamiento, radares de control de velocidad y tarjetas inteligentes. Este último es el de mayor relevancia para este proyecto. Las tarjetas inteligentes permiten hacer un seguimiento de la movilidad de los distintos usuarios que utilizan los transportes públicos (e incluso podrían permitir accesos a otros servicios, generando así mayor cantidad de datos). Permiten identificar cuándo un usuario ha tomado un transporte público y en qué lugar, cuál fue su recorrido y el tiempo que transcurre, cuantos usuarios había al momento de utilizar el transporte, y otra gran

cantidad de datos.

Con respecto a la privacidad de los datos, es muy importante ocultar (anonimizar) información sensible, permitiendo seguir trabajando con los datos pero sin conocer la identidad (aunque sea de forma parcial[1]) de los usuarios.

2.1.1. Sistema de Transporte Metropolitano (STM)

Con el fin de modernizar y reorganizar el transporte público de la ciudad, en 2010, la IM presentó un Plan de Movilidad (PM) que incluye, entre otras cosas, la creación y adaptación de carriles exclusivos y preferenciales en las calles principales; mantenimiento y ampliación de la red de carreteras; mejoras en terminales de transbordo e intercambiadores, así como el reacondicionamiento urbano de sus áreas de influencia; sincronización y expansión de la red de semáforos, junto con la instalación de cámaras para controlar el tráfico. Como parte fundamental del PM, se estableció el Sistema de Transporte Metropolitano (STM) para unificar el sistema de transporte público, principalmente compuesto por buses y operado por cuatro empresas privadas.

En el contexto del STM, los buses fueron equipados con unidades GPS a bordo y máquinas de venta de boletos que funcionan principalmente con tarjetas inteligentes. Estos dispositivos generan gran cantidad de datos útiles para el Centro de Gestión de Movilidad (CGM)[2], organismo encargado de la planificación y gestión de la movilidad en Montevideo. La combinación de los datos del STM y el monitoreo en tiempo real de los semáforos y cámaras de tráfico permite al CGM tener una visión integral de la movilidad en la ciudad y tomar decisiones basadas en datos para mejorar su eficiencia y sostenibilidad.

Un componente importante del STM son las tarjetas inteligentes, que son tarjetas de recarga sin contacto y están vinculadas a la identidad del usuario. Con estas tarjetas, los pasajeros pueden comprar diferentes tipos de boletos, como boletos regulares, boletos preferenciales para líneas con mejores vehículos y rutas más rápidas, boletos locales para viajes cercanos entre zonas específicas de la ciudad y boletos para ser usados dentro del centro de la ciudad. Además, hay dos tipos de boletos que permiten transbordos: los boletos de una hora, que permiten abordar dos buses en una hora, y los boletos de dos horas, que permiten hacer transbordos ilimitados dentro de dos horas. También es posible pagar en efectivo para aquellos que no tengan tarjeta STM, pero solo se permiten boletos sencillos sin transbordos.

La información recogida por las máquinas de venta de boletos, como el tipo de boleto adquirido, la fecha, hora y parada de ascenso, y la línea de bus, están disponibles para el público. Sin embargo, otros datos más sensibles, como la ubicación por GPS del bus o el identificador de la tarjeta (que permitirían identificar al titular de la misma), no están disponibles públicamente.

[1] Haciendo un análisis exhaustivo y considerando información no presente en los datos (e.g que linea de ómnibus utiliza una persona para ir al trabajo) se podría conocer la identidad de un usuario que en principio está anonimizada.

[2] Centro de Gestión de Movilidad. https://montevideo.gub.uy/centro-de-gestion-de-movilidad

2.2. Procesos de negocio y minería de procesos

Es un aspecto natural y esperado en las organizaciones el buscar mejorar su modo de operar. Como se describe en (van der Aalst, 2016), los procesos de negocio son medios por los cuales una organización intenta alcanzar sus objetivos, y la minería de procesos es una técnica que permite analizar y evaluar estos procesos con el fin de mejorar la eficacia y eficiencia de la organización.

2.2.1. Procesos de negocio

Un proceso de negocio es un conjunto de actividades o tareas interrelacionadas que se llevan a cabo para lograr un objetivo específico en el contexto de una organización (van der Aalst, 2016). Los procesos de negocio pueden ser muy diversos, y pueden incluir desde procesos de producción y distribución de productos o servicios hasta procesos de atención al cliente, gestión de proyectos y toma de decisiones.

Los procesos de negocio son esenciales para la operación y el funcionamiento de una organización, ya que permiten a las empresas llevar a cabo sus actividades de manera eficiente y efectiva. La mejora de los procesos de negocio puede tener un impacto positivo en la productividad, la calidad, los costos y la satisfacción del cliente.

En general, un proceso de negocio incluye la identificación de un objetivo o necesidad, la definición de los pasos y actividades necesarias para lograr ese objetivo, y la ejecución y control de esos pasos y actividades para garantizar que se cumplan de manera eficiente y efectiva. La minería de procesos es una técnica que se utiliza para analizar y optimizar los procesos de negocio.

Como se menciona en (Delgado y Calegari, 2022), usualmente estos procesos de negocio presentan algunas particularidades que aumentan la dificultad de su abordaje:

- Son muy grandes y complejos, involucrando varias secciones o áreas de una organización o incluso varias organizaciones distintas.
- Su duración además de ser variable, puede llegar a ser muy larga (semanas, meses, años).
- Están muy adaptados a un dominio en específico (salud, viajes, contabilidad, etc), dificultando su generalización.
- Por último y no menos importante, estos procesos pueden ser manuales, automatizados o una mezcla de ambos, y en ocasiones estar implícitos en los propios sistemas de automatización, haciendo difícil su modelado de forma explícita.

Modelado de procesos

Existen muchas notaciones para el modelado de procesos en la actualidad (BPMN, Petri nets, YAWL, EPC entre otros), lo cual demuestra la relevancia del modelado de procesos. Algunas organizaciones pueden usar sólo modelos de procesos informales para estructurar discusiones y documentar procedimientos.

Sin embargo, las organizaciones con cierto nivel de madurez en modelado de procesos, utilizan modelos que pueden analizarse y utilizarse para establecer procesos operativos. Hoy en día, la mayoría de los modelos de procesos se realizan de forma manual y no se basan en un análisis riguroso de los datos de procesos existentes.

Yendo a lo concreto, los modelos de proceso representan un conjunto de actividades y las restricciones de ejecución entre ellas. Entre otras cosas muestran:

- Actores involucrados en el proceso (roles, áreas).
- Actividades operativas distinguibles y su secuencia.
- Entradas, salidas, recursos, eventos.

Lenguajes de modelado

Como se mencionó, existen varios lenguajes de modelado. A continuación se explican tres de ellos, Petri net, BPMN y Directly-Follows Graph.

Petri net

Las Petri nets son el lenguaje de modelado de procesos más antiguo y mejor investigado que permite el modelado de actividades concurrentes. Aunque la notación gráfica es intuitiva y simple, las Petri nets son ejecutables y se pueden aplicar muchas técnicas de análisis sobre ellas.

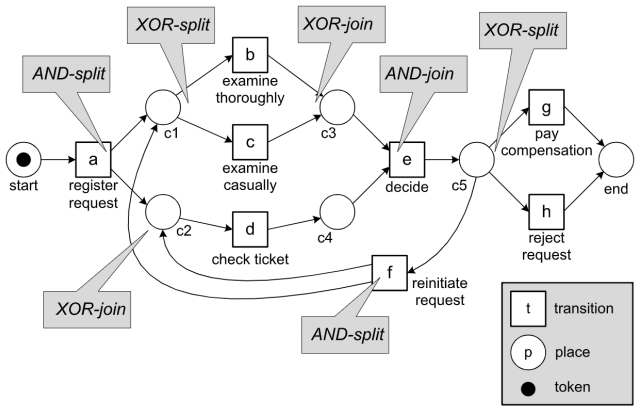


Figura 2.1: Petri net relacionada al proceso de manejo de una solicitud de compensación dentro de una aerolínea. Extraída de (van der Aalst, 2016).

La Figura 2.1 muestra el proceso de manejo de una solicitud de compensación dentro de una aerolínea, con algunos conceptos importantes resaltados. Una Petri net es un gráfico bipartito que consta de *places* y *transitions*. El estado de una Petri net está determinado por la distribución de *tokens* en los *places*. El avance de los tokens sobre la red (ejecución del proceso) se realiza cuando se cumple la regla de activación, denominada *firing rule*:

- Previo a la ejecución de una transición, los *places* inmediatamente anteriores deben tener un *token*.
- Luego de ejecutada la transición, se consumen los *tokens* de los *places* anteriores, y se generan nuevos *tokens* en los *places* siguientes.

Para controlar las bifurcaciones (XOR) y la concurrencia (AND) de un proceso en una Petri net, se definen los siguientes conceptos:

- **AND-split:** cada transición con más de un arco saliente, puede ejecutarse en paralelo.
- **AND-join:** cada transición con más de un arco entrante, debe esperar la finalización de la ejecución de cada transición (generadas por el AND-split) para proseguir con la ejecución del proceso.
- **XOR-split:** de cada transición con más de un arco saliente, puede ejecutarse solamente una.
- **XOR-join:** cada transición con más de un arco entrante, debe esperar la finalización de la ejecución de una única transición (generada por el XOR-split) para proseguir con la ejecución del proceso.

Por ejemplo, en la Figura 2.1 las transiciones *Check Ticket* y *Examine Casually* pueden ejecutarse en paralelo. Sin embargo, esto no ocurre con las transiciones *Examine Casually* y *Examine Thoroughly*, que no pueden ser ejecutadas concurrentemente.

BPMN 2.0

BPMN (Business Process Modeling Notation, en inglés), es uno de los lenguajes más utilizados para modelar procesos de negocio. La Figura 2.2 muestra el mismo proceso de la Figura 2.1, pero utilizando notación BPMN. En este caso, las actividades del proceso se representan con rectángulos y las bifurcaciones y concurrencia del proceso se representan con rombos.

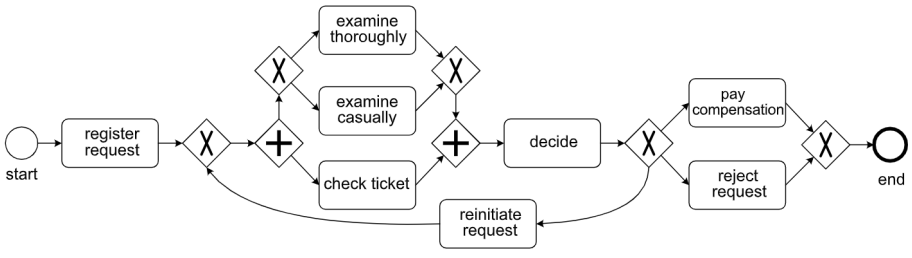


Figura 2.2: Modelo de Figura 2.1, con notación BPMN. Extraída de (van der Aalst, 2016).

Directly-Follows Graph

El Directly-Follows Graph (DFG) es un lenguaje de modelado utilizado en la práctica en la mayoría de las herramientas de minería de procesos, incluyendo Disco[3]. En este grafo existen nodos y arcos, que representan las actividades y las relaciones entre ellas, respectivamente. Un arco entre dos nodos cualesquiera representa que la actividad que recibe el arco está seguida de la otra. En la Figura 2.3 se muestra el mismo proceso de la Figura 2.1 utilizando DFG.

El DFG no tiene elementos para representar control de flujo (AND/XOR) en los modelos, como sí lo tienen las Petri Nets y el BPMN 2.0. DFG presenta problemas con actividades que son concurrentes (van der Aalst, 2019), en estos casos se forman loops entre estas aunque se ejecuten como máximo una sola vez en el proceso, tendiendo a modelos menos estructurados.

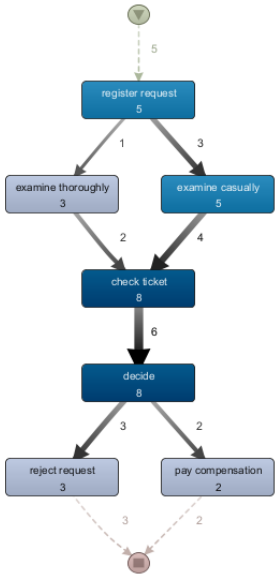


Figura 2.3: Modelo de Figura 2.1 utilizando DFG.

2.2.2. Minería de procesos de negocio

Acorde a (van der Aalst, 2016), la minería de procesos es una técnica de análisis de datos que se utiliza para descubrir patrones y tendencias en grandes conjuntos de información, con el fin de identificar oportunidades de mejora y optimización en los procesos de negocio. Esta técnica se aplica a menudo en el contexto de la gestión de procesos de negocio, donde se busca mejorar la eficiencia y la calidad de los procesos mediante el análisis de datos que se generan a lo largo de su ejecución.

La minería de procesos se basa en el uso de técnicas de análisis de datos para analizar grandes conjuntos de información, típicamente registros (logs) de eventos, y extraer conocimientos y patrones útiles. Estos conocimientos y patrones pueden utilizarse para optimizar los procesos de negocio, identificar oportunidades de mejora, predecir futuros comportamientos y tomar decisiones informadas.

Se considera un evento como una actividad relacionada con un caso en particular (instancia de un proceso). Normalmente, en los logs de eventos se puede

[3]Disco. https://fluxicon.com/disco/

almacenar información correspondiente con distintas actividades que conforman el proceso como tal. Por ejemplo, entre esta información se puede tener: un número de identificador del caso, el nombre de la actividad, el ejecutor, la fecha y hora de inicio y finalización de la actividad.

En general, la minería de procesos se utiliza para mejorar la eficiencia y la calidad de los procesos de negocio, reducir costos y aumentar la productividad. También puede utilizarse para identificar oportunidades de innovación y para desarrollar nuevos productos o servicios.

Habiendo definido la minería de procesos y los procesos de negocios, se puede tener entonces un panorama más claro de lo que es la minería de procesos de negocio. Dentro de la misma se destacan tres actividades principales:

- **Descubrimiento:** a partir de sólamente un registro de eventos como entrada, genera como salida un modelo de proceso de negocio, como por ejemplo la Petri net de la Figura 2.1. Esto es posible mediante la utilización de diferentes algoritmos de minería de procesos (alpha, inductivos, fuzzy, heurísticos, entre otros).
- **Conformidad:** Contrastar la realidad observada a partir del registro de eventos, con el modelo de proceso en el que está basado y viceversa. Esto permite por ejemplo encontrar desviaciones entre el modelo y la ejecución real del proceso.
- **Extensión:** A partir de un modelo de proceso de negocio ya existente, y junto con la nueva información obtenida sobre su ejecución real (participación de roles, tiempos, desviaciones) realizar mejoras o extensiones en el mismo.

A su vez, la minería de procesos de negocio puede ser llevada a cabo desde diferentes perspectivas:

- **Organizacional:** enfocada en los propios recursos (sistemas, personas, etc) que participan en el proceso.
- **Temporal:** enfocada en el tiempo y frecuencia con la que los eventos ocurren durante el proceso.
- **Flujo de control:** se enfoca en el orden de la ejecución de las actividades definidas en el flujo de control del proceso.
- **Flujo de datos:** se enfoca en los datos y el valor de los mismos durante el avance del flujo por cada instancia, por ejemplo, los atributos de los eventos.

2.2.3. Procesamiento de datos

Como se mencionó previamente, a partir del log de eventos es posible realizar las tres actividades principales de la minería de procesos: descubrimiento, conformidad y extensión. En esta sección, se presenta el log de eventos y se explica en qué consiste cada una de estas actividades.

Log de eventos

La Tabla 2.1 muestra un fragmento de log del proceso de la Figura 2.1, se muestra la información que típicamente se encuentra en un log de eventos utilizado para la minería de procesos.

Tabla 2.1: Fragmento de log de eventos. Extraído de (van der Aalst, 2016).

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...

Un log de eventos contiene datos relacionados con un único proceso, y cada entrada en el log debe hacer referencia a una única instancia de proceso, a la que se le denomina caso, esto puede verse en la Tabla 2.1, donde cada entrada está relacionada con un caso, por ejemplo, el caso 1. A su vez, cada registro puede tener alguna actividad asociada, como las actividades *register request*, *check ticket* y *reject request* del ejemplo. Para identificar un proceso en minería de procesos, es necesario poder asociar una actividad a una instancia del proceso, en el ejemplo quienes permiten realizar eso son las columnas "Case id" y "Activity". Otro requerimiento de los registros es que deben poder ordenarse. Por ejemplo, el evento 35654425 ocurre antes que el evento 35654426, (la actividad *check ticket* se ejecuta antes que *decide*). De no poder ordenarse los datos, sería imposible identificar las distintas dependencias en los modelos de procesos. De esta manera, se puede definir la *traza* de un proceso como la secuencia ordenada de actividades realizadas por una instancia de dicho proceso. Por ejemplo, la traza del caso 1 es *register request* - *examine thoroughly* - *check ticket* - *decide* - *reject request*.

Otros dos campos (también llamados *atributos*) útiles, son los campos "Timestamp" y "Resource". El primero se utiliza para ordenar los eventos en el log, también es de utilidad para analizar propiedades relacionadas al rendimiento, como por ejemplo el tiempo de espera entre dos actividades. El segundo es útil para identificar, por ejemplo, las personas encargadas de llevar a cabo una tarea.

La Figura 2.4 muestra la estructura de árbol del log de eventos presentado en la Tabla 2.1. En ella pueden reforzarse visualmente las definiciones recién mencionadas:

- Un *proceso* se compone por un conjunto de *casos*.

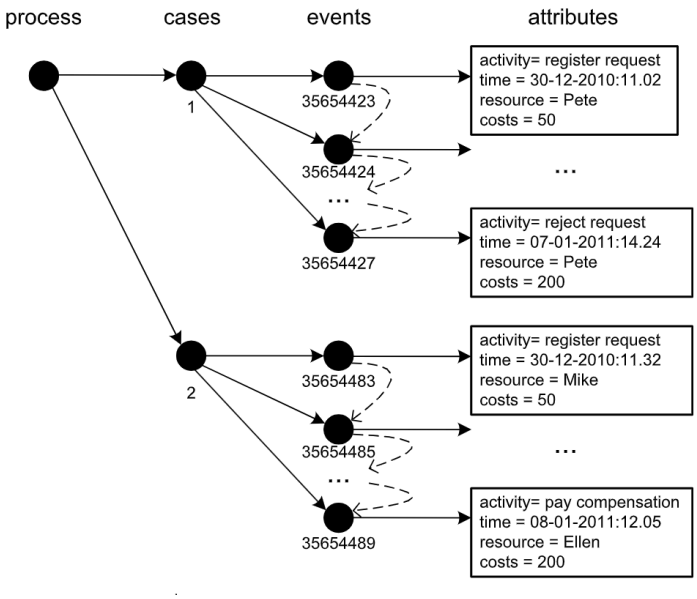


Figura 2.4: Estructura de árbol basada en log de eventos de 2.1. Extraído de (van der Aalst, 2016).

- Un *caso* está compuesto por *eventos*, en donde cada evento se relaciona a un sólo *caso*.
- Los eventos relacionados a un *caso* están ordenados.
- Los eventos pueden tener *atributos*. Ejemplos de nombres de atributos típicos son actividad, tiempo, costos y recursos.

No es necesario que todos los eventos tengan los mismos atributos, no obstante los eventos asociados a la misma actividad comparten sus atributos.

Preprocesamiento

El armado del log se realiza obteniendo datos que en ocasiones se encuentran en múltiples fuentes, y potencialmente con diferentes semánticas, lo que complejiza la tarea. Por este motivo, es importante tener en cuenta algunas consideraciones para el correcto armado del mismo, para que pueda ser de utilidad para la correcta aplicación de las técnicas de minería de procesos. Algunas de estas consideraciones son:

- El log puede tener casos no contemplados, como instancias de procesos en ejecución que no han finalizado.
- Comportamiento no frecuente, trazas que son ejecutadas en una ínfima proporción de veces.
- Datos incompletos, algunas instancias podrían no tener todos los datos necesarios.

Es necesario inspeccionar el log para identificar estos problemas, eliminando instancias incompletas y comportamientos no frecuentes.

Extracción

Luego de construido el log, es necesario asegurar ciertos criterios de calidad y consistencia sobre el mismo para obtener mejoras significativas durante la aplicación de las técnicas de minería de procesos. Algunos de estos criterios son:

- Definir el alcance de la información obtenida, y en base a eso seleccionar los eventos relevantes.
- Cada registro de eventos debe estar relacionado a un único proceso, a su vez, cada evento de este registro debe referenciar a una única instancia del proceso. Estas referencias deben ser unívocas a través de un identificador (único o compuesto).
- Debe poder establecerse un orden entre los eventos, preferentemente mediante una marca de tiempo.
- Los valores de los atributos deben ser precisos, de lo contrario se debe explicitar esta situación.
- Mantener la misma estructura de los registros, con la finalidad de ser comparables a lo largo del tiempo y para diferentes variantes de procesos.

También es recomendable tener nombres de referencia y atributos con semánticas claras, es decir, que tengan el mismo significado para todas las personas involucradas en la creación y análisis de los datos de los eventos; así como tener una colección estructurada de nombres de referencias y atributos. También es recomendable de ser posible, guardar información transaccional sobre el evento (fecha inicio, fecha de fin, etc), realizar chequeos automatizados de consistencia y correctitud para asegurar la correctitud sintáctica del log de eventos, y asegurar la privacidad sin perder las correlaciones significativas.

Estándar XES

XES es un formato estándar utilizado para trabajar sobre los logs de eventos, específicamente en su guardado e intercambio. La Figura 2.5 es un metamodelo de XES presentado en formato UML.

Un registro (log) contiene trazas y cada traza contiene eventos. Los registros, las trazas y los eventos tienen atributos, los cuales pueden ser de varios tipos, siendo los principales *String*, *Date*, *Int*, *Float* y *Boolean*. Las extensiones pueden ser utilizadas para definir nuevos atributos, los cuales deben estar especificados en el registro. Los atributos globales son atributos que se declaran obligatorios. Estos atributos pueden estar a nivel del evento o de traza, así como también pueden estar anidados. Los clasificadores de eventos se definen para el registro y asignan una "etiqueta" (por ejemplo, el nombre de la actividad) a cada evento.

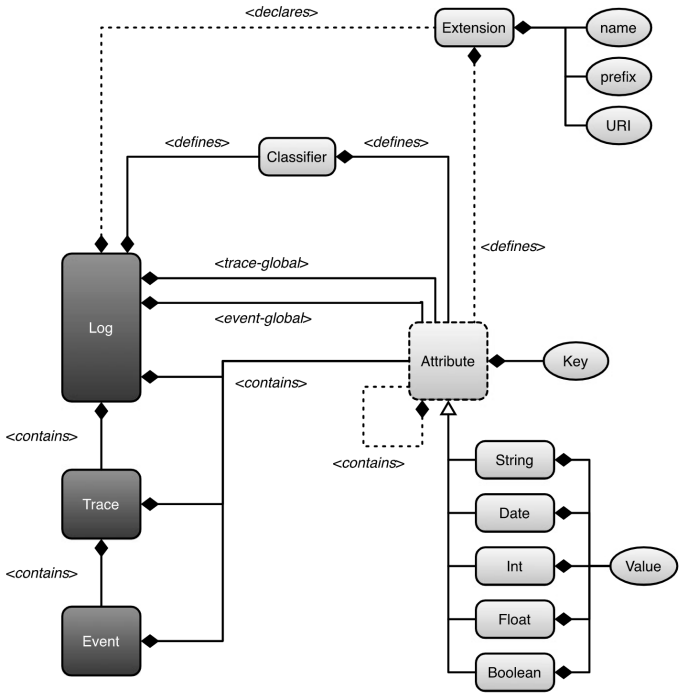


Figura 2.5: Metamodelo de XES. Extraído de (van der Aalst, 2016).

Descubrimiento

El modelado de procesos consiste en construir un modelo de proceso, a partir del comportamiento observado en los logs de eventos. Se pueden generar modelos como por ejemplo, un BPMN o una Petri net. Estos modelos son generados a partir de diversos algoritmos, de los cuales algunos de ellos son el inductivo, alpha y Fuzzy Miner.

Algoritmo Alpha

Este algoritmo toma como entrada el log de eventos y produce como salida una Petri net. Puede generar el modelo a partir de una recorrida del log de eventos, buscando patrones de comportamiento entre las relaciones de eventos contiguos. Los patrones de comportamiento que pueden encontrarse son los de secuencia XOR-split, XOR-join, AND-split y AND-join.

Este algoritmo es muy conocido debido a ser de los primeros en lograr, de forma exitosa, visualizar los conceptos y problemas relacionados al descubrimiento de modelos. A pesar de esto, debido a su antigüedad ya no es tan utilizado y en la práctica es conveniente utilizar mejoras del algoritmo como *alpha+*, *alpha++*, *alpha$*.

Algoritmo inductivo

Tomando como base el log de eventos, esta técnica de descubrimiento genera un árbol de proceso, un modelo con una estructura que puede ser transformado a una *Petri net* del tipo *workflow net*, que tiene la particularidad de poseer un estado inicial y un estado final, en donde para cada estado puede encontrarse un camino desde el estado inicial hasta el estado final. Esta técnica posee varios algoritmos, desde los muy básicos a los más avanzados, que pueden incluir diversos filtrados.

El algoritmo básico de Inductive Mining se basa en separar el log en diferentes subgrupos con eventos disjuntos, relacionándolos entre sí mediante operadores que describen el comportamiento entre ellos. Este proceso se realiza de forma iterativa y finaliza cuando cada grupo contiene un sólo evento. Estos operadores son:

- Secuencia ($a \rightarrow b$): luego del evento a ocurre el evento b.
- Mutua exclusión ($a \times b$): sólo puede ocurrir el evento a o el evento b.
- Paralelo ($a \parallel b$): pueden ocurrir tanto a como b en cualquier orden.
- Loop: ($a \circlearrowleft b$): la secuencia de eventos ab puede repetirse indefinidamente.

Gracias a su flexibilidad, eficacia y escalabilidad, esta técnica actualmente es una de las más utilizadas en la etapa de descubrimiento.

Fuzzy miner

Otra técnica utilizada para realizar la etapa de descubrimiento es el Fuzzy Miner. Este algoritmo propuesto en (Günther y van der Aalst, 2007) es muy utilizado en la práctica, ya que permite ejecutar abstracciones y generalizaciones sobre el proceso. El resultado del algoritmo es un Fuzzy Model y puede ser simplificado y generalizado con parámetros configurados por el usuario. Disco, el software de minería de procesos utilizado en el proyecto, utiliza este algoritmo para realizar el descubrimiento de modelos.

Propiedades de los modelos

Independientemente del algoritmo utilizado para descubrir el modelo del proceso, existen diferentes cualidades de los modelos que permiten compararlos entre ellos y, por tanto, comparar los algoritmos que los generan. Las cualidades que existen para los modelos son las siguientes:

- **Fitness:** un modelo con alto nivel de fitness permite ejecutar el comportamiento que se observa en el log de eventos.
- **Precisión:** un modelo con alto nivel de precisión no permite ejecutar comportamiento muy distinto a lo observado en el log.
- **Generalización:** un modelo que generaliza permite ejecutar comportamiento no visto, pero que se asemeja a lo que se encuentra en el log.

- **Simplicidad:** el modelo debe ser lo más simple posible.

En general, estas cualidades de los modelos están inversamente relacionadas. Por ejemplo, cuando aumenta el fitness de un modelo, es de esperar que la precisión sea menor. Como el modelo permitirá mayor comportamiento, también permitirá mayor comportamiento completamente distinto (menor precisión).

Otra propiedad importante de los modelos es lo que se conoce como soundness. Un modelo sound es un modelo que tiene las siguientes propiedades:

- **Safeness:** *places* (petri net) no pueden tener multiples tokens al mismo tiempo.
- **Proper completion:** al momento de finalizar no quedan *tokens* que no estén en el final del proceso para esa instancia.
- **Option to complete:** cualquier instancia del proceso debe poder ser completada.
- **Absence of dead parts:** cualquier actividad debe poder ser ejecutada en alguna instancia del proceso.

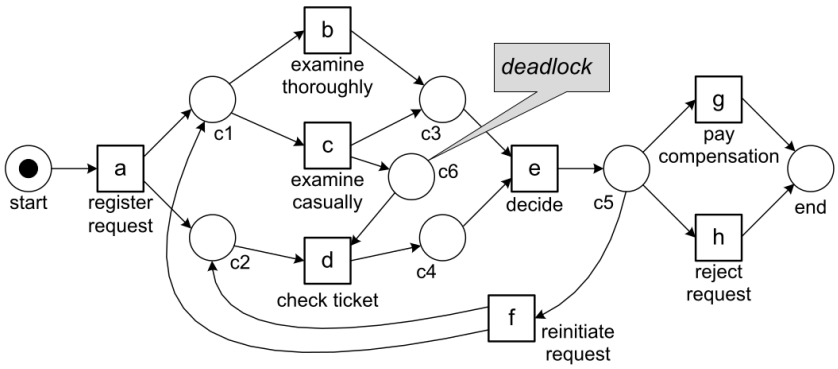


Figura 2.6: Visualización de deadlock en ejecución de modelo basado en Figura 2.1. Extraído de (van der Aalst, 2016).

El modelo de la Figura 2.6 no es *sound* al no cumplir con la propiedad *Option to complete*, ya que al ejecutar $< a, b >$, se tiene el marcado *[c2,c3]* (un token en *c2* y un token en *c3*), por lo que *d* no podrá ser ejecutada debido a que precisa un token en *c6* y esto no es posible. Por lo tanto *e* no será ejecutada y la instancia del proceso estará en deadlock.

Idealmente, se busca un modelo que tenga un buen grado de fitness y precisión, que pueda generalizar el comportamiento observado en el log, que sea simple y *sound*. El *soundness* no es fácil de conseguir, por lo que en general se relaja la búsqueda, sin tomar el *soundness* como cualidad necesaria.

Benchmark de los algoritmos

El modelo que se descubre depende del algoritmo utilizado. Por lo tanto, las cualidades de los modelos dependen de los algoritmos que se utilicen. A continuación se detallan las fortalezas y debilidades de algunos algoritmos para el descubrimiento de modelos, extraídos del análisis realizado en (Augusto y cols., 2019). Este análisis toma como entrada algoritmos de descubrimiento (2013-2017), logs (públicos y propietarios) y como resultado se analizan y se comparan las dimensiones de calidad de cada modelo generado: fitness, precisión, f-score, simplicidad, soundness y tiempo de ejecución del algoritmo utilizado.

En las pruebas realizadas (sin optimización de parámetros) destacaron los algoritmos Inductive Miner (IM), Evolutionary Tree Miner (ETM), y Split Miner (SM), los cuales presentaron los mejores valores (por encima de los otros algoritmos en la mayoría de las ejecuciones) en las métricas mencionadas anteriormente.

Otros algoritmos como Fodina (FO) y Structured HM (S-HM) tuvieron problemas para obtener modelos sound. Sin embargo presentan buenos valores de fitness. De los modelos *sound* que se obtuvieron en S-HM, 9 de 16 obtuvieron los mejores resultados para fitness y generalización, lo que lo coloca como la mejor opción junto a IM si se consideran estas medidas.

El algoritmo alpha$ mostró problemas de escalabilidad al hacer time-out en 8 logs de 24. Sin embargo, presentó buen fitness y precisión en general.

Resultado interesante del análisis de la ejecución con optimización de parámetros es que FO y S-HM descubren modelos sound y tienen buenos resultados a costa de gran consumo de recursos y cerca de 24 horas de ejecución para algunos logs reales de gran tamaño, lo cual IM, ETM y SM no pudieron manejar.

Chequeos de conformidad

Otra tarea que se realiza en la minería de procesos es el chequeo de conformidad. A diferencia del descubrimiento de modelos, en esta etapa, se busca determinar si un modelo es conforme al log de eventos y/o viceversa. El modelo puede ser el generado en la etapa de descubrimiento, o puede ser un modelo ya diseñado por expertos en el dominio. En cualquier caso, lo que se busca es encontrar desviaciones, obtener información del proceso y analizar perspectivas organizacionales, de performance, entre otras. Todas las respuestas o conocimiento que se produzca, generan una acción a realizar. La acción puede ser corregir el modelo viendo que este no refleja la realidad. También se puede tomar la acción de corregir el proceso, haciendo ajustes para que el proceso sea más eficiente en términos de tiempo o recursos.

La ejecución de los procesos de las organizaciones se respaldan en personas y sistemas informáticos. Las personas son las encargadas de realizar los pasos manuales del proceso, generar cambios de estado en el mismo y confirmar estos cambios a través de los sistemas informáticos. A su vez, estos sistemas, además de persistir los cambios del proceso en bases de datos, realizan actividades automáticas interactuando con otros sistemas. Este ecosistema que respalda al pro-

ceso no es suficiente para garantizar que los logs y los modelos correspondientes sean conformes. En primer lugar, los responsables de realizar las actividades del proceso son humanos, esto hace que involuntariamente o no, se cometan ciertos errores al realizar las tareas del proceso, por ejemplo, el no chequeo de ciertos cumplimientos al solicitante de un préstamo bancario. Incluso si los encargados de realizar las tareas no cometen errores, muchas veces el sistema informático que respalda el proceso no es lo suficientemente flexible para adaptarse al mismo. Por ejemplo, actualmente existen sistemas utilizados por las organizaciones que contienen un conjunto de soluciones para distintos tipos de procesos. Si bien estos sistemas suelen adaptarse muy bien a distintos procesos, cada organización tiene sus peculiaridades, haciendo que estos sistemas genéricos no se adapten en su totalidad al proceso de la organización, generando así una brecha entre el sistema informático y el proceso real. Esta brecha, indudablemente, se verá reflejado en la no conformidad entre el log de eventos y el modelo, debido a las acciones ad-hoc que se deban tomar para poder continuar con la ejecución del proceso cuando el sistema no lo permita.

El chequeo de conformidad permite visualizar este tipo de disconformidades entre el log y el modelo. Esto hace más sencillo la auditoría de procesos o que los stakeholders puedan determinar si los procesos se están realizando dentro de los límites acordados. Este último aspecto no fue de los principales en el contexto de este proyecto y por lo tanto no se incluyó en su ejecución.

Extensión de modelos

Como se mencionó previamente, la minería de procesos de negocio puede ser llevada a cabo desde diferentes perspectivas: organizacional, temporal, flujo de control y flujo de datos. Si bien el enfoque principal del descubrimiento de procesos está en la perspectiva del flujo de control, los log de eventos contienen mucha información relacionada con estas otras tres perspectivas.

Perspectiva organizacional

La perspectiva organizacional se puede utilizar para obtener información sobre patrones de trabajo más comunes, estructuras dentro de la organización y también redes sociales, puede ser útil para comprender cómo un modelo particular puede aplicarse y adaptarse a diferentes contextos organizacionales. Se basa en la información sobre los recursos (*resources*) contenidos en los logs, es decir, qué actores (como personas, roles y departamentos) están involucrados y cómo se relacionan. El objetivo es estructurar la organización, clasificando a las personas en términos de funciones y unidades organizativas.

Por ejemplo, si se está considerando la implementación de un modelo de gestión de proyectos en una empresa, es importante tener en cuenta cómo está estructurada la misma y cuáles son sus procesos y procedimientos existentes. La perspectiva organizacional puede ayudar a comprender cómo el modelo de gestión de proyectos puede integrarse con la estructura y los procesos de la

empresa, y cómo puede adaptarse para satisfacer las necesidades específicas de la organización.

Perspectiva temporal

La mayoría de los log de eventos tienen asociada una marca de tiempo (*timestamp*), con una precisión que puede variar, desde solo fechas hasta milisegundos. Dentro de la perspectiva temporal, estas marcas de tiempo, así como las frecuencias de las actividades, juegan un rol fundamental, permitiendo identificar cuellos de botella, analizar tiempos de servicio, hacer un seguimiento de la utilización de recursos y predecir tiempos de procesamiento restantes para instancias en proceso.

Perspectiva de flujo de datos

La perspectiva de flujo de datos se centra en las propiedades de los casos. Cada caso se caracteriza por sus propios atributos, los atributos de sus eventos, la traza generada y la información de su rendimiento. Lo importante de este enfoque es cómo los datos fluyen a través de una organización y cómo estos son utilizados para tomar decisiones y mejorar el rendimiento de la empresa.

Aplicación de las perspectivas

Un ejemplo del uso de la extensión de modelos basado en estas perspectivas puede verse tomando como base el log de eventos de la Tabla 2.1. Como se mencionó, en el mismo existen los atributos *resource*, *timestamp* y *cost*. Estos atributos son claves para extender el modelo con las perspectivas organizacional, temporal y de flujo de datos, respectivamente.

La Figura 2.7 muestra la forma en que un modelo orientado al flujo de control puede extenderse con estas tres perspectivas principales. Por ejemplo, basado en el atributo *resource* del log de eventos de la Tabla 2.1 se puede ver que Sara es la única que realiza las actividades *decide* y *reinitiate request*, lo cual apunta a que hay un rol de *Manager* y que Sara es la única que tiene este rol. La actividad *examine thoroughly* solo es realizada por Sue y Sean, lo cual sugiere la existencia de un rol de *Expert* asociado a esta actividad. Las actividades restantes son ejecutadas por Pete, Mike y Ellen, lo cual puede indicar un rol de *Assistant*. Al explotar la información de recursos en el registro, la perspectiva organizacional se puede agregar al modelo de proceso. De manera similar, la información sobre marcas de tiempo se puede usar para agregar información relacionada con el rendimiento al modelo. La Figura 2.7 muestra que es posible medir el tiempo que transcurre entre un examen (actividades b o c) y la decisión real (actividad e). Si este tiempo es muy largo, se puede utilizar la minería de procesos para identificar el problema y descubrir las posibles causas. Si el registro de eventos contiene información relacionada con el caso, se puede usar para analizar más a fondo los puntos de decisión en el proceso. Por ejemplo, a través del análisis

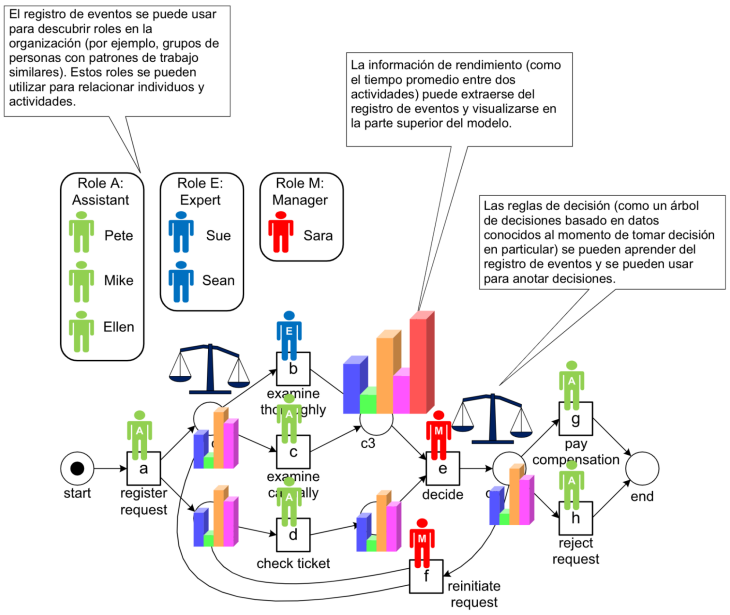


Figura 2.7: Modelo de proceso de la Figura 2.1 ampliado con las perspectivas organizacional (roles y personas), flujo de datos (reglas de decisión) y temporal (rendimiento). Extraído de (van der Aalst, 2016).

de puntos de decisión se puede conocer que las solicitudes de compensación de más de 800 ($cost > 800$) tienden a ser rechazadas.

Una vez obtenidos estos modelos, si se realizan análisis sobre los mismos, se podrían llegar a obtener conclusiones interesantes, como por ejemplo: "las solicitudes examinadas por Sean tienden a ser rechazadas con más frecuencia", "las solicitudes para las que se verifica el ticket después del examen tienden a demorar mucho más", "las solicitudes de menos de 500 tienden a completarse sin ninguna iteración adicional".

2.2.4. Herramientas

Para poder llevar a cabo la disciplina de minería de procesos es necesario contar con herramientas que se adecuen a la misma. Se necesitan herramientas para la extracción y preprocesamiento de los datos, para la generación de los modelos y el posterior análisis de estos. Para la extracción y preprocesamiento, las herramientas clásicas de minería de datos son de gran utilidad para esta fase del proceso. Herramientas como hojas de cálculo y sistemas de bases de datos son utilizadas para la visualización de los datos y el preprocesamiento de los mismos. También son utilizados herramientas ad-hoc, como puede ser un programa creado específicamente para realizar manipulaciones de datos de interés, que no vienen incorporados (o es muy complejo de realizarlos) en los sistemas clásicos.

Si bien el proyecto tiene una gran parte de manipulación de datos y preprocesamiento del log, el objetivo principal es generar modelos y analizarlos. Para esto existen muchas herramientas que pueden ser analizadas desde diferentes perspectivas. Una visión que se le puede dar a estas herramientas es desde la licencia que tienen, si son open source o comerciales. Esto determina en gran parte el foco que tiene cada herramienta, la flexibilidad y soporte de cada una.

También se las puede ver con la perspectiva de a qué grupo de usuarios está dirigida la herramienta. Por ejemplo, puede estar dirigida a usuarios que necesitan de un rápido análisis, para esto incluyen pre configuraciones para que el usuario no consuma tiempo en realizarlas, y otras herramientas pueden estar dirigidas a usuarios que necesitan realizar un mayor análisis, estos usuarios no tienen como foco principal el análisis rápido, sino más bien un análisis más en detalle de los procesos y por esto mismo, las herramientas permiten crear configuraciones por parte de los usuarios.

Otro punto a tener en cuenta, que está intrínsecamente relacionado con la cualidad de si una herramienta es comercial o no, es que las herramientas comerciales, como su nombre lo indica, tienen calidad comercial y, por lo tanto, están testeadas rigurosamente, lo que indica que la mayoría de las funcionalidades que se ofrecen en la herramienta, funcionarán correctamente.

Al momento de este informe, existen varias herramientas, cada una con un enfoque particular medido desde las perspectivas mencionadas anteriormente. Concretamente, existen dos herramientas que se analizaron en este proyecto, y que son utilizadas tanto en lo académico como en lo industrial. Estas herramientas son ProM[4] y Disco.

ProM es una herramienta open source, de origen académico para la minería de procesos. Esta herramienta es muy flexible debido a su arquitectura de plug-in. La herramienta cuenta con cientos de plugins, cada uno enfocado en algún análisis o modelo en particular. Además, cada uno puede crear su propio plugin, lo que lo hace más flexible aún. La contraparte de esta herramienta es que no todos los plug-ins tienen calidad comercial. Permite la importación de archivos XES, MXML y CSV. La Figura 2.8 muestra vistas del espacio de trabajo y acciones disponibles en ProM.

Por otro lado, se encuentra Disco, una herramienta de ámbito comercial, muy enfocada al análisis y generación rápida de modelos. No tiene la flexibilidad de ProM, pero es muy sencilla de utilizar, genera modelos rápidamente (utiliza una versión modificada del algoritmo Fuzzy Mining[5]). El pasaje de un log de eventos a un modelo, se hace a través de un par de clics, y permite analizar distintas perspectivas del modelo, como pueden ser métricas con respecto al log, performance, analizar visualmente como el modelo ejecuta el log, entre otros análisis. También permite la importación de archivos XES, MXML y CSV. La interfaz de análisis de Disco se separa principalmente en tres secciones. La vista de mapa ("Map") que muestra el modelo de proceso descubierto y da una idea de los flujos y tiempos del proceso, la vista de estadísticas ("Statistics") que

[4]ProM Tools. https://promtools.org/
[5]Disco Tour. https://fluxicon.com/disco/files/Disco-Tour.pdf

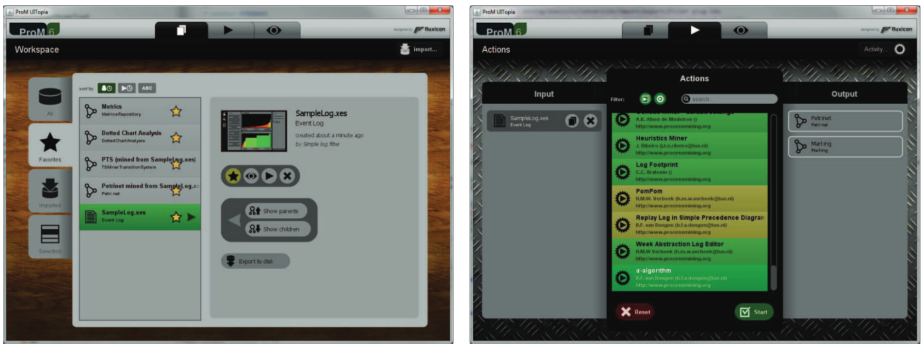


Figura 2.8: Interfaz de ProM. A la izquierda la vista de espacio de trabajo, a la derecha la vista de acciones.

muestra métricas de rendimiento detalladas sobre el proceso y la vista de casos ("Cases") que muestra datos crudos sobre los casos y las variantes.

En la Figura 2.9 se puede ver un ejemplo de la vista de mapa. Esta vista permite al usuario variar el nivel de detalle (cantidad de actividades y transiciones mostradas) de forma interactiva, observar los controles deslizantes ("Activities", "Paths") en la barra lateral derecha de la pantalla de la Figura 2.9. También permite cambiar las métricas desplegadas, teniendo dos tipos de visualizaciones, una de frecuencias ("Frequency") y una de rendimiento ("Performance"). El modelo de proceso desplegado para ambas visualizaciones es el mismo, lo que cambia son valores que se muestran.

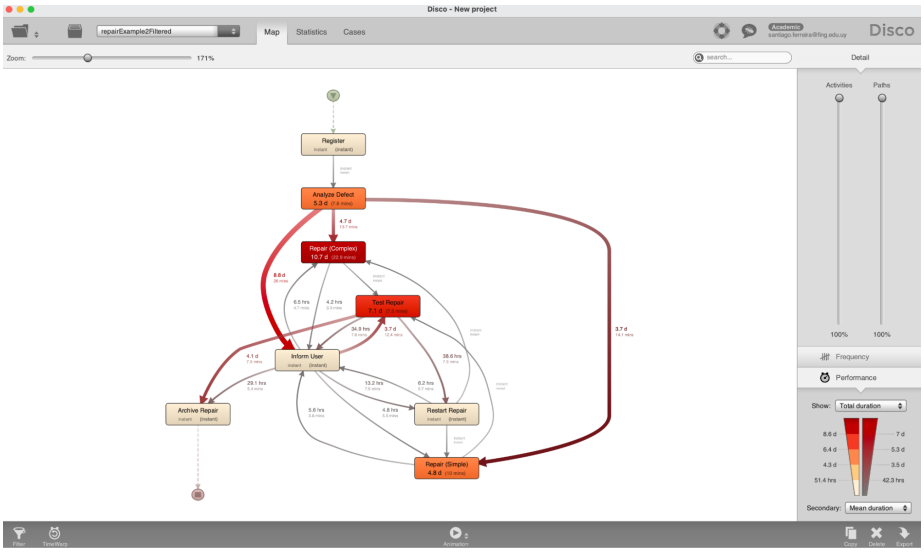


Figura 2.9: Interfaz de Disco. Vista de mapa, visualización de rendimiento.

En la visualización de frecuencias se brinda información sobre la cantidad de

ocurrencias de las actividades y transiciones. Se puede ver la cantidad de apariciones en total en el log ("Absolute frequency"), sin contar las repeticiones en un mismo caso ("Case frequency"), la máxima cantidad de repeticiones para algún caso ("Max. repetitions") y el porcentaje de casos de los que es partícipe ("Case coverage"). En la visualización de rendimiento se brinda información sobre la duración total, media, mediana, máxima y mínima de las actividades y transiciones. Para ambas visualizaciones se permite seleccionar una métrica primaria y otra secundaria para mostrar. En la Figura 2.9 se seleccionó la duración total como métrica principal y la media como secundaria. La métrica principal es la que se toma en cuenta para las intensidades de colores de las actividades así como para los grosores de flechas que representan las transiciones. La métrica secundaria se muestra junto a la principal con tamaño de fuente más pequeño.

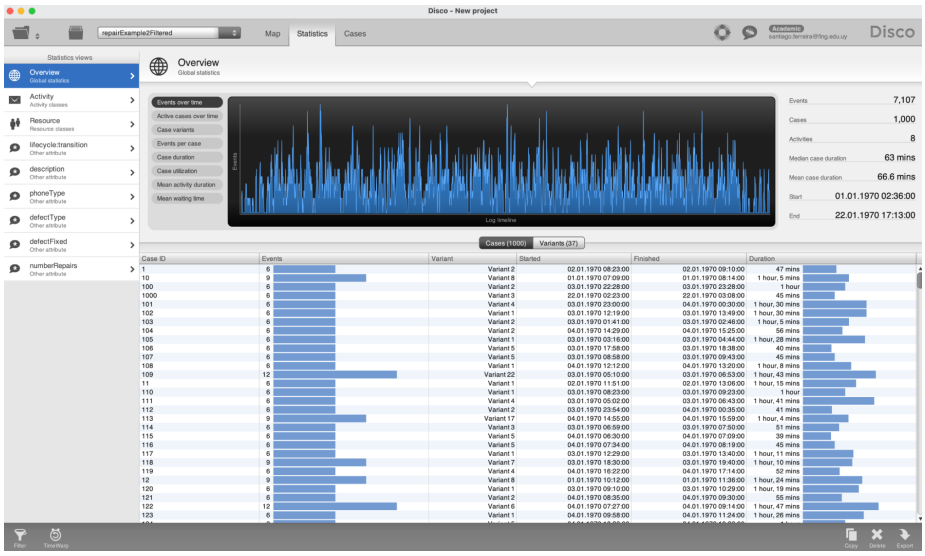


Figura 2.10: Interfaz de Disco. Vista de estadísticas de casos.

En la Figura 2.10 se muestra la vista de estadísticas, específicamente en la pantalla de "Overview" donde se puede ver por ejemplo la cantidad total de eventos y casos, la cantidad de actividades diferentes, la duración media y mediana de los casos y el rango de tiempo considerado en los datos. También se muestran gráficos pregenerados con información relevante como eventos a lo largo del tiempo, casos activos a lo largo del tiempo y cantidad de casos en función de variante, cantidad de eventos, duración del caso y otros. En la pantalla de actividades ("Activities") de la vista de estadísticas se presenta información más detallada de frecuencia y rendimiento de las actividades del proceso. La pantalla de recursos ("Resources") es exactamente igual a la de actividades pero muestra información relativa al atributo que se haya configurado como recurso durante la importación de datos. Por último también están las pantallas de estadísticas de atributos que muestran información sobre cualquier atributo seleccionado

durante la importación de datos que no se haya establecido como actividad, caso, recurso o tiempo (observar en la Figura 2.10 los seis atributos debajo de "Resources").

Debido a las cualidades de cada una de estas herramientas, es que en general se utiliza Disco para realizar un análisis rápido del log y generar modelos. ProM, en cambio, se utiliza para un análisis más en detalle, debido a la cantidad de plugins que existen. Por ejemplo, se puede elegir qué tipo de algoritmo utilizar para descubrir un modelo, también contiene más tipos de filtros que Disco y una gran variedad de otras opciones que en Disco no se encuentran (e.g análisis del proceso desde el punto de vista organizacional).

2.3. Análisis de problemas de movilidad

Una de las posibilidades que brindan los datos disponibles de sistemas de movilidad, como el STM, es la obtención de datos exactos de demanda por hora y zona con la granularidad deseada tanto en espacio como en tiempo, a diferencia de los datos de encuestas que no son tan exactos y requieren un gran esfuerzo. Esto es de gran importancia para el planeamiento estratégico.

Estos datos además permiten rastrear el comportamiento individual, teniendo las horas y ubicaciones de subida de los pasajeros y en algunos casos también las de bajada o una estimación en su defecto. Con estos datos es posible realizar diversos tipos de análisis como identificación de paradas populares de transbordo, combinaciones frecuentes de líneas, etc.

A su vez los datos de las tarjetas permiten obtener información sobre el recorrido de los ómnibus ya que se tiene las horas de pasada por paradas a lo largo del recorrido. En base a esto se pueden analizar desfases (adelantos o atrasos) y amontonamiento de ómnibus por ejemplo. Si además se cuenta con la información o estimación de bajada de los pasajeros es posible tener información tal como la cantidad de gente que viaja sobre el ómnibus, dando lugar a otros análisis de interés.

Una de las dificultades que se tiene al trabajar con datos de tarjetas inteligentes es el volumen de datos (big data), el cual puede complejizar su procesamiento. Otra dificultad puede ser directamente el acceso a los datos ya que las tarjetas pueden contener información sensible. También puede haber información faltante debido a varias causas, por ejemplo el dato de bajada del pasajero el cual no se tiene en el STM y varios otros sistemas así como otros datos socio-demográficos que podrían ser de interés. Es frecuente que las tarjetas prepagas como la STM incluyan pocos o ningún dato socio-demográfico. Por último también puede existir en ciertos lugares la dificultad de que un porcentaje alto de personas no utilicen la tarjeta por lo que sea difícil lograr una representación fiable de la población entera.

Hay varios estudios en el área de utilización de los datos de las tarjetas inteligentes de transporte. En (Pelletier, Trépanier, y Morency, 2011) se presentan varios de estos estudios definiendo los datos utilizados, el análisis realizado y los potenciales beneficios. Se categoriza a los estudios en 3 tipos: de nivel es-

tratégico, útiles para planeamiento a largo plazo como estimación de demanda; de nivel táctico, por ejemplo para ajustes de frecuencia y rutas; y de nivel operacional como pueden ser estadísticas de cantidad de pasajeros, indicadores de rendimiento (ej: adelantos, atrasos), etc.

En (Li, Sun, Jing, y Yang, 2018) se presenta un resumen de los estudios enfocados en la estimación del destino de los pasajeros. Se dividen los estudios en tres tipos según el modelo de inferencia utilizado (encadenamiento de viajes, probabilístico y aprendizaje profundo). Se realiza una comparación de estos estudios teniendo en cuenta factores tales como los tipos y la cantidad de datos utilizados, los distintos problemas encontrados, el proceso de preparación de datos y el método de validación de resultados. Los métodos de validación de resultados se basan en general en comparación contra encuestas, usualmente utilizados en sistemas con solo el dato de subida al ómnibus, o comparación con el dato real, utilizado para sistemas donde se tiene la subida y la bajada de los pasajeros.

Algunos de los objetivos de los estudios existentes sobre utilización de datos de tarjetas inteligentes incluyen la estimación del comportamiento de los pasajeros, dentro de la cual se tiene la estimación del destino de pasajeros, la generación de la matriz origen-destino, la inferencia de actividad de los pasajeros (viajes dirigidos a realizar una cierta actividad), análisis de patrones de viaje entre grupos de pasajeros (por ejemplo mediante clustering de trayectorias) y otros.

El libro (Kurauchi y Schmoecker, 2021) recopila gran cantidad de información de los estudios sobre el uso de datos de las tarjetas inteligentes para el transporte. Está dividido en tres secciones donde la primera se enfoca en los métodos de estimación del comportamiento de los pasajeros, la segunda en los métodos y potencial utilidad de combinación de los datos de las tarjetas con otros como los de encuestas y la tercera se focaliza en la utilización de los datos para la evaluación del sistema de transporte desde la perspectiva del usuario, como por ejemplo la evaluación de los cuellos de botella en las paradas.

En (Agard, Morency, y Trépanier, 2006) se utilizan conceptos de minería de datos para obtener detalles sobre los datos generados en un sistema SCAFC (Smart Card Automated Fare Collection), mismo tipo de sistema que STM. A través de algoritmos de clustering se segmentan los usuarios en grupos de usuarios similares respecto a los viajes que realizan. Luego se discriminan cada uno de estos grupos con respecto al tipo de usuario (ADULT, STUDENT, ELDERLY) y se realizan observaciones. También se analiza la variabilidad de los grupos obtenidos a través del clustering en un rango de tiempo de 3 meses (tomando sólo usuarios de tipo STUDENT). Resultados interesantes se desprenden, al notar por ejemplo que en una semana en particular uno de los grupos es radicalmente distinto que en las otras semanas, teniendo como causa que esa semana eran vacaciones escolares.

Un análisis tradicional y que es de interés en todos los sistemas de transporte es el conocer los patrones de movimiento de los usuarios. Esto permite que se puedan tomar mejores estrategias al hacer la planificación del transporte público, generar acciones concretas (e.g agregar más recursos en determinadas

zonas de una ciudad). El gran volumen de datos que se genera en los sistemas SCAFC hacen que este análisis pueda ser realizado. Numerosos trabajos (Alsger, Mesbah, Ferreira, y Safi, 2015; Berlingerio y cols., 2013; Farzin, 2008; Barry, Newhouser, Rahbee, y Sayeda, 2002; Nassir, Khani, Lee, Noh, y Hickman, 2011; J. Zhao, Rahbee, y Wilson, 2007) han intentado estimar la matriz Origen-Destino (O-D Matrix). Esta matriz cuenta la cantidad de viajes realizados desde un origen a un destino, posibilitando el análisis de patrones de movimientos. La matriz O-D se puede generar con base en encuestas, pero hacerlo mediante datos de los sistemas SCAFC tiene varias ventajas. Las encuestas son costosas, sobre todo si se quieren realizar con cierta periodicidad para tener datos actualizados, mientras que con los datos de sistemas SCAFC la matriz O-D se podría calcular diariamente. Además estos datos permiten generar la matriz O-D por rangos de días u horarios y sobre distintas divisiones geográficas de interés. Más allá de esto, las encuestas pueden obtener datos de interés que no son fáciles de inferir en base a los datos presentes en el sistema SCAFC, tales como los motivos de los viajes.

Una complejidad que se añade en muchos de estos sistemas SCAFC (incluido el STM) es que la única información que se genera es la de subida de un pasajero a un transporte público, dificultando así la correlación de viajes que tienen muchos tramos (varias subidas a transportes distintos). Muchos de estos trabajos intentan resolver esta dificultad de diferentes formas, pero el algoritmo base es el mismo, un algoritmo heurístico que toma ciertos parámetros (tiempo de un viaje y tiempo entre dos viajes) para reconstruir viajes de múltiples tramos. Variando estos parámetros se obtienen distintos resultados. Este análisis puede tener distintos enfoques, por ejemplo analizar por áreas de la ciudad o teniendo en cuenta factores adicionales, brindando mayor información a los expertos.

También existen varios artículos que utilizan la disciplina de minería de datos para extraer información en el sector del transporte. Por ejemplo, en (K. Zhao, Tarkoma, Liu, y Vo, 2016) se analizan registros de recorridos de taxis dando posibles aplicaciones como predicción de tráfico o detección de zonas de la ciudad (residencial, negocios, educación). De forma similar en (Wang, Lo, y Liu, 2015) se analizan varias estaciones de metro en base a los registros de las tarjetas y analizando la cantidad de usuarios, los períodos de tiempo de mayor uso y otros datos más, se establecen relaciones y características entre las estaciones. Por ejemplo, movimiento de flujo entre estaciones en áreas de negocios a áreas de ocio, diferentes horarios picos según su ubicación y día de la semana, etc. También en (Foell y cols., 2013) se construyen modelos para tratar de predecir los viajes que van a realizar los usuarios, recomendarles acciones, así como también predecir desvíos en el tránsito.

Estos últimos trabajos están enfocados en otra área importante de la minería de procesos, como son las predicciones y realización de recomendaciones. Sin embargo, la mayoría de los estudios encontrados a la fecha utilizan minería de datos y pocos utilizan minería de procesos, abriendo la posibilidad a una nueva línea de investigación de movilidad urbana utilizando minería de procesos (Bădică, Bădică, Buligiu, y Ciora, 2022).

Capítulo 3

Extracción

En esta sección se presenta la estructura inicial de los datos disponibles, así como también el proceso realizado mediante la implementación y evaluación de un algoritmo, con el objetivo de reconstruir datos faltantes.

3.1. Datos disponibles

En el Catálogo Nacional de Datos Abiertos[1], filtrando por "Categoría = Transporte" y "Publicador = Intendencia de Montevideo" se pueden encontrar varios conjuntos de datos del STM que se analizaron como posibles datos de interés para los objetivos planteados, destacando en particular el registro de ascensos a ómnibus en las distintas paradas y los datos de recorridos y horarios de ómnibus. En el Catálogo de datos geográficos de Montevideo[2] se puede obtener variedad de capas geográficas públicas, siendo las de mayor interés las capas de zonificaciones (barrios, códigos postales, municipios, etc) utilizados en la extensión de logs, que se encuentra explicado en siguientes secciones. A continuación se describen los conjuntos de datos más relevantes que se encontraron.

[VROSTM] Viajes realizados en los ómnibus del STM[3]. Contiene todos los registros de ascenso de pasajeros en las líneas de transporte colectivo urbano de Montevideo. La información suministrada proviene de la totalidad de registros procesados por las máquinas de validación de viajes del STM. Los datos de ascensos están disponibles en formato CSV, por mes y se cuenta con registros desde octubre del 2020 hasta el último mes finalizado. Este es el conjunto de datos más importante para el trabajo realizado.

- *id_viaje:* identifica de forma única a un viaje dentro de un mes. Definiendo por viaje, todos los tramos realizados por el/los pasajeros con un único pago.

[1] https://catalogodatos.gub.uy/

[2] https://geoweb.montevideo.gub.uy/geonetwork

[3] https://catalogodatos.gub.uy/dataset/intendencia-montevideo-viajes-realizados-en-los-omnibus-del-stm

- *ordinal_de_tramo:* para viajes con tarjeta, ordinal del tramo dentro del viaje.
- *cantidad_pasajeros:* cantidad de personas que realizan el tramo
- *fecha_evento:* fecha y hora en la cual se registra el ascenso.
- *con_tarjeta:* indica si se utilizó o no tarjeta para abonar el viaje.
- *codigo_parada_origen:* código de la parada en la que se registra el ascenso.
- *cod_empresa:* código de empresa transportista a la cual pertenece el ómnibus.
- *linea_codigo:* código de la línea del ómnibus.
- *sevar_codigo:* código de la variante de la línea del ómnibus.
- *grupo_usuario:* grupo de usuario (ej. Estudiante, Jubilado, Usuario Corriente).
- *grupo_usuario_específico:* subgrupo específico dentro del grupo de usuario.
- *tipo_viaje:* tipo de viaje (ej. 1 hora, 2 horas, céntrico, común).

[HOUP] Horarios de ómnibus urbanos, por parada - STM[4]. Contiene los horarios de ómnibus del transporte colectivo urbano de Montevideo para cada una de las paradas. Estos son los horarios teóricos estimados en los que pasará una determinada línea de ómnibus por una cierta parada a lo largo de su recorrido. Estos datos se estiman en base a los horarios predefinidos de las líneas, la velocidad promedio de las unidades de transporte y la distancia entre paradas. Este juego es actualizado diariamente y no se cuenta con el histórico, solo con la última versión. Este conjunto de datos también fue utilizado.

- *tipo_dia:* día hábil, sábado o domingo.
- *cod_variante:* código de variante de la línea.
- *frecuencia:* identifica mediante la hora de salida a la frecuencia de la variante para el tipo de día.
- *cod_ubic_parada:* código de la parada.
- *ordinal:* número ordinal de la parada dentro del recorrido de la variante.
- *hora:* hora estipulada de pasada del ómnibus por la parada.
- *dia_anterior:* indica si el ómnibus (frecuencia) comienza el recorrido el día anterior al que pasa por la parada.

[LOOD] Líneas de ómnibus, origen y destino[5]. Contiene información de origen y destino de las variantes de cada línea así como información geográfica con los recorridos de cada variante.

[EUCI] Estadísticas de uso: Cómo Ir[6]. Contiene las estadísticas de uso de la aplicación Cómo Ir desarrollada por la IM. Los datos incluyen tanto las consultas de ruteo (de un punto a otro de la ciudad), como las de horarios (saliendo o llegando a una determinada hora).

[4] https://catalogodatos.gub.uy/dataset/intendencia-montevideo-horarios-omnibus-urbanos-por-parada-stm

[5] https://catalogodatos.gub.uy/dataset/intendencia-montevideo-lineas-omnibus-origen-y-destino

[6] https://catalogodatos.gub.uy/dataset/intendencia-montevideo-estadisticas-de-uso-comoir

[**EOD**] **Encuesta Origen-Destino Montevideo**[7]. Contiene datos de las encuestas origen destino realizadas en los años 2009 y 2016.

[**SPM**] **Shapefile de paradas de Montevideo**[8]. Mapa digital que contiene la ubicación de las paradas de ómnibus de Montevideo. Esta información es proporcionada por el Plan de Movilidad Urbana de Montevideo.

[**SCPM**] **Shapefile de códigos postales de Montevideo**[9]. Mapa digital que contiene la delimitación geográfica y el número correspondiente a los Códigos Postales de la ciudad de Montevideo, según información provista por la Administración Nacional de Correos.

[**SMM**] **Shapefile de municipios de Montevideo**[10]. Mapa digital que contiene los límites geográficos correspondientes a los Municipios de Montevideo según Decreto No. 33227 de la Junta Departamental de Montevideo.

[**SBM**] **Shapefile de barrios de Montevideo**[11]. Mapa digital que contiene los límites correspondientes a los Barrios de la ciudad de Montevideo según definición del Instituto Nacional de Estadística (INE). Se actualiza coincidentemente con los Censos Nacionales.

Además de los datos disponibles públicamente, fueron importantes para el estudio los datos de ascensos con información extendida brindados por la IM para el mes de mayo de 2022. Estos datos son iguales a los presentes en VROSTM pero tienen datos que no están disponibles públicamente.

[**VROSTM_F**] **Viajes realizados en los ómnibus del STM con datos de frecuencia.** Contiene los mismos datos que VROSTM más los siguientes:

- *nro_frecuencia:* equivalente a 'frecuencia' de HOUP. Identifica la frecuencia de la variante.
- *numero_evento_recorrido:* id único de una instancia de un recorrido de una frecuencia, es decir, identifica el viaje de una frecuencia para cierto día.

[**VROSTM_FT**] **Viajes realizados en los ómnibus del STM con datos de frecuencia y tarjeta.** Contiene los mismos datos que VROSTM_F más el siguiente.

- *id_tarjeta:* id anonimizado de la tarjeta utilizada en el ascenso

3.2. Algoritmo de identificación de buses

Un problema encontrado en los datos de ascensos (VROSTM) es que no tienen una identificación del ómnibus con la que se pueda distinguir un ómnibus

[7] https://catalogodatos.gub.uy/dataset/intendencia-montevideo-encuesta-origen-destino-montevideo

[8] https://geoweb.montevideo.gub.uy/geonetwork/srv/spa/catalog.search#/metadata/c6ea0476-9804-424a-9fae-2ac8ce2eee31

[9] https://geoweb.montevideo.gub.uy/geonetwork/srv/spa/catalog.search#/metadata/3e637c6d-b59b-4d87-abe0-d39bafc99ac9

[10] https://geoweb.montevideo.gub.uy/geonetwork/srv/spa/catalog.search#/metadata/b0a2cf85-af7a-4aac-998f-da124ac7d073

[11] https://geoweb.montevideo.gub.uy/geonetwork/srv/spa/catalog.search#/metadata/1277c8cd-3e7a-4afd-8289-aeae893ce0db

particular de los demás ómnibus de la misma variante. El primer objetivo era poder describir los recorridos de los ómnibus como procesos y para esto se debe primero poder identificar un ómnibus. Por este motivo se decidió implementar un algoritmo para intentar identificar los distintos ómnibus de una misma variante.

Más adelante se logró conseguir los datos con la frecuencia del ómnibus de cada registro para el mes de mayo de 2022 (ver conjunto de datos VROSTM_F). Esto permitió realizar un análisis más exacto de los datos y también comparar los resultados de inferencia del algoritmo implementado respecto a los datos reales.

Para identificar los distintos ómnibus de una misma variante se probó con distintos enfoques, pero todos ellos apoyados en la información de las horas estipuladas de salida y pasada por cada parada de cada variante. Un problema encontrado en los datos de frecuencias y horarios (HOUP) es que son actualizados diariamente, con lo cual puede haber diferencias importantes si se contrastan contra los registros de ascensos (VROSTM) anteriores o posteriores. Otra complejidad es que un ómnibus podría alcanzar o incluso adelantar a otro. Del análisis de datos se sabe que para algunas variantes existen frecuencias contiguas con poca diferencia entre sus horarios de salida. Además, hay que contemplar que en ocasiones en una parada sube mucha gente, con lo cual el último en marcar ascenso en esa parada podría estarlo haciendo incluso cuando ya comenzó a subir gente de la siguiente. Durante el análisis se plantearon varios enfoques, ninguno de los cuales contemplaba el posible adelanto de un ómnibus sobre otro. A continuación se describe el algoritmo que fue finalmente utilizado.

3.2.1. Detalles de implementación

El algoritmo recorre todas las variantes de una lista de variantes que se le especifiquen procesando de a un día. Para cada una de estas variantes se obtienen todas sus frecuencias y se las recorre. Para cada frecuencia se recorren todos los registros de ascenso de la variante en el día analizado. Un ascenso se considera que pertenece a la frecuencia si la hora a la que se produce está dentro de cierto umbral (máximo adelanto o retraso permitido) respecto a la hora de pasada estipulada de dicha frecuencia por la parada en la que se produce el ascenso y siempre que se siga el orden correcto de paradas según el recorrido que realiza la variante respecto a los ascensos ya asociados a esta frecuencia.

La selección del umbral es importante dado que un umbral pequeño ocasionará que haya más ascensos sin frecuencia asignada, pero un umbral muy grande generará ascensos mal clasificados. Por este motivo, el algoritmo comienza con un valor de umbral pequeño para cada día, que luego irá incrementando al final de cada iteración de todas las frecuencias de la variante actual. El algoritmo pasará al siguiente día cuando haya logrado asignar una frecuencia de ómnibus a cada ascenso de la variante actual o bien cuando el valor del umbral alcance un valor definido como cota superior. Se utilizó 5 minutos como umbral inicial, así como incremento y se utilizó 45 minutos como umbral máximo.

Un problema del algoritmo se da cuando dos ómnibus de la misma variante se aproximan mucho. En estos casos la prioridad para asignación del ascenso la

tendrá el primero, ya que las frecuencias se recorren en orden de salida. Otra ventaja del algoritmo es el tiempo que demora. Los primeros incrementos del umbral logran clasificar la mayoría de los ascensos, pero en general se continúa hasta llegar al umbral máximo. Son mínimos los casos en los que se logran clasificar todos los ascensos para una variante en un día antes de llegar al umbral máximo. Los umbrales grandes logran clasificar ciertos ascensos con gran desfase pero coherentes en el desfase a lo largo del recorrido del ómnibus. Esto es probablemente resultado de la discrepancia entre los datos de horarios actualizados y los horarios reales de la variante. En el anexo IB1 se presenta un pseudocódigo del algoritmo implementado reducido al procesamiento de una variante en un día, ya que el procesamiento del resto es equivalente.

3.2.2. Posibles mejoras

Uno de los problemas del algoritmo anterior ocurre cuando dos frecuencias se aproximan mucho. Una posible mejora para esto es manejar un tiempo mínimo y máximo entre los ascensos en una parada.

Otra posible mejora es utilizar umbrales independientes por parada, teniendo en cuenta los horarios estipulados de pasada por parada de la frecuencia que se está analizando la anterior y la que le sigue, realizando un incremento del umbral no fijo, sino como un porcentaje de la diferencia entre los tiempos de pasada por parada estipulados de las frecuencias contiguas.

También se puede definir una diferencia máxima aceptable entre los adelantos o retrasos en las paradas consecutivas de un recorrido. Para establecer el máximo se pueden tener en cuenta diversos factores tales como la distancia entre las paradas, velocidad promedio, etc.

3.2.3. Análisis de la ejecución

Como se vio, el algoritmo no solo deja ascensos sin frecuencia de variante asociada, sino que es propenso a varios errores en la inferencia de las frecuencias. Es necesario tener un método de estimación de cuán confiables son los resultados. El método más exacto y sencillo es lógicamente comparar contra los datos reales. Como ya se mencionó, avanzado el proyecto se pudo conseguir los datos de ascensos con el dato de frecuencia para el mes de mayo de 2022 y, por lo tanto, se utilizaron para ver cuán certero era el algoritmo. Previo a la obtención de los datos se pensaron posibles criterios y métricas para la evaluación de los resultados, estos pueden verse en detalle en el anexo IB2.

La comparación de los resultados del algoritmo con los datos reales se hizo para tres corredores por separado. Para Av. Italia, el algoritmo asignó frecuencia a 3.623.635 de un total de 3.730.394 ascensos (97,1 %). De estas, un 85,4 % fueron asignaciones correctas. Para Gral. Flores, el algoritmo asignó frecuencia a 4.247.272 de un total de 4.435.698 ascensos (95,75 %). De estas, un 86,2 % fueron correctas. Por último, para el corredor Agraciada/Garzón, el algoritmo asignó 9.814.360 de un total de 10.450.582 ascensos (93,91 %). De estas, un 78,39 % fueron correctas.

Hay que tener en cuenta que el algoritmo se basa en los datos de recorridos y horarios estipulados que, como se comentó, se actualizan periódicamente y presentan diferencias respecto a los datos de ascenso analizados. Analizando los datos reales y comparando con los datos de horarios, se vio que del total de ascensos de mayo de 2022, hay un 9,6 % para los cuales el par {*variante, frecuencia*} no se encuentra en la tabla de horarios. Restringiendo a Av. Italia, por ejemplo, este porcentaje es de 8,4 %. De esto se concluye que el porcentaje real de frecuencias correctamente inferidas sería mayor en caso de contar con los datos de horarios correctos, en el caso de Av. Italia, por ejemplo, el porcentaje de acierto podría llegar a 93,8 %.

Capítulo 4

Inspección y limpieza

En esta sección se presenta un análisis de los datos contenidos en los logs, así como también se desarrolla el proceso de limpieza, filtrado y enriquecimiento de los logs. También se mencionan brevemente algunos errores encontrados en los datos.

4.1. Inspección general de los datos

Se realizó un análisis exploratorio de los datos de viajes (VROSTM) para mayo de 2022, que como ya se mencionó es el mes para el cual se consiguió información de frecuencia de ómnibus (VROSTM_F) y tarjeta de pasajero (VROSTM_FT) y es por ese motivo el mes sobre el que se realizaron los estudios.

Para mayo se tiene un total de 25.009.694 de registros, donde el primero es de las 22:01:02 del 01/05/2022 y el último es de las 23:59:58 del 31/05/2022.

Al analizar la cantidad de ascensos por día del mes se encuentran algunas particularidades. En la Figura 4.1 se muestran los ascensos por día del mes. Se observa cómo el primero de mayo (feriado del día del trabajador) no hay prácticamente ascensos, lo cual es lógico. También se puede ver cómo en general la cantidad de ascensos de lunes a viernes es pareja, pero hay algunas excepciones tales como los días 16, 17 y en menor medida el día 25. El lunes 16 de mayo hubo un feriado laborable por la batalla de Las Piedras, el cual podría ser el motivo de la importante disminución de ascensos observada. En cuanto al 17 y al 25, no hay motivos tan evidentes que expliquen la cantidad de ascensos reducida.

En la Figura 4.2 se muestra la cantidad de ascensos por día de la semana. Para esto se excluye el día primero de mayo y se toma hasta el 29 de mayo inclusive. De este modo se tienen cuatro semanas completas de lunes a viernes, contando cuatro veces cada día.

En la Figura 4.3 se puede ver específicamente la distribución de ascensos entre días hábiles, sábados y domingos. El intervalo de tiempo considerado es desde el 2 de mayo hasta el 29.

Figura 4.1: Cantidad de ascensos por día del mes para mayo de 2022.

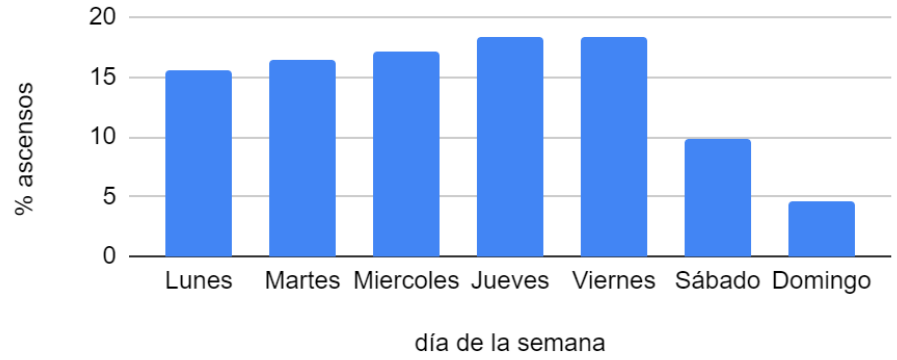


Figura 4.2: Cantidad de ascensos por día de la semana para mayo de 2022.

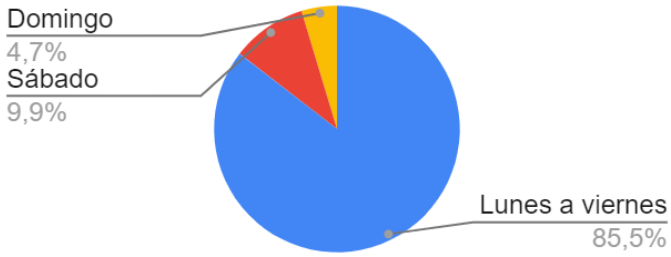


Figura 4.3: Cantidad de ascensos según día hábil, sábado o domingo para mayo de 2022.

En la Figura 4.4 se muestra una gráfica con la cantidad de ascensos por hora del día, separando para días de lunes a viernes, sábados y domingos. Nuevamente, se considera entre el 2 de mayo y el 29 de mayo para considerar 4 veces cada día. La diferencia más evidente es la mayor cantidad de ascensos durante la mañana los días de lunes a viernes, lo cual es totalmente lógico. También se puede ver una disminución más pronunciada en la cantidad de ascensos de lunes a viernes a partir de las 18 respecto a sábados y domingos. Es interesante observar como los ascensos a horas muy tempranas de la mañana (3 a 5 am)

son proporcionalmente iguales para los tres casos.

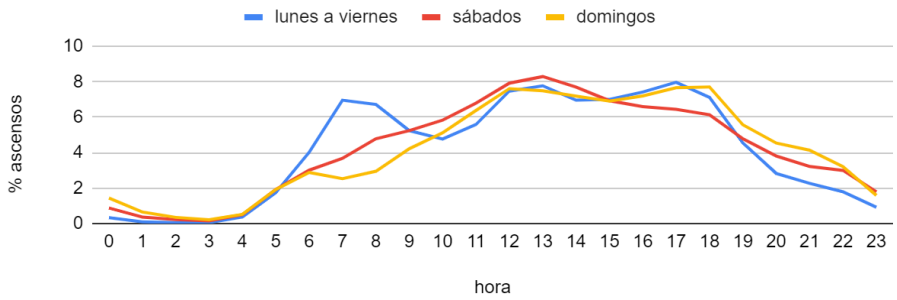


Figura 4.4: Cantidad de ascensos por hora del día para mayo de 2022.

En la Figura 4.5 se puede ver la distribución de ascensos según las distintas empresas de transporte. Observar como CUTCSA registra 2 de cada 3 ascensos.

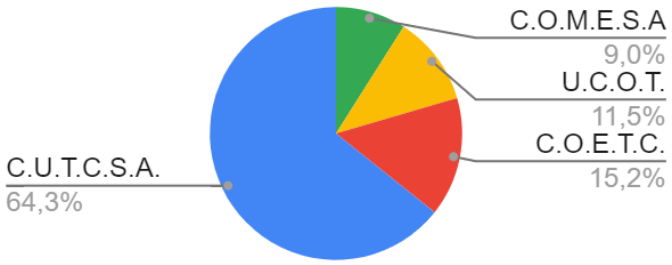


Figura 4.5: Distribución de ascensos según empresa de transporte.

Los viajes en ómnibus se pueden realizar con o sin tarjeta. Para mayo de 2022 la gran mayoría de los ascensos registrados son con tarjeta (ver Figura 4.6).

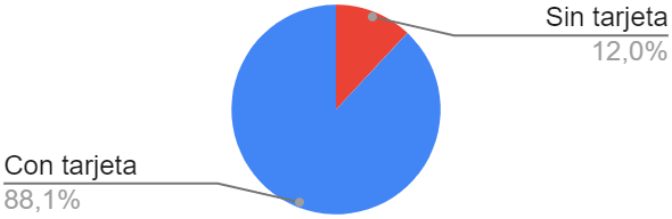


Figura 4.6: Cantidad de registros de ascensos con y sin tarjeta.

Un boleto puede ser utilizado para más de un pasajero. Como se puede ver en la Figura 4.7, un 97 % de los pasajes son para una persona. Hay un 2,5 % de pasajes para 2 personas. Luego se observan pasajes de entre 3 y 30 personas que representan una fracción mínima del total (0,2 %).

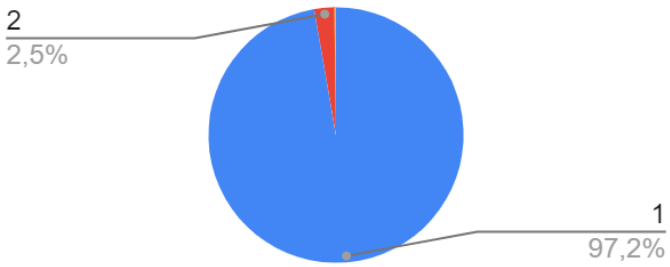


Figura 4.7: Cantidad de registros de ascensos según cantidad de pasajeros.

Al realizar un viaje con tarjeta se puede con el mismo boleto tomar 1, 2 o más ómnibus dependiendo de la tarjeta y del tipo de viaje solicitado. En caso de efectuar uno o más transbordos con la misma tarjeta dentro del mismo viaje pago, se guarda registro del número (ordinal) de tramo del viaje correspondiente al ascenso. El primer ascenso siempre tiene ordinal de tramo 1, a partir de aquí el tramo incrementa en 1 en cada transbordo sucesivo del mismo viaje pago. En la Figura 4.8 se puede ver la distribución de viajes con tarjeta según cantidad de tramos.

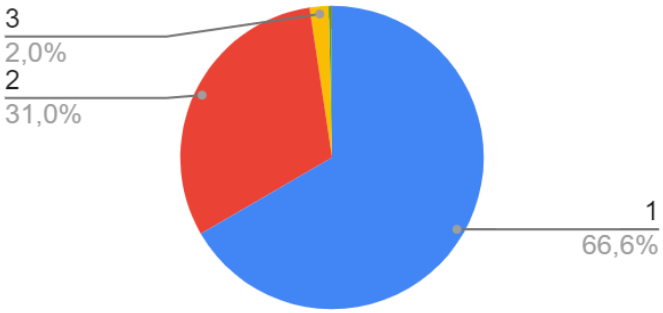


Figura 4.8: Cantidad de viajes según cantidad de tramos.

Los registros de ascenso con tarjeta también guardan el dato de *grupo_usuario* que corresponde al tipo de tarjeta utilizada. En la Figura 4.9 se puede ver la distribución en los ascensos de los principales grupos de usuario. Asociado al grupo de usuario está el grupo de usuario específico cuya distribución se puede ver en la Figura 4.10.

Los pasajeros pueden con una misma tarjeta seleccionar entre distintos tipos de boleto, esto se registra en los datos de ascensos como *tipo_viaje*. En la Figura 4.11 se puede ver la distribución de los tipos de viajes para los viajes sin tarjeta y en la Figura 4.12 para los viajes con tarjeta.

Se analizan también las principales líneas y variantes presentes en los registros de viajes de mayo. En la Tabla 4.1 se pueden ver las 10 líneas con más

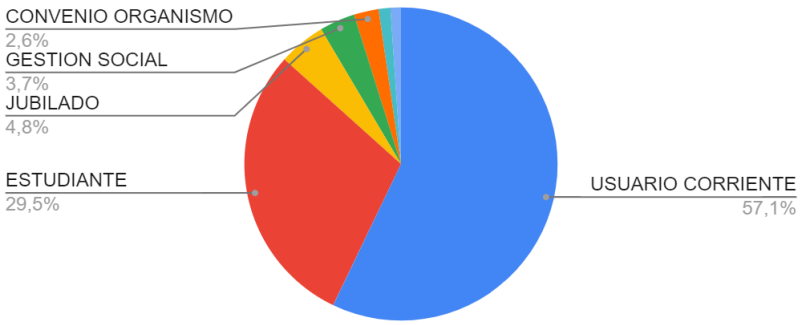


Figura 4.9: Distribución de ascensos según grupo de usuario.

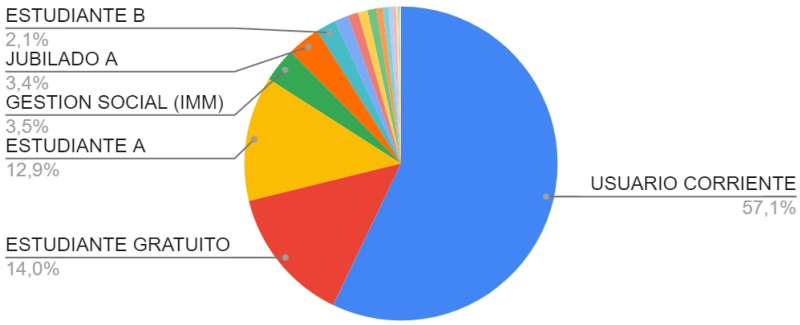


Figura 4.10: Distribución de ascensos según grupo de usuario específico.

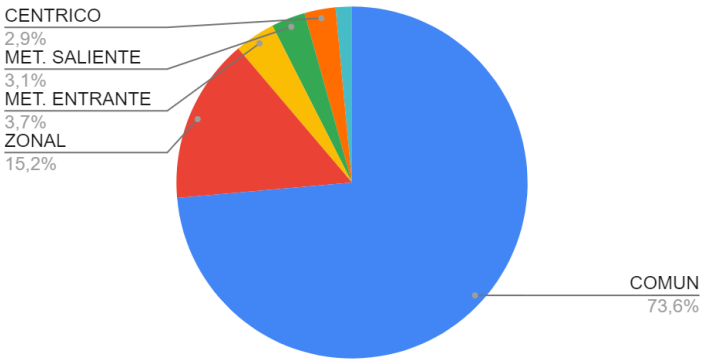


Figura 4.11: Distribución de ascensos para los viajes sin tarjeta.

ascensos, la columna "% Ascenso" muestra el porcentaje total de ascensos respecto al total de ascensos del mes. En la Tabla 4.2 se muestran las 10 variantes que registraron más ascensos.

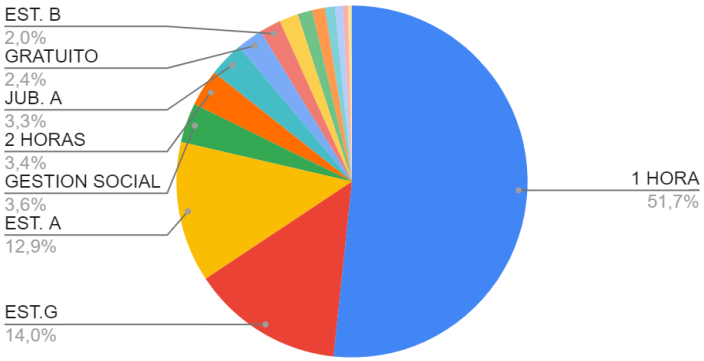


Figura 4.12: Distribución de ascensos para los viajes con tarjeta.

Tabla 4.1: Top 10 líneas con más ascensos en mayo 2022.

Línea	% de ascensos
103	3,5
G	2,7
185	2,4
306	2,4
183	2,4
145	2,3
181	2,3
300	2,2
163	2,0
405	2,0

Tabla 4.2: Top 10 variantes con más ascensos en mayo 2022.

Variante	Línea	% de ascensos
7603	181	1,27
8389	183	1,25
8401	183	1,10
1761	306	1,09
8398	300	1,09
8385	300	1,07
7666	306	1,05
1347	185	1,04
7602	181	0,99
1337	185	0,98

En la Figura 4.13 se puede ver el porcentaje de ascensos abarcado según el porcentaje de líneas (tomando de una tabla ordenada por cantidad decreciente

de ascensos). Se ve cómo el 20 % de líneas con más ascensos concentra más de la mitad de los ascensos y que el 50 % abarca más del 90 %. La Figura 4.14 es equivalente pero para las variantes. Notar cómo un porcentaje pequeño de variantes abarca la mayoría de los ascensos, hay muchas variantes con muy poca cantidad de ascensos.

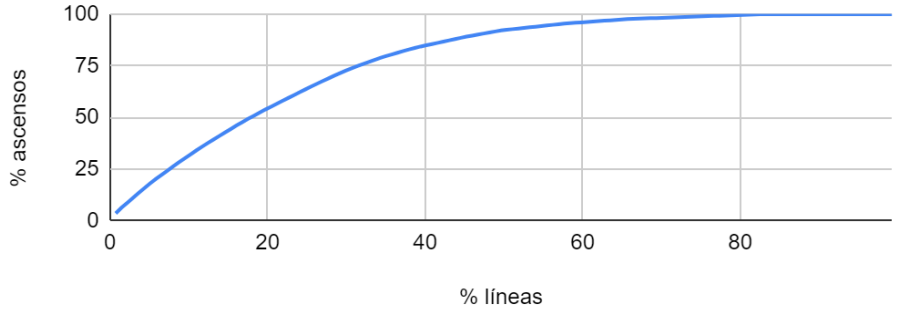


Figura 4.13: Porcentaje de ascensos según porcentaje de líneas.

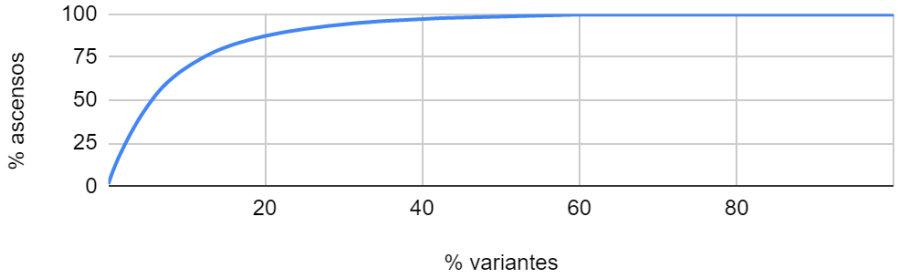


Figura 4.14: Porcentaje de ascensos según porcentaje de variantes.

4.2. Errores encontrados en los datos

Analizando más en profundidad los datos de viajes de mayo (VROSTM), se encontraron algunos errores, de los cuales unos afectan a un porcentaje muy pequeño de registros y otros bastante más frecuentes.

Se encontraron 14 viajes con cantidad de pasajeros igual a cero, que es una cantidad menor y cuyos otros datos no presentan problemas. Otro error menor fueron los viajes con {*id_viaje, ordinal_de_tramo*} repetido, encontrado en solamente 16 registros (8 casos de 2 registros cada uno). La cantidad de registros afectados es muy mínima en ambos casos, pudiendo ignorar el segundo caso y asumir cantidad de pasajeros igual a 1 en el primero con error insignificante.

Existen 427 viajes con tramos faltantes, es decir, viajes multitramo donde la cantidad de registros es menor al máximo ordinal de tramo para ese ID de viaje. Los registros correspondientes a estos viajes suman 1049.

En cuanto a las paradas, se encontraron 58 paradas en los datos de viajes (VROSTM) que no están presentes en los datos de recorridos y horarios (HOUP), de las cuales 57 tampoco se encuentran en los datos de ubicaciones geográficas de las paradas (SPM). Estas paradas aparecen en 44.886 registros de la tabla de viajes (0.18 % del total de registros). Esto impide tomar estos registros en cuenta para los análisis que toman en cuenta la ubicación geográfica. Existen maneras de aproximar algunas o todas las ubicaciones de estas paradas, por ejemplo buscando en los registros las paradas conocidas previas y posteriores más cercanas entre sí. Esto no se hizo dado que la cantidad de registros afectados es mínima.

Un error encontrado más frecuentemente es el de pares {*variante, frecuencia*}, presentes en la tabla de viajes (VROSTM), pero no en la tabla de recorridos y horarios (HOUP). De un total de 35.660 pares {*variante, frecuencia*} existentes en la tabla de viajes, el 9 % (3.377) no se encuentran en la tabla de recorridos y horarios. Estos pares de variante y frecuencia aparecen en 2.394.598 registros (9,6 % del total). Esto impide calcular el desfase (adelanto o retraso) para dichas frecuencias.

Un error relacionado con el anterior, pero más específico, son algunas tuplas {*variante, frecuencia, parada*} encontradas en la tabla viajes, pero no en la de recorridos y horarios. Lo cual al igual que en el caso anterior, impide el cálculo de desfase para la frecuencia, pero en este caso solo en cierta parada. Se encontraron en esta condición 4.829 tuplas de un total de 1.535.414 (0,31 %) que aparecen en 101.960 registros (0,4 %).

Como se puede ver, los errores encontrados en el conjunto de datos de viajes (VROSTM) son mínimos. El error que afecta a un porcentaje considerable de registros solo influye en el cálculo de desfase. Ninguno de los errores encontrados requiere limpieza de datos, los registros del conjunto de datos de viajes (VROSTM) se utilizan en su totalidad para la conformación de los logs utilizados.

En la sección 4.3 se describen los datos agregados a los registros de viajes (VROSTM) para la conformación del principal log de eventos utilizado. En la sección 4.4 se describe la agregación de eventos realizada para la creación del segundo log de eventos utilizado.

4.3. Enriquecimiento de logs

En la mayoría de los análisis es necesario extender el log con datos obtenidos de otras fuentes. Esto se da cuando los datos básicos no proveen información suficiente para realizar los análisis, por lo que se necesita extender el log agregando nuevas columnas.

El proceso para extender el log con datos externos es muy similar en todos los casos. Por ejemplo, para generar el log de la Tabla 4.3 es necesario tener otra

Tabla 4.3: Log de eventos inicial, extendido con la columna *Barrio de la Parada*.

ID del recorrido	ID de la parada	Fecha y Hora	Barrio de la Parada	Otros atributos
1	1000	01/05/2022 5:16	Unión	...
1	1000	01/05/2022 5.17	Unión	...
2	2000	01/05/2022 5.19	Tres Cruces	...
1	1001	01/05/2022 5.21	La Blanqueada	...
...	...	...	...	...

fuente de datos como el de la Tabla 4.4. Luego se realiza un JOIN por *ID de la parada* para obtener el atributo *Barrio* de la Tabla 4.4.

Tabla 4.4: Estructura de datos adicional, relacionado a las paradas.

ID de parada	Calle	Esquina	Barrio	Código Postal	Otros atributos
1	Canelones	Carlos Quijano	Centro	11100	...
2	Canelones	Julio Herrera y Obes	Centro	11100	...
3	Canelones	Paraguay	Centro	11100	...
4	Canelones	Convención	Centro	11100	...
...	...	...	...	...	...
1000	Comercio	Agustin Sosa	Unión	11400	...
...	...	...	...	...	...

Los datos de la Tabla 4.4 pueden ser obtenidos utilizando capas del tipo "Geographical Information System (GIS)" de uso libre, como los que provee la IM. En otras ocasiones suele ser necesario crear manualmente los datos, ya que son muy particulares al análisis.

Para este estudio se agregó a los datos de viajes (VROSTM) información geográfica relacionada a la parada en donde se produce el ascenso. Para esto primero se realizó un JOIN por el identificador de la parada entre los datos de viajes (VROSTM) y la capa geográfica de paradas (SPM) para tener la ubicación geográfica de la parada. Teniendo la ubicación geográfica se pudo agregar a los datos los siguientes atributos:

- *cod_postal:* código postal en donde se encuentra la parada. Obtenido mediante JOIN espacial con la capa de códigos postales de Montevideo (SCPM).
- *municipio:* municipio en donde se encuentra la parada. Obtenido mediante JOIN espacial con la capa de municipios de Montevideo (SMM).
- *barrio:* barrio donde se encuentra la parada. Obtenido mediante JOIN espacial con capa de barrios de Montevideo (SBM).

Además de agregar datos externos, en muchas ocasiones es necesario o de utilidad extender el log con atributos derivados de la información del log en sí mismo. Para este estudio se agregó a los datos de viajes (VROSTM) ciertos campos que facilitaban algunos de los análisis. De este modo se agregó a los datos de viajes (VROSTM) los siguientes datos:

- *dia_mes:* día del mes. Obtenido de *fecha_evento*.

- *dia_sem:* día de la semana. Obtenido de *fecha_evento*.
- *hora:* hora del día entre 0 y 23. Obtenido de *fecha_evento*.
- *tipo_dia:* hábil, sábado o domingo. Obtenido de *fecha_evento*.

Además de los atributos anteriores, se agregaron a los datos de viajes (VROSTM) también otros tres que ayudaron en el análisis de corredores de Montevideo, el cual se expone en la sección 5.2:

- *tramo:* identifica un tramo (o segmento de calle) dentro de un corredor. Obtenido en función de la ubicación de la parada y la separación en tramos realizada (ver sección 5.2 para más detalle sobre la separación en tramos). Para las paradas que están fuera del corredor se asigna tramo nulo o un valor que indica desde qué tramo se salió del corredor, esto dependiendo del log de eventos.

- *sentido:* sentido de la línea de ómnibus sobre el corredor (ej. para Av. Italia se tienen viajes en sentido este y oeste, para 8 de Octubre se tiene noreste y suroeste). El sentido queda determinado por la parada y el corredor.

- *en_corredor:* es un campo booleano que indica si la parada se encuentra en el corredor analizado.

4.4. Agregación de eventos

Para analizar tiempos entre paradas y tramos de un recorrido suele ser mejor utilizar un agregado de eventos para cada traza. Como ejemplo simple de agregado de eventos es posible pensar en agrupar para cada recorrido de un ómnibus todos los eventos consecutivos en cada parada, dejando un solo evento por parada con fecha de inicio igual al primer ascenso registrado en esa parada y fecha de fin igual al último ascenso. Para observar este ejemplo de manera más visual se puede ver cómo con esta agregación se pasaría del log de la Tabla 4.5 al de la Tabla 4.6.

Tabla 4.5: Estructura de datos inicial.

ID del recorrido	ID de la parada	Fecha y Hora	Otros atributos
1	1000	01/05/2022 5:16	...
1	1000	01/05/2022 5.16	...
1	1000	01/05/2022 5.17	...
1	1001	01/05/2022 5.19	...
1	1001	01/05/2022 5.19	...
1	1001	01/05/2022 5.19	...
1	1002	01/05/2022 5.25	...
...	...	...	...

La agregación anterior puede entre otras cosas ayudar por ejemplo a estudiar el tiempo transcurrido en el cobro de boletos de las distintas paradas. También

Tabla 4.6: Estructura de datos de la Tabla 4.5, luego de realizar una agregación de eventos.

ID del recorrido	ID de la parada	Fecha y Hora inicio	Fecha y Hora fin	Otros atributos
1	1000	01/05/2022 5:16	01/05/2022 5:17	...
1	1001	01/05/2022 5.19	01/05/2022 5.19	...
1	1002	01/05/2022 5.25	01/05/2022 5.25	...
...	...	...	...	...

utilizando la agregación anterior pero, ignorando la fecha y hora de fin durante el descubrimiento de procesos, se ignora el tiempo transcurrido en el cobro de boletos (loop en las paradas), traspasándolo a la transición entre paradas, lo que puede también facilitar algunos análisis.

En este caso, para la generación de uno de los logs se utilizó una agregación similar a la del ejemplo anterior, pero en lugar de agrupar por recorrido y parada se hizo por recorrido y tramo, es decir, tomando para cada ómnibus solo el primer y último ascenso registrado en cada tramo para cada día, lo que ayuda en el análisis de tiempos sobre los tramos definidos. Para la generación de este log se utilizó un script que además de hacer la agregación se encarga de asignar nombres de tramo particulares a los ascensos fuera del corredor, de modo de poder identificar en qué tramos se abandonó el corredor. En la sección 5.2 se comenta la utilidad que presenta el estudio mediante la separación en tramos. La definición de los tramos sobre los corredores puede verse en la sección 5.2.1 para Av. Italia y 5.2.2 para Av. 8 de Octubre.

4.5. Logs utilizados

Durante el transcurso del estudio se pasó por diversas versiones del log de eventos, principalmente por disponibilidad de los datos y porque en diversas ocasiones se realizaron extensiones o alteraciones para análisis específicos. Más allá de esto, los logs relativos a los estudios presentados en el informe se pueden resumir en tres distintos.

Como ya se mencionó, el recurso principal para el análisis es el conjunto de datos de viajes (VROSTM). Los logs de eventos utilizados son construidos sobre estos datos de viajes, específicamente sobre el conjunto de datos VROSTM_FT que es equivalente pero con datos de frecuencia y tarjeta y solo para mayo de 2022, que es el mes sobre el cual se realizó el estudio. Los logs utilizados son los siguientes:

- **[LOG_BASE]:** obtenido de agregar los atributos *cod_postal*, *municipio*, *barrio*, *dia_mes*, *dia_sem*, *hora*, *tipo_dia*, *tramo* y *sentido* a los registros de VROSTM_FT como se explica en la sección 4.3. A los registros de ascenso en paradas que no están en los corredores se les asigna tramo nulo.
- **[LOG_AGG_TRAZAS]:** obtenido de LOG_BASE, realizando agregación de eventos por recorrido de ómnibus (*numero_evento_recorrido*) y tramo mediante la utilización de un script, como se explica en la sección 4.4. Antes de agregar los eventos, el script identifica de manera particular los

ascensos fuera del corredor, estableciendo un identificador de tramo que indica desde qué tramo se produjo la salida del corredor, de modo de agregar también estos ascensos dependiendo desde donde se produjo la salida del corredor.

- **[LOG_MULTI8]:** obtenido de agregar los atributos *barrio* y *en_corredor* a los registros de VROSTM_FT, como se explica en la sección 4.3. Solo se consideran los ascensos que pertenecen a viajes multitramos (ver sección 5.2.3) con al menos uno de los ascensos registrado en el corredor 8 de Octubre. A las paradas del Intercambiador Belloni no se las considera como paradas dentro del corredor 8 de Octubre (*en_corredor* = Falso).

Destacar que los logs LOG_BASE y LOG_AGG_TRAZAS fueron filtrados considerando los datos necesarios para cada análisis. Por ejemplo, cuando se trabaja en el corredor de Av. Italia, solo se consideran las líneas que tienen alguna parada en el corredor de Av. Italia. De forma análoga se filtran los ascensos para el corredor 8 de Octubre.

Capítulo 5

Análisis de datos

En esta sección se explica cómo las técnicas de minería de procesos pueden ser útiles para analizar la movilidad urbana, se presentan los principales análisis realizados utilizando estas técnicas, una posible forma de mejorar su visualización a través de mapas, y un análisis más tradicional con los datos disponibles, sin la utilización de minería de procesos.

5.1. Introducción

Como se vio en la sección 2.2, la minería de procesos se utiliza para examinar procesos corporativos o procesos menos estructurados. Los procesos corporativos suelen tener una estructura clara y organización definida, conocidos como procesos "lasagna", mientras que los procesos menos estructurados, conocidos como procesos "spaghetti", carecen de una organización y estructura clara y no tienen una secuencia lineal definida. En el caso del transporte público, depende del proceso en cuestión, pero en general son procesos de tipo "spaghetti", lo que dificulta su análisis. A continuación se describe brevemente cómo la minería de procesos es útil en este ámbito, con algunos ejemplos.

Comenzando con el análisis de un recorrido específico de un ómnibus, partiendo del log LOG_BASE y utilizando *codigo_parada_origen* como actividad del proceso, se pueden descubrir modelos como el mostrado en la Figura 5.1. Del mismo, se observan varias cosas interesantes. Una de ellas es que se obtiene un modelo que refleja el comportamiento real del recorrido del ómnibus, donde cada actividad (nodo en la figura) es una parada del recorrido y cada flecha (transición) es el movimiento del ómnibus de una parada a otra. Esto es interesante y se debe a que la minería de procesos analiza el proceso (actividades relacionadas con un fin) y no los datos individuales. Además, se puede aumentar o reducir el nivel de detalle de las actividades y transiciones, lo que permite enfocar el análisis donde sea necesario.

Por ejemplo, la Figura 5.2 representa el mismo recorrido de la Figura 5.1, pero aumentando el nivel de detalle a las actividades y transiciones.

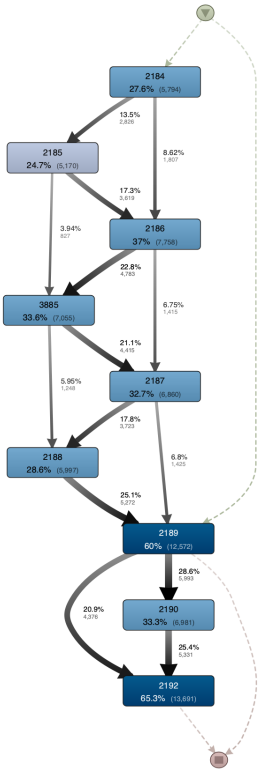


Figura 5.1: Modelo de recorrido de ómnibus con paradas como actividades, con detalle al 20 %.

Cuando se analizan los recorridos, se puede variar el nivel de detalle de los modelos generados hasta lograr alcanzar una comprensión del problema que se está abordando. Las herramientas utilizadas para crear estos modelos proporcionan una forma muy sencilla (y rápida) de obtener mayor o menor granularidad en el modelo. Esto es algo de suma importancia, ya que la fase de descubrimiento de modelos en la minería de procesos es una actividad visual, lo que facilita el análisis o la búsqueda de resultados específicos.

La capacidad de visualización que proporcionan los modelos se puede ver fácilmente reflejada en los modelos anteriores, donde cada transición que comienza y termina en la misma parada indica que hay más de una persona subiendo en esa parada en los recorridos. También se puede observar que hay paradas que tienen este tipo de transiciones y otras que no, e incluso entre las que sí tienen transiciones, hay diferencias en el grosor de las mismas, lo que indica que en esa parada sube más gente. La fortaleza de la minería de procesos en este contexto es que, por ejemplo, el dato de ascensos de pasajeros en una parada para un recorrido determinado se puede descubrir simplemente observando un modelo, sin tener que realizar consultas a bases de datos. Por otro lado, aunque este modelo representa un recorrido en particular, se puede modificar fácilmente para utilizar otro recorrido y obtener los mismos datos.

Otra información fácil de observar es el número total de ascensos en una parada para un recorrido en general. Esto se puede ver por la opacidad del color de la actividad (parada). La opacidad del color de la actividad es directamente proporcional al número de ascensos en esa parada. También es sencillo visualizar las paradas donde, en general, un recorrido tiene sus primeros y últimos ascensos, ya que corresponden a las primeras y últimas transiciones en el modelo, respectivamente, como se observa en las figuras 5.3 y 5.4.

Aunque la relevancia de esta información la determina finalmente un analista, lo que ofrece este método es una forma sencilla de obtenerla y visualizarla. Por ejemplo, si se realiza un análisis previo de recorridos que tienen un tiempo de duración prolongado (esto se puede conseguir parcialmente con minería de procesos) y se visualizan los modelos correspondientes, se puede utilizar esto como entrada para ver si hay recorridos a los que se les puede eliminar las primeras paradas, asumiendo que como son las primeras paradas, tampoco hay personas que las utilizan para bajarse.

La Figura 5.5 muestra un modelo de proceso similar al de la Figura 5.1,

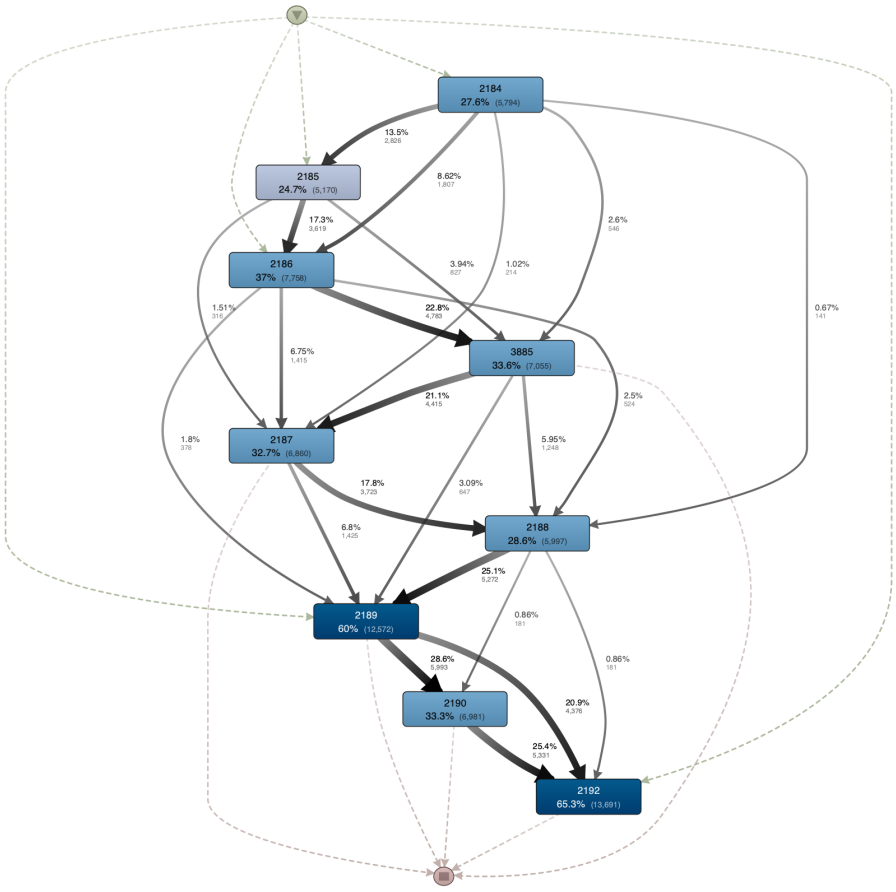


Figura 5.2: Modelo de recorrido de ómnibus con paradas como actividades, con detalle al 50 %.

pero con datos de performance. Al igual que la opacidad de las actividades, el grosor de una transición es directamente proporcional a la duración de la misma. Esto permite visualizar para un recorrido entre cuáles paradas hay más retraso. Además, se puede obtener el retraso medio entre cada parada, entre otros datos.

Hasta ahora se han mostrado ejemplos con paradas, pero esto puede generalizarse. En lugar de utilizar las paradas como actividades, podría utilizarse la zona (código postal, barrio, municipio, etc.) de Montevideo en la que se encuentra la parada. Como cada parada se encuentra en una zona, se podría conocer los transbordos (el usuario subió en una parada X, y luego en la parada Y) más frecuentes con respecto a zonas de Montevideo. Por ejemplo, la Tabla 5.1 refleja los transbordos más frecuentes entre códigos postales. En este caso, la mayor cantidad de transbordos se dan dentro (es tanto origen como destino) del código postal 12800. La simple generalización de un atributo como actividad del

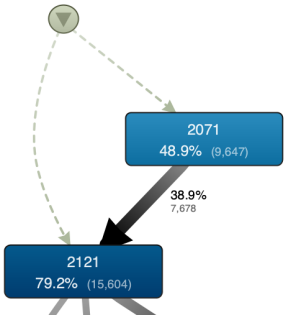


Figura 5.3: Primeras transiciones del recorrido.

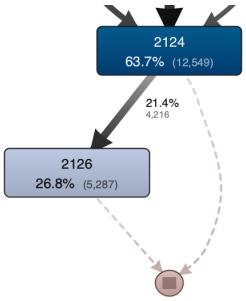


Figura 5.4: Últimas transiciones del recorrido.

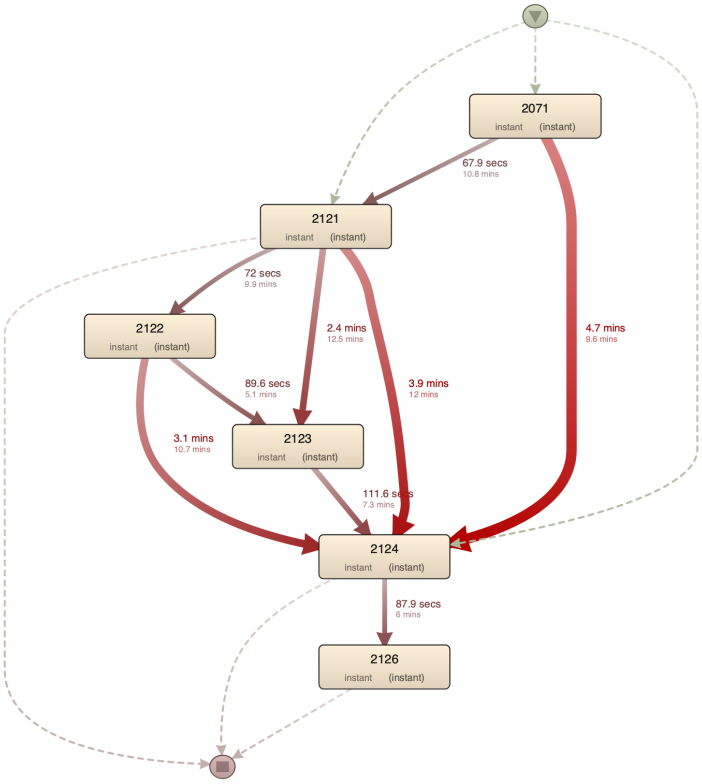


Figura 5.5: Imagen de modelo de performance.

proceso permite tener más o menos detalle en los análisis.

Otra forma de generalización es considerar que las paradas se encuentran en tramos definidos a conveniencia, como se muestra en la Figura 5.6. Aquí, se

Tabla 5.1: Transbordos más frecuentes entre códigos postales de Montevideo.

CP origen	CP destino	Frecuencia	Proporción
12800	12800	52.198	3,92 %
11900	11900	43.475	3,40 %
13000	12000	34.070	2,76 %
11600	11600	29.350	2,45 %

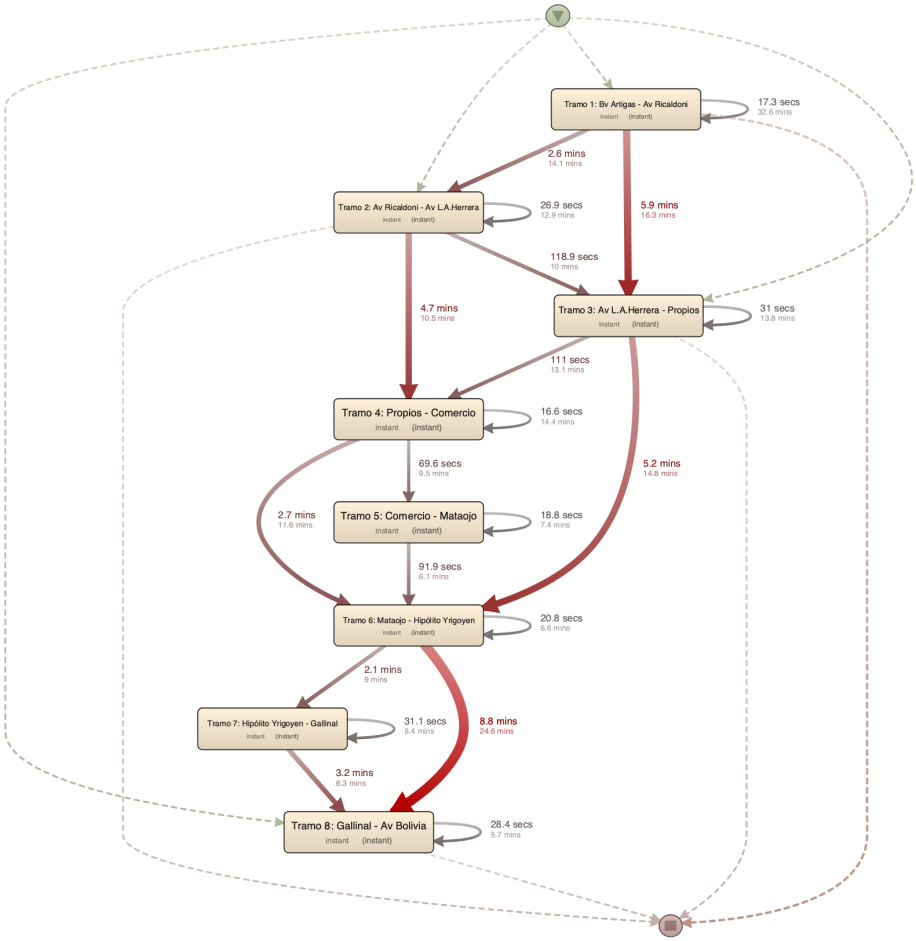


Figura 5.6: Modelo de performance con tramos como actividades.

toma como actividad del proceso los tramos de las paradas, donde cada parada se encuentra en alguno de los tramos definidos. Desde el punto de vista del tiempo, se puede observar el tiempo promedio que ocupa un recorrido entre un tramo y otro. Algo que se considera relevante, pero que no se puede observar en el modelo, es conocer el tiempo que un ómnibus permanece en un tramo

determinado. En este caso, esto no es posible obtenerlo, ya que el log utilizado no tiene la fecha de cuándo un ómnibus comienza y termina un tramo, sino que solo tiene la fecha del primer ascenso en el tramo (primera parada que se ubica en ese tramo) y la fecha del último ascenso del tramo (última parada que se ubica en el tramo). A partir de estos datos, se puede manipular el log para tener en cuenta esta información. De cierta manera, lo que se realiza es un agregado de trazas por tramo. Un ejemplo de esto es el de la Figura 5.7. Como se puede apreciar, se tiene el dato de cuánto tiempo demora un ómnibus en cada tramo directamente en el modelo.

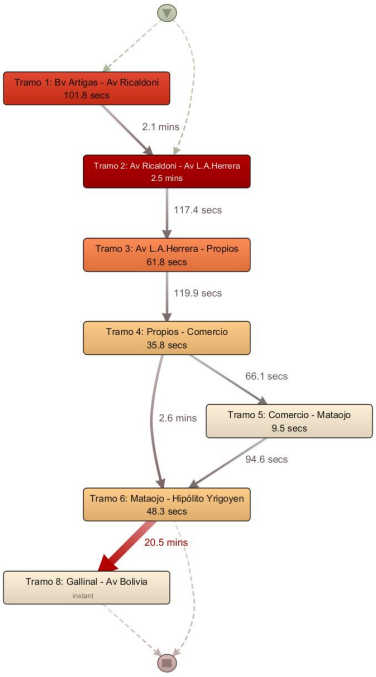


Figura 5.7: Modelo de performance con tramos como actividades, luego de realizar una agregación de trazas.

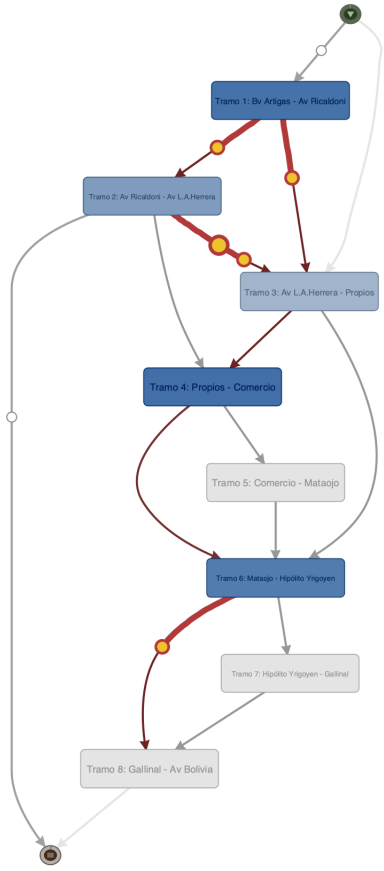


Figura 5.8: Comportamiento dinámico del proceso.

La minería de procesos, y en particular las herramientas que se dedican a esta disciplina, permiten reproducir las instancias del proceso a partir de los datos del log mediante una animación. Utilizando el ejemplo de la Figura 5.7, se puede ver cómo se comportaron los ómnibus en su recorrido a lo largo de diferentes tramos. Esto permite visualizar un comportamiento dinámico en lugar de simplemente

un comportamiento estático, lo que permite observar embotellamientos o la utilización de diferentes tramos a distintas horas del día, como se puede ver en la Figura 5.8.

Nuevamente, es importante señalar que las herramientas de minería de procesos proporcionan esta visualización directamente, sin necesidad de realizar procesamientos adicionales o utilizar otras herramientas de visualización. Además, cada análisis puede ser examinado desde distintas perspectivas y los datos utilizados pueden ser filtrados, lo que aumenta su valor, ya que reduce el tiempo necesario para ejecutar esa tarea.

En resumen, por todo lo visto en esta sección, la minería de procesos podría ser capaz de responder las siguientes preguntas:

1. **¿Cuál es el recorrido que realiza un bus?** Como se menciona y ejemplifica en la Figura 5.1, es posible generar una visualización de un modelo de referencia del recorrido de un bus, modelando las paradas como actividades.

2. **¿Existen demoras en los corredores?** Siguiendo el ejemplo de la figuras 5.6 y 5.7, y filtrando las líneas de buses que recorren los corredores, se pueden calcular tiempos máximos y promedios de transición entre paradas y tramos para encontrar posibles demoras.

3. **¿Cuáles son las paradas y zonas más concurridas?** Es posible utilizar los modelos generados mediante minería de procesos (como los mencionados en esta sección) para visualizar rápidamente según la opacidad de las actividades, paradas y tramos más concurridos (con más ascensos).

4. **¿Cuáles son los horarios pico?** Analizando los datos como proceso, y segmentando los mismos por hora, pueden visualizarse los horarios con más ascensos, por lo que se puede deducir los horarios donde existe mayor demanda de viajes.

5. **¿Entre qué zonas se realizan más transbordos?** Mediante el uso del enriquecimiento de logs agregando información del barrio, como se describe en la sección 4.3, es posible conocer los barrios de origen y destino de los viajes que realizan transbordos (viajes con boletos de 1 hora o 2 horas).

6. **¿Desde/hacia qué zonas se ingresa/egresa de los corredores?** De forma similar a la pregunta 5, se puede conocer las zonas donde se registra un ascenso previo al ingreso del corredor, y las zonas donde se registra un ascenso posterior al egreso del corredor.

5.2. Análisis de corredores

Uno de los puntos de interés del proyecto, fue realizar análisis de movilidad sobre los corredores definidos en el Plan de Movilidad (PM) de Montevideo, los cuales se observan en la Figura 5.9. Según el PM, los corredores son vías

de circulación de tránsito que alojan una ruta troncal, servida por buses de gran capacidad, funcionando en un carril exclusivo o con circulación preferencial, con paradas predefinidas y terminales o intercambiadores de transferencia. En particular, en el alcance del proyecto se analizaron dos de estos corredores, el corredor de Av. Italia, y el corredor Av. 8 de Octubre, dichos análisis pueden verse en profundidad en los anexos "Análisis Av. Italia" y "Análisis 8 de Octubre", respectivamente.

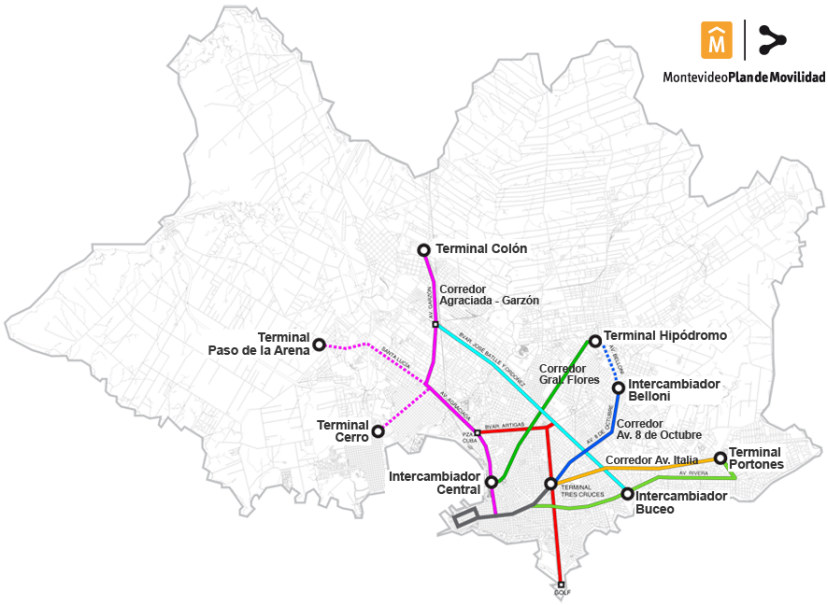


Figura 5.9: Corredores del Plan de Movilidad de Montevideo.

Para estudiar el comportamiento y flujo del transporte público dentro del corredor, se decidió dividir al mismo en tramos, agregando esta información al log de eventos inicial mediante el proceso de enriquecimiento de logs, descrito en la sección 4.3. Los modelos presentados son descubiertos utilizando el log LOG_BASE, con sus respectivos filtros por corredor explicados en la sección 4.5, con el campo *numero_evento_recorrido* como ID de Caso, el campo *tramo* como actividad, y el campo *fecha_evento* como marca de tiempo. Dado que no se cuenta con datos de ubicación de los ómnibus en el tiempo, sino de transacciones (boleto de pasajero) en paradas, es importante que cada tramo tenga al menos una parada en cada sentido, de lo contrario, nunca se vería el pasaje por dicho tramo en ese sentido.

La división en tramos tiene ventajas respecto al análisis por paradas. Uno de los problemas de analizar el flujo por paradas es que son muchas, lo cual dificulta el análisis. Además, no todas las variantes pasan por las mismas paradas, lo que hace difícil la comparación. Al dividir en tramos es más probable que las distintas variantes pasen (tengan paradas) por los mismos. El análisis por paradas puede

ser preferible cuando se desea hacer un análisis más granular. Se decidió entonces dividir los corredores entre cruces con calles importantes, de modo de tener en cuenta lo mejor posible las consideraciones previas, intentando equiparar de la mejor manera posible la cantidad de paradas por tramo y sentido. Como se verá, la mayoría de las líneas ingresan o salen del corredor en una de las calles definidas como límites de tramo, lo que ayuda en la comparación.

Los análisis también dividen los corredores en 2 partes, que son los sentidos de circulación (hacia el este y oeste en Av. Italia, hacia el suroeste y el noreste en Av. 8 Octubre), agregados también con enriquecimiento de logs.

Buscando responder las preguntas 1 a 4 planteadas en la sección 5.1, se analizan datos en general como cantidad de ascensos por tramo, empresa, línea y variante, también por tipos de días (hábiles, sábados, domingos) así como por diferentes rangos horarios, identificando las horas pico de cada tipo de día. También se hacen observaciones sobre la performance de los procesos de viaje, así como un análisis de casos subdividido por conjuntos de tramos, analizando el proceso con las paradas como actividades (utilizando el campo *codigo_parada_origen* como actividad), en ambos sentidos.

Los ascensos registrados en los corredores de Av. Italia y Av. 8 de Octubre representan aproximadamente un 10 % del total de 25 millones mencionado en la sección 4.1, correspondiendo un 8 % a 8 de Octubre, y un 2 % a Av. Italia. La Figura 5.10 muestra la distribución de estos ascensos según corredor y sentido.

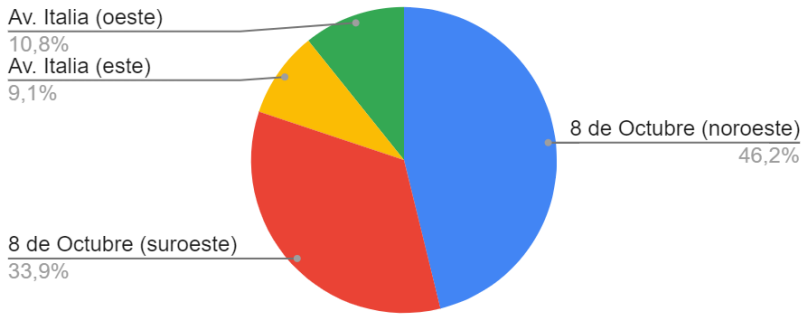


Figura 5.10: Distribución de ascensos por corredor y sentido para 8 de Octubre y Av. Italia.

5.2.1. Av. Italia

El corredor abarca Av. Italia, desde Bv. Artigas hasta Av. Bolivia, teniendo un largo de 7,8 km y un total de 49 paradas, 26 hacia el este y 23 hacia el oeste. En la Figura 5.11 se puede ver el corredor entero dividido en tramos. Los círculos verdes son las paradas del corredor. Para Av. Italia se definieron 8 tramos delimitados por las siguientes 9 calles: Bv. Artigas, Av. Ricaldoni, L.A. de Herrera, Propios, Comercio, Mataojo, Hipólito Yrigoyen, Av. Gallinal, Av. Bolivia. Se numeran los tramos del 1 al 8 comenzando desde el extremo oeste.

El tramo 1 en sentido oeste tiene una sola parada, aun así el 85 % de las variantes que recorren el corredor hacia el oeste paran en esa parada. El resto de tramos tiene entre 2 y 5 paradas en cada sentido.

Figura 5.11: Corredor Av. Italia dividido en 8 tramos.

La división en 2 sentidos dio como resultado el sentido este (desde el Bv. Artigas hacia la Av. Bolivia) y el sentido oeste (desde Av. Bolivia hacia Bv. Artigas). Tomando como ejemplo el sentido este, puede notarse que la mayor concentración de ascensos se da en los tramos 1 y 2, como muestran la Figura 5.13 y la Tabla 5.2.

Linea \ Tramo	1	2	3	4	5	6	7	8
21								
D10								
407								
64								
D9								
370								
151								
174								
300								
71								
115								
143								
494								
495								
187								
405								
112								

Figura 5.12: Tramos recorridos por cada línea en Av. Italia.

En los datos de mayo, se observa que por el corredor se desplazan 17 líneas de ómnibus con un total de 114 variantes. Para tener un mejor entendimiento del corredor se clasifican las líneas que lo atraviesan según los tramos por los que se desplazan.

En la Figura 5.12 se muestran los tramos recorridos por cada línea. El color celeste indica que la línea recorre el tramo en su completitud y el color amarillo indica un recorrido parcial. Notar que la mayoría de las líneas recorren tramos enteros, esto se debe a la división del corredor en calles particularmente importantes. Se observa a simple vista cómo la mayoría de las líneas recorren un intervalo continuo de tramos entre el tramo 1 y 8.

Las líneas 21 y D10 recorren los tramos 1 al 8, es decir, el corredor entero.

Las líneas 407, 64 y D9 recorren los tramos 1 al 6. Las líneas 370 y 151 recorren los tramos 1 al 4, además la línea 370 recorre también parte del tramo 8 (desde Alberto Zum Felde hasta Av. Bolivia). Las líneas 71, 174 y 300 recorren los tramos 1 y 2. Las líneas 495, 115, 143 y 494 recorren el tramo 1 entero y parte del tramo 2 (hasta Av. Garibaldi). La línea 187 recorre solamente el tramo 1 y la 112 solo el tramo 8. La línea 405 recorre los tramos 3 y 4.

Tabla 5.2: Cantidad de ascensos por tramo, sentido este.

Tramo	Frecuencia	Proporción
Tramo 1: Bv Artigas - Av Ricaldoni	97.543	43,10 %
Tramo 2: Av Ricaldoni - Av L.A. Herrera	71.292	31,50 %
Tramo 3: Av L.A. Herrera - Propios	25.174	11,12 %
Tramo 4: Propios - Comercio	13.509	5,97 %
Tramo 6: Mataojo - Hipólito Yrigoyen	12.696	5,61 %
Tramo 8: Gallinal - Av Bolivia	2.702	1,19 %
Tramo 5: Comercio - Mataojo	2.483	1,10 %
Tramo 7: Hipólito Yrigoyen - Gallinal	893	0,39 %

Con respecto a la pregunta 1, la Figura 5.14 presenta la línea 21 con sus paradas para la variante 7198, la cual recorre el corredor en sentido oeste. En la Figura 5.14a se puede ver el recorrido publicado en el sitio del STM[1], con parada de origen 2169 en Av. Bolivia, y parada de destino 6274 en Plaza Independencia (fuera del corredor). En la Figura 5.14b puede verse el modelo teórico representado con minería de procesos en Disco, mostrando las primeras 8 paradas sobre el corredor de Av. Italia, desde la 2169 que es el origen del recorrido. Se observa que es secuencial debido a que tiene un registro en cada parada. En la Figura 5.14c se puede ver el modelo generado teniendo en cuenta el registro de viajes realizados durante mayo, donde puede verse la menor opacidad de las primeras 3 paradas, lo que indica una menor cantidad de ascensos en las mismas.

Para analizar las salidas y entradas al corredor se utilizó el log LOG_AGG_TRAZAS, detallado en la sección 4.5.

En la Figura 5.15 se puede ver el porcentaje de cubrimiento de casos de cada actividad y transición. Para que el modelo fuese más claro se nombran los tramos solo por su número. Las actividades con una U y luego un número de tramo (e.g "U8"), corresponden a las actividades de salida de tramo antes mencionadas. Notar cómo luego de salir de un tramo hacia afuera del corredor, ningún ómnibus (o línea) vuelve al corredor, salvo una mínima cantidad que vuelve al tramo 8 que por la clasificación realizada anteriormente se sabe que corresponden a las líneas 21 y 370.

Recordar que los datos con que se trabaja son ascensos a ómnibus y no datos de GPS, la salida de tramo puede ser producto de que efectivamente un ómnibus se desvió en ese tramo, pero también puede ocurrir que no haya parado en ningún tramo posterior y su siguiente parada haya sido fuera del

[1]STM Líneas y horarios. https://www.montevideo.gub.uy/app/stm/horarios/

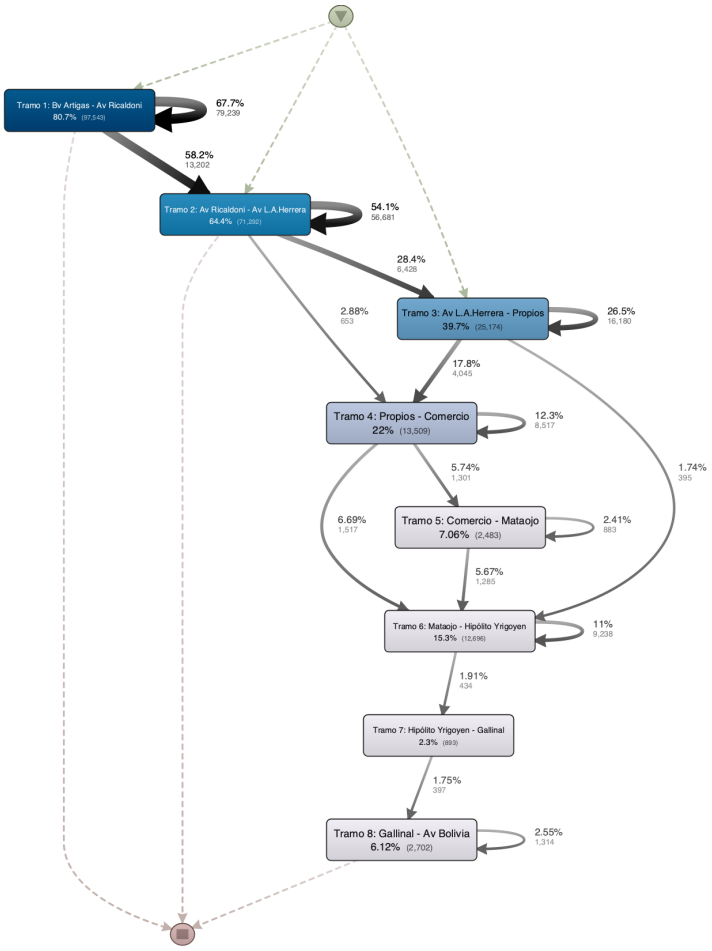


Figura 5.13: Modelo de proceso de Av. Italia con tramos como actividades, en sentido este. Con nivel de detalle de actividades al 100 % y de transiciones al 30 %.

corredor. Por lo tanto, algunas de las transiciones observadas no corresponden a los recorridos de ninguna línea en particular, lo cual es evidente, por ejemplo, para la transición entre el tramo 4 y el 6. La transición desde U1 hacia 8 por ejemplo, puede corresponder a la línea 21 o 370 para los casos en los que se registró al menos un ascenso en el tramo 1, luego alguno fuera del corredor y luego al menos uno en el tramo 8. Observando la clasificación de líneas de la Figura 5.12, es fácil ver que este puede perfectamente ser el caso para un viaje hecho por la línea 370, pero no resulta tan directo notar para la línea 21, que recorre todo el corredor. Esto es porque la línea 21 tiene variantes que abandonan el corredor al final del tramo 6 (Hipólito Yrigoyen) y vuelven a entrar en el tramo 8, cerca del comienzo (Alberto Zum Felde), dato que la clasificación de la figura 5.12 no muestra al estar enfocada en las líneas en general y no en

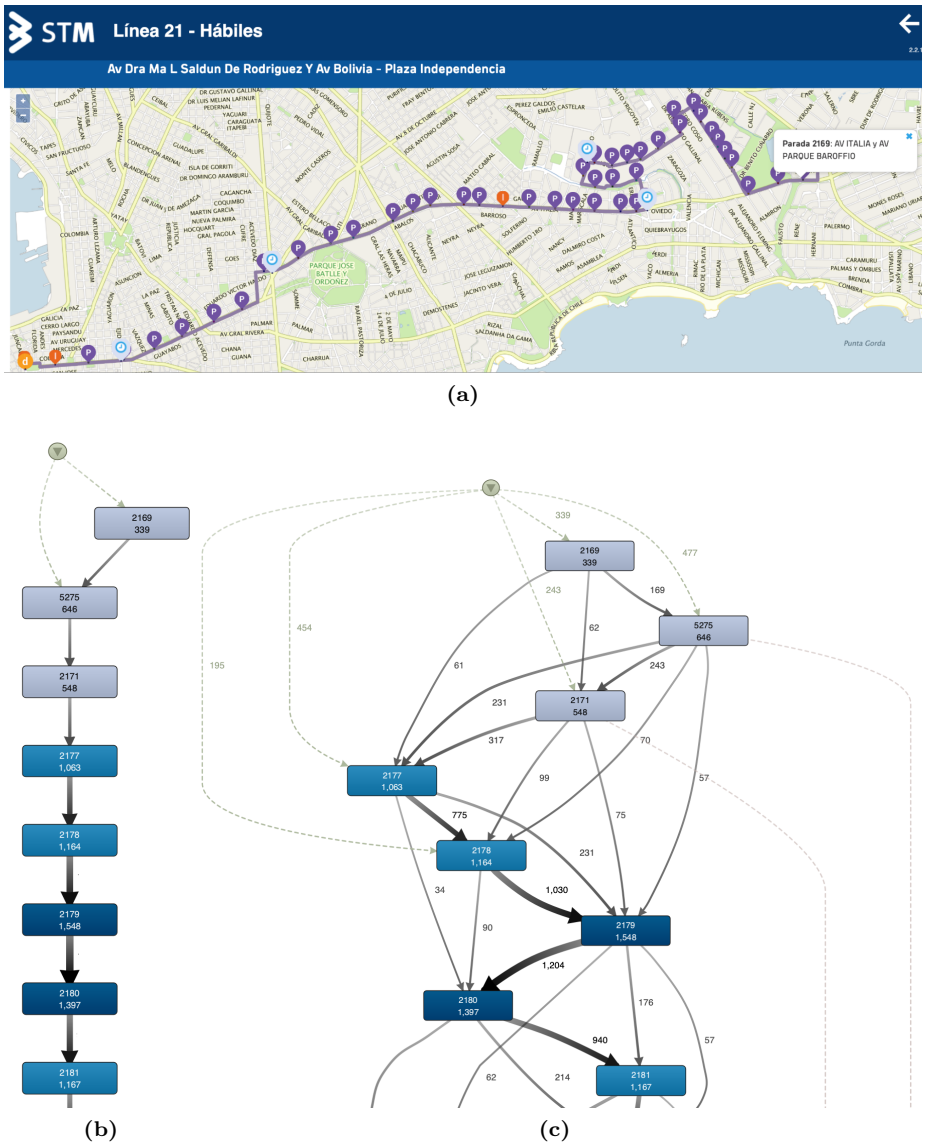


Figura 5.14: Recorrido de la línea 21 con las paradas para la variante 7198: (a) horarios publicados en STM, (b) fragmento del modelo de referencia y (c) fragmento de viajes realizados con STM.

cada variante.

Desde el tramo 1 la siguiente parada más frecuente es en el tramo 2, esta transición está presente en el 58 % de los casos, pero también hay bastantes casos en los que la siguiente parada es fuera del corredor (18 %). De estos últimos, pocos vuelven al corredor en el tramo 8 y como ya se mencionó, esto ocurre en

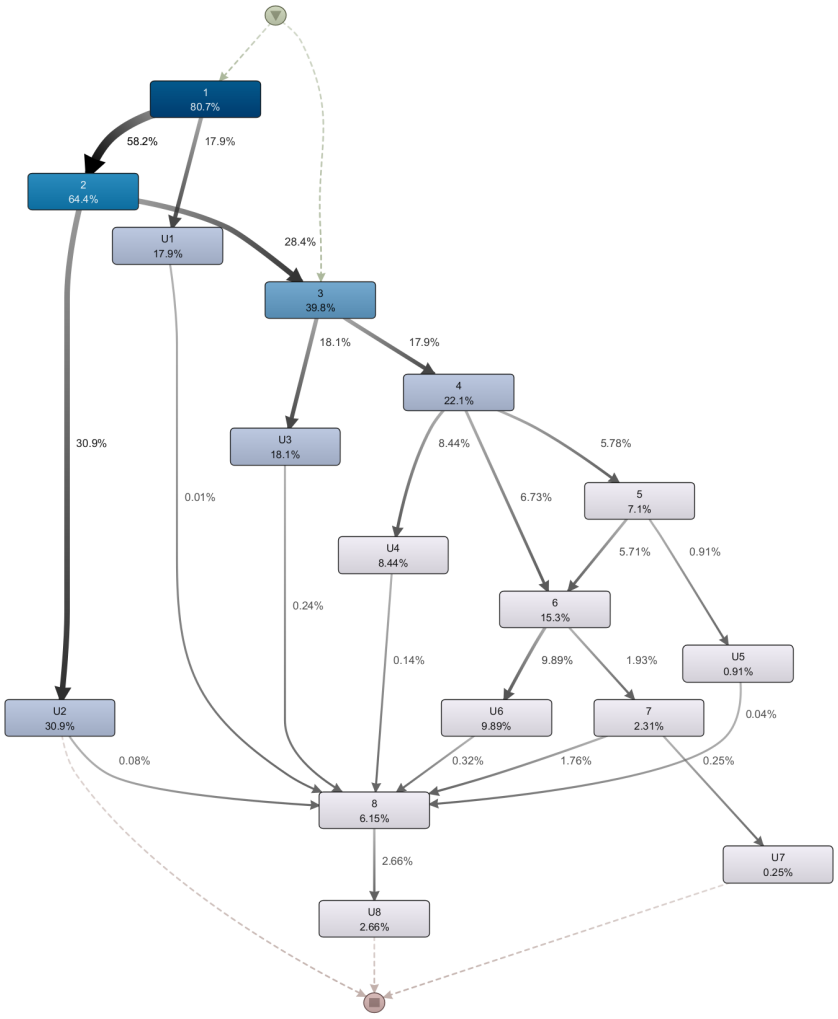


Figura 5.15: Modelo de proceso para el flujo de ómnibus con dirección este sobre tramos de Av. Italia. Se muestra el porcentaje de cobertura de casos.

general para todos los casos en que el ómnibus abandona el corredor. Luego del tramo 2 hay una cantidad mayor de ómnibus que tienen su siguiente parada fuera del corredor a los que la tienen en el tramo 3, esto tiene sentido dada la cantidad de líneas que solo recorren los tramos 1 y 2. Lo mismo sucede en el tramo 3, con una división casi igual entre los casos en que se abandona el corredor y aquellos en los que se vuelve a parar en el tramo 4. Desde el tramo 4 se pueden ver 3 transiciones, las 2 usuales (hacia el tramo siguiente o abandono

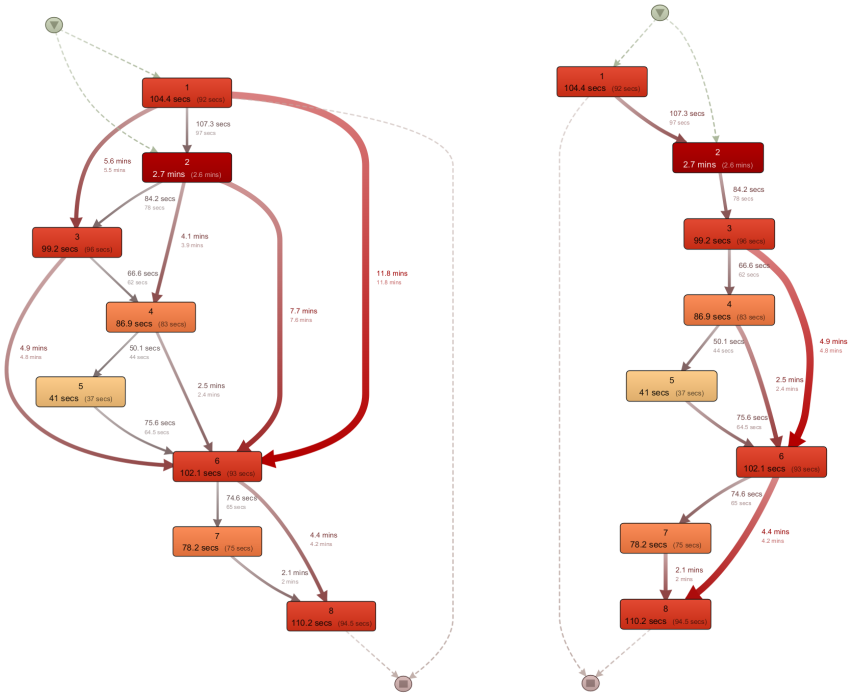
del corredor) más una transición hacia el tramo 6. Nuevamente, la transición más frecuente es la de abandono del corredor, en este caso con más proporción que en los casos del tramo 2 y 3. Observar cómo la transición hacia el tramo 6 es más frecuente que la transición hacia el tramo 5, indicando que muchas veces los ómnibus no paran en el tramo 5, lo cual era de esperarse viendo la cantidad de ascensos en el tramo 5. En el tramo 5 la gran mayoría de las transiciones son hacia el tramo 6, muy pocos casos en que el siguiente ascenso es fuera del corredor en comparación. Desde el tramo 6, en la mayoría de los casos se registra el siguiente ascenso fuera del corredor, son pocas (pero no despreciables) las transiciones hacia el tramo 7, recordar que el tramo 7 es el que menos ascensos registra (0,9 % del total). Desde el tramo 7 la mayoría de transiciones son hacia el tramo 8, siendo igualmente una transición poco frecuente por el mismo motivo anterior, la poca cantidad de ascensos en el tramo 7 en general. Por último se observa que una pequeña cantidad de ómnibus tiene ascensos fuera del corredor luego del tramo 8, estos corresponden a las líneas 21 y D10 que son las únicas que pasan por el tramo 8 y siguen por Av. Italia pasando el fin del tramo (Av. Bolivia) y hasta Av. a la Playa (6,5 km aprox. pasando Av. Bolivia).

Con respecto a la pregunta 2, en la Figura 5.16 se presenta un modelo generado sobre el log LOG_AGG_TRAZAS. El tiempo que demora una actividad es el tiempo desde la primera parada en el tramo correspondiente a la actividad hasta la última parada en dicho tramo. Por lo tanto, si un ómnibus para una sola vez en un tramo, se obtendrá una duración nula. Por este motivo, para generar este modelo se filtraron los datos, removiendo los eventos en los cuales un ómnibus tiene una sola parada en un tramo. Esto evita actividades con duración nula permitiendo obtener tiempos más cercanos a los reales.

La duración de una transición corresponde al tiempo entre la última parada en un tramo y la primera parada fuera de este tramo, por lo cual no se tiene el problema anterior. Observar que aunque se eliminen las actividades de duración nula, siguen existiendo casos que no son los óptimos, pero aplicar un filtro para quedarse con los casos óptimos implica perder demasiados datos y quedarse con casos más bien particulares. Notar cómo en estos casos, la duración no se pierde sino que se traslada de actividad a transición. Por ejemplo, si un ómnibus para en las primeras dos paradas de un tramo, se saltea el resto y luego para en la tercera parada del siguiente, la duración de las actividades será pequeña y el tiempo extra estará en la transición entre ellas. Por este motivo es conveniente observar el tiempo de las actividades junto con el de las transiciones.

El filtro aplicado cambia la frecuencia de aparición de ciertas transiciones y actividades, por lo que se pueden ver algunas diferencias respecto a otros modelos. Otro inconveniente del filtro son los tramos con una sola parada en algún sentido, en el caso de Av. Italia se tiene el caso del tramo 1 oeste que se pierde. Para este caso el mejor dato que se tiene es el de transición entre el tramo 2 y el tramo 1 de los datos sin filtrar, la duración de la actividad correspondiente al tramo 1 hacia el oeste siempre será nula por tener una sola parada.

Volviendo a la Figura 5.16 puede verse que tanto el tramo 1 como el 2 son los que más tiempo demoran. Si bien el tramo 2 resalta particularmente, hay que tener en cuenta que el tramo 1 es de 640 metros, el tramo 2 de 1,1



(a) **(b)**

Figura 5.16: Modelo de proceso para el flujo de ómnibus con dirección este sobre tramos de Av. Italia con características de performance. Se muestra la media y mediana de las duraciones. Se muestra la media y mediana de las duraciones (a) completo (b) transiciones más frecuentes.

km aproximadamente, y ambos tienen 3 paradas hacia el este, es decir que la velocidad media del ómnibus es bastante similar en ambos casos. También resalta la duración de la transición del tramo 1 al 2. Es el valor más grande luego de la transición entre el tramo 7 y el 8. Los tramos 7 y 8 no solo son más grandes, sino que además reciben muchos menos ascensos, con lo cual es posible que varios ómnibus salten paradas en su recorrido, lo cual aumenta el tiempo de la transición. Notar que el tiempo entre los tramos 4, 5 y 6 es bastante más rápido que el resto. También observar que el tiempo tanto al parar en el tramo 5 como no hacerlo (parar en 4 y luego en 6) es casi el mismo.

Se observa cómo en las transiciones donde se saltean tramos, en general la duración si bien es menor, también está muy cercana a la duración del camino por tramos. En el caso del salto del tramo 1 al 6 por ejemplo, si se suma la duración del camino entero, se obtienen 12,8 minutos, solo 1 minuto más que la transición directa.

Al analizar los tiempos de transiciones correspondientes a saltos de tramo hay que tener presente que en algunos casos esto puede deberse a que a un ómnibus no lo pararon durante uno o más tramos, pero también se puede deber a que una línea (o una variante de una línea) se desvía y vuelve a ingresar al corredor. Al no separarse por línea, estos tiempos se promedian juntos. Este es el caso de, por ejemplo, algunas variantes de la línea 21 que salen del corredor en el tramo 6 y vuelven a entrar en el 8, por lo que demoran bastante más que si siguieran por el corredor, aportando más tiempo a la transición entre los tramos 6 y 8.

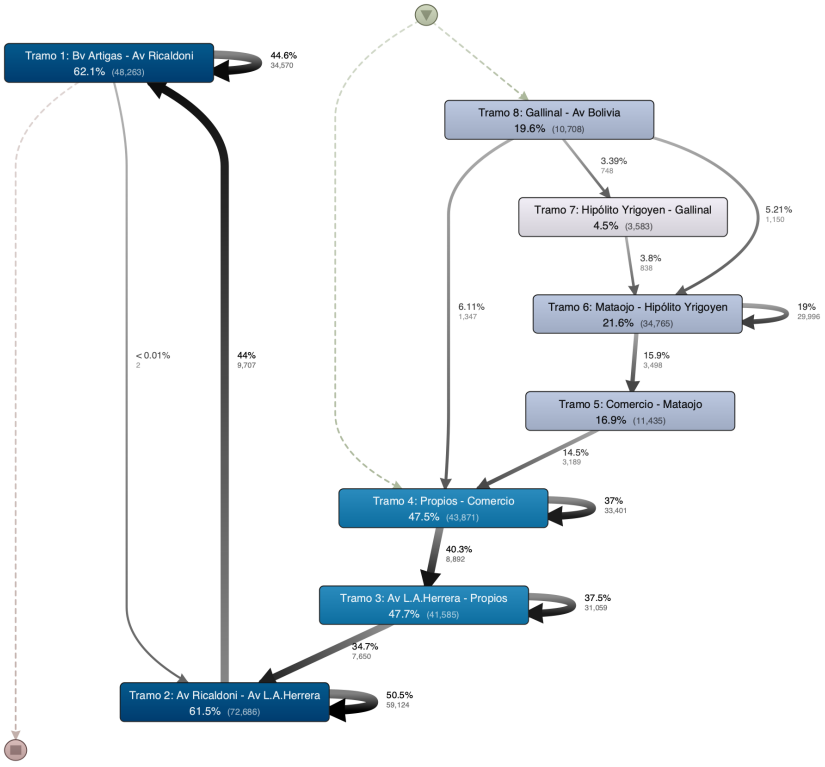


Figura 5.17: Modelo de proceso del corredor Av. Italia con tramos como actividades, en sentido oeste. Con nivel de detalle de actividades al 100 % y de transiciones al 10 %.

En relación a la pregunta 3, como se mencionó, en la Figura 5.13 y Tabla 5.2 se observa que hacia el este la mayor concentración de ascensos se da en los tramos 1 y 2. Por otro lado, en la Figura 5.17 puede verse que en el sentido oeste nuevamente los tramos 1 y 2 son los más concurridos, a pesar de estar al "final" de los viajes en este sentido. Respecto a las paradas, las que mayor cantidad de ascensos registran se ubican frente a dos grandes hospitales, el Hospital de Clínicas y la Médica Uruguaya, como se observa en la Figura 5.18. Estas paradas concentran el 56 % de ascensos en el sentido este, y el 40 % en el sentido oeste.

En la Figura 5.17 también se puede ver que los viajes en el sentido oeste están más distribuidos en todos los tramos, con una marcada diferencia entre los tramos 1 al 4, respecto a los tramos 5 a 8. También hay una diferencia importante en la proporción de ascensos en el tramo 1 entre este y oeste que se puede deber a varios factores.

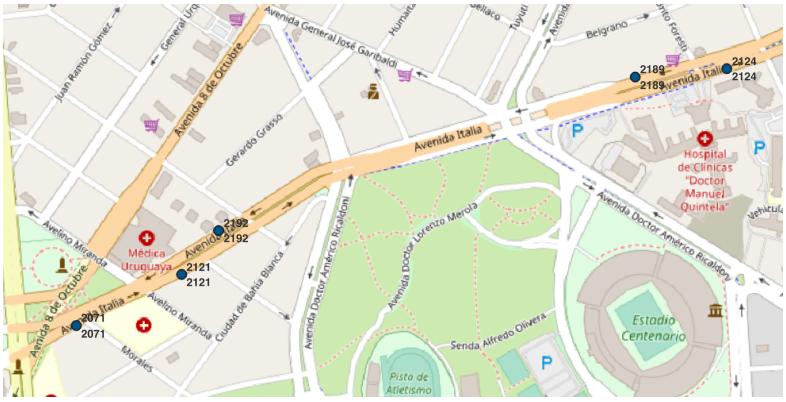


Figura 5.18: Paradas con más ascensos en el corredor Av. Italia, ubicadas frente al Hospital de Clínicas y a la Médica Uruguaya.

Un posible factor es que el tramo 1 solo tiene una parada hacia el oeste, pero de todos modos por esa parada pasan el 85 % de las variantes, como ya se mencionó. También puede ser que por algún motivo mucha gente que se desplaza hacia el oeste prefiera la parada anterior (tramo 2) o la siguiente, es decir, alguna de las primeras paradas hacia el oeste del fin del corredor (esquina de Tres Cruces o 18 de Julio). Otra posibilidad es que a mucha gente que viaja hacia el oeste le quede mejor tomar el ómnibus en 8 de Octubre, que a esa altura de Av. Italia está muy cerca.

Los tramos 5, 7 y 8 son los que tienen menos cantidad de ascensos en ambos sentidos, particularmente para el este (2,6 % entre los 3). Estos 3 tramos reciben aproximadamente 4 veces más ascensos hacia el oeste que hacia el este. También los tramos 4 y 6 reciben bastante más ascensos hacia el oeste (3 veces más aproximadamente).

En cuanto a los ascensos por tipo de día (hábil, sábado, domingo/feriado), en la Figura 5.19 se observa que, como era de esperar, la mayor cantidad de ascensos se dan en días hábiles, alcanzando un 86,97 % en el sentido este, y 88,66 % en el oeste. La distribución de ascensos es seguida en ambos sentidos por los días sábado, con una proporción en el entorno del 7,85 %, y los domingos con alrededor de 4,3 %. Se encuentra la particularidad también que en ambos sentidos el día con mayor ascensos es el martes, con alrededor del 19,4 %, y el menor los miércoles, con cerca de 16,25 %.

Respecto a la pregunta 4, la Figura 5.20 muestra que, teniendo en cuenta la cantidad de ascensos en el tiempo, se tiene que las horas con más ascensos ocurren entre las 15 y las 18 horas, registrando la mayor cantidad de eventos

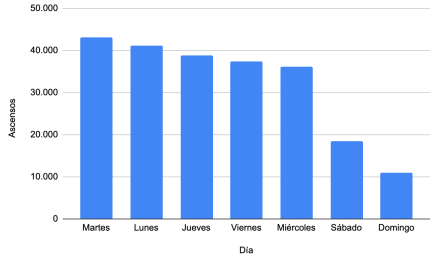


Figura 5.19: Cantidad de ascensos por día de semana en corredor Av. Italia, sentido este.

Figura 5.20: Ascensos por hora en el corredor Av. Italia en días hábiles, sentido este.

alrededor de las 17 horas, con un promedio de 15 ascensos por minuto a lo largo de todo el corredor (Figura 5.21).

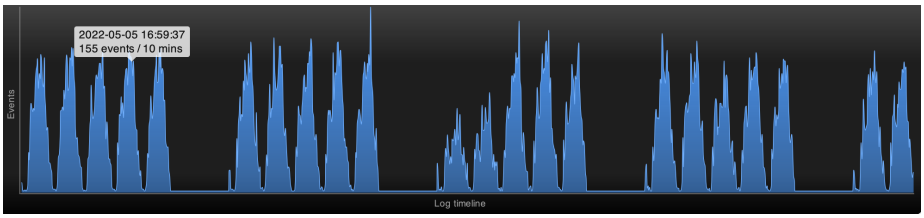


Figura 5.21: Cantidad de ascensos ocurridos en todos los tramos de Av. Italia en días hábiles, sentido este.

5.2.2. Av. 8 de Octubre

El corredor abarca la Av. 8 de Octubre en toda su extensión, desde Av. 18 de Julio hasta Av. José Belloni, teniendo un largo de 6,0 km y un total de 57 paradas, 29 hacia el noreste y 28 hacia el suroeste. En la Figura 5.22 se puede ver el corredor entero dividido en tramos. Los círculos verdes son las paradas del corredor. Para Av. 8 de Octubre se definieron 7 tramos delimitados por las siguientes 7 calles: 18 de Julio, L.A. de Herrera, Comercio, Larravide, Pan de Azúcar, Cno. Corrales y Belloni, adicional a estos cruces con calles, se consideró la manzana del Intercambiador Belloni como un tramo más, dada su importancia en ascensos y transbordos, como más adelante se verá. Se numeran los tramos del 1 al 7 comenzando desde el extremo suroeste. En el tramo 1 se observa un trayecto cercano a 18 de Julio sin paradas. La parada más cercana a 18 de Julio en el corredor es en la esquina de Presidente Berro, a 2 cuadras del túnel. También se observa un trayecto considerable sin paradas cerca del intercambiador.

La división en 2 sentidos dio como resultado el sentido suroeste (desde el intercambiador Belloni hacia la Av. 18 de Julio) y el sentido noreste (desde

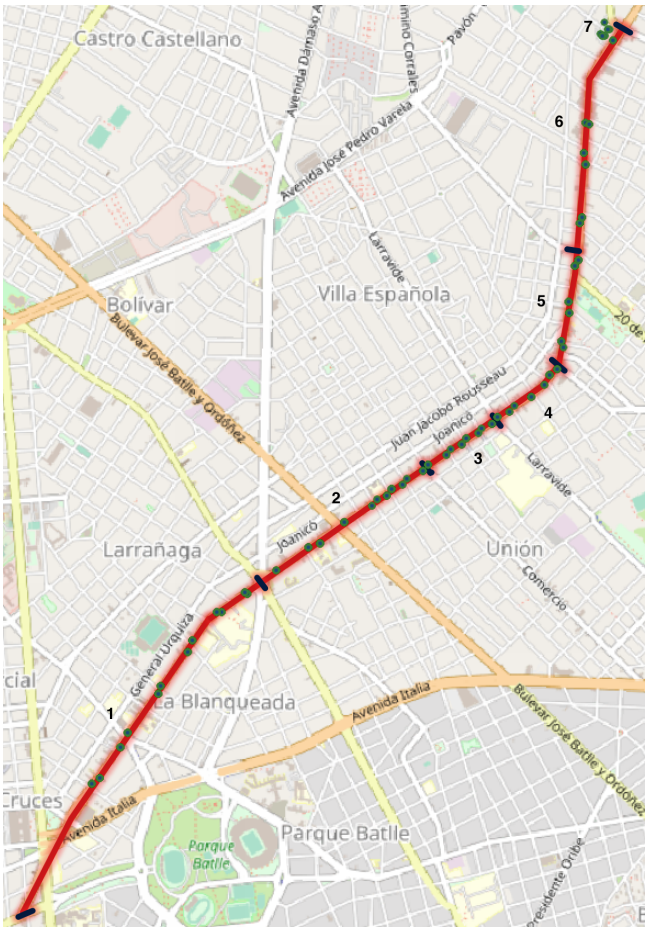


Figura 5.22: Corredor 8 de Octubre dividido en 7 tramos.

la Av. 18 de Julio hacia el intercambiador Belloni). Tomando como ejemplo el sentido suroeste, puede notarse que la mayor concentración de ascensos se da en los tramos 1 y 2, como muestran las figuras 5.23 y 5.24, y la Tabla 5.3.

En los datos de mayo, se observa que por el corredor se desplazan 27 líneas de ómnibus con un total de 299 variantes. Para tener un mejor entendimiento del corredor se clasifican las líneas que lo atraviesan según los tramos por los que se desplazan.

En la Figura 5.25 se muestran los tramos recorridos por cada línea. El color celeste indica que la línea recorre el tramo en su completitud y el color amarillo indica un recorrido parcial. Para la clasificación de la figura se toma en cuenta el recorrido de todas las variantes de cada línea en cada sentido. En la mayoría de los casos las líneas hacen el mismo recorrido de calles a la ida y a la vuelta, existiendo casos particulares como por ejemplo la línea 405, cuya variante hacia

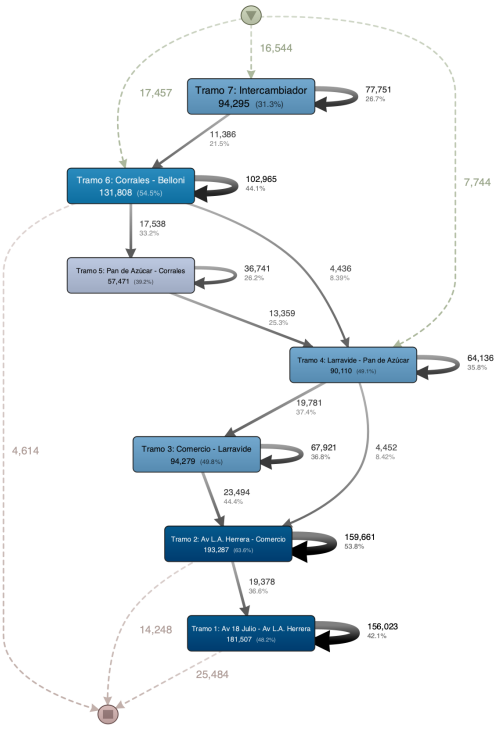


Figura 5.23: Modelo de proceso de Av. 8 de Octubre con tramos como actividades, en sentido suroeste. Con nivel de detalle de actividades al 100 % y de transiciones al 40 %.

Tabla 5.3: Cantidad de ascensos por tramo en 8 de Octubre, sentido suroeste.

Tramo	Frecuencia
Tramo 2: Av L.A. Herrera - Comercio	193.296
Tramo 1: Av 18 Julio - Av L.A. Herrera	181.515
Tramo 6: Corrales - Belloni	131.808
Tramo 7: Intercambiador	94.295
Tramo 3: Comercio - Larravide	94.279
Tramo 4: Larravide - Pan de Azúcar	90.111
Tramo 5: Pan de Azúcar - Corrales	57.471

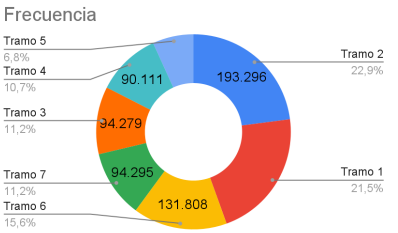


Figura 5.24: Distribución de ascensos por tramo en 8 de Octubre, sentido suroeste.

el suroeste no recorre el tramo 3 ni el 4. Para el intercambiador (tramo 7) se marcaron en celeste todas las líneas que utilizan al menos una parada.

De esto se desprende que de las 27 líneas, 6 recorren los 7 tramos del corredor, 4 recorren 6 tramos, 5 recorren 5 tramos, y el resto recorre 4 o menos tramos, llegando a haber 6 líneas que recorren solo 1 tramo.

Haciendo uso de log LOG_AGG_TRAZAS y utilizando la misma nomenclatura que en la Figura 5.15, en la Figura 5.26 se puede ver el porcentaje de cubrimiento de casos de cada actividad y transición, donde se observa que ninguna línea vuelve a ingresar al corredor luego de abandonarlo. Las salidas del corredor más frecuentes ocurren en los tramos 1 y 2, lo cual es lógico si se observa que las líneas con más ascensos recorren alguno de estos 2 tramos y solo 8 de 27 líneas no recorren los tramos 1 y/o 2. Además, el tramo 1 es el último en sentido suroeste, por lo cual todos los ómnibus que pasan por él luego abandonan el corredor. Desde los tramos 5, 6 y 7 se observa una cantidad similar de casos en los que se abandona el corredor y en los tramos 3 y 4 prácticamente no hay salidas del corredor.

Desde cualquier tramo del corredor la siguiente parada más frecuente es en

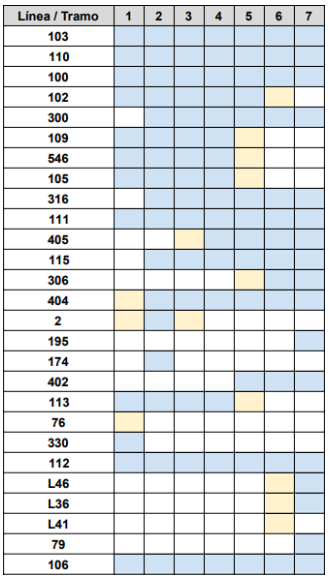


Figura 5.25: Tramos recorridos por cada línea en 8 de Octubre.

el siguiente tramo (hacia fuera del corredor en el caso del tramo 1 que es el último hacia el suroeste). Los saltos de tramo más frecuentes ocurren del tramo 6 al 4 y desde el tramo 4 al 2. Los tramos más comunes en los viajes hacia el suroeste son el 2 y el 6, pero en general se ve cómo todos los tramos son bastante frecuentes, siendo el intercambiador (7) el menos común.

Con respecto a la pregunta 2, en la Figura 5.27 se presenta un modelo generado sobre el log LOG_AGG_TRAZAS. El intercambiador (tramo 7) no aparece en el modelo, ya que cada variante utiliza como máximo una de las paradas en el intercambiador.

Observando el modelo destaca la duración de los tramos 1, 2 y 6, lo cual no es raro dado que son los tramos más largos. El tramo 5 es el que tiene menor duración, a pesar de que es un poco más largo que los tramos 3 y 4, los cuales tienen el mismo largo.

Mirando las transiciones llama la atención la duración de los pasajes entre el tramo 2 y 1 que es el más alto, así como entre el tramo 4 y 3 que es el más pequeño, los demás son muy cercanos entre sí. Una transición no incluida en el modelo por los filtros mencionados es desde el tramo 1 hacia afuera del corredor, la cual tiene una duración media de 3,3 minutos (mediana 2,6), superando al resto por bastante. Esto tiene sentido observando que el tramo 1 tiene un trayecto importante sin paradas (antes y después del túnel) que es parte del tiempo de la transición.

De las duraciones en los saltos de tramos se observa que siempre son un poco menores a la suma de las duraciones medias del recorrido tramo a tramo, siendo mayor la diferencia cuantos más tramos se saltean. Esto es lo esperado dado que

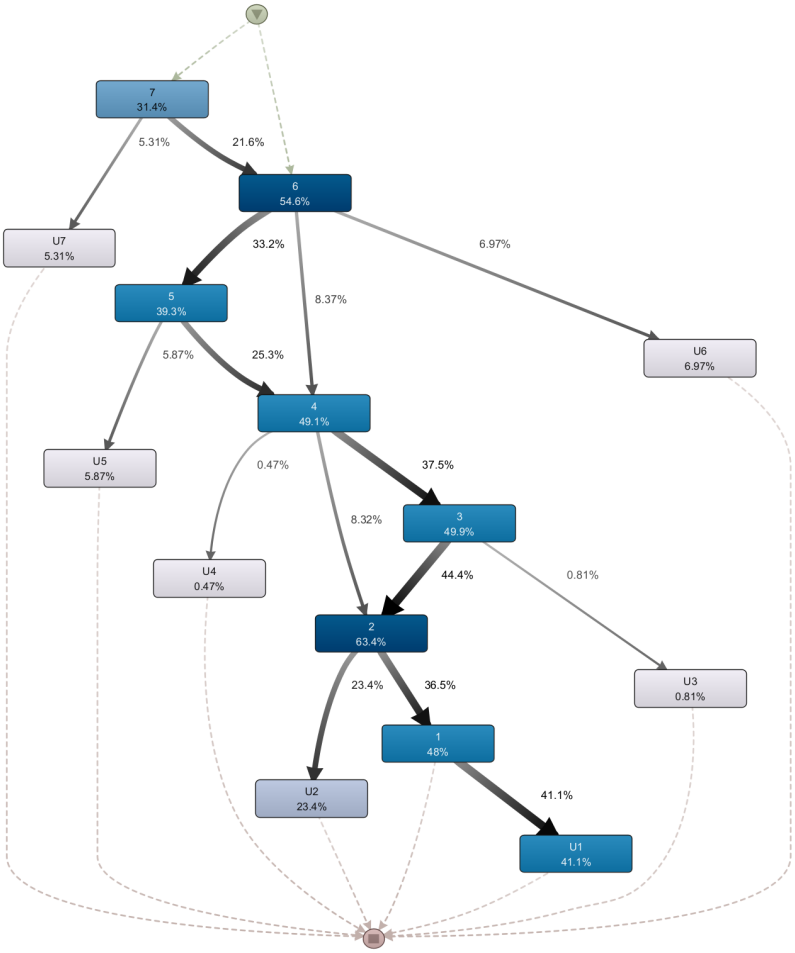


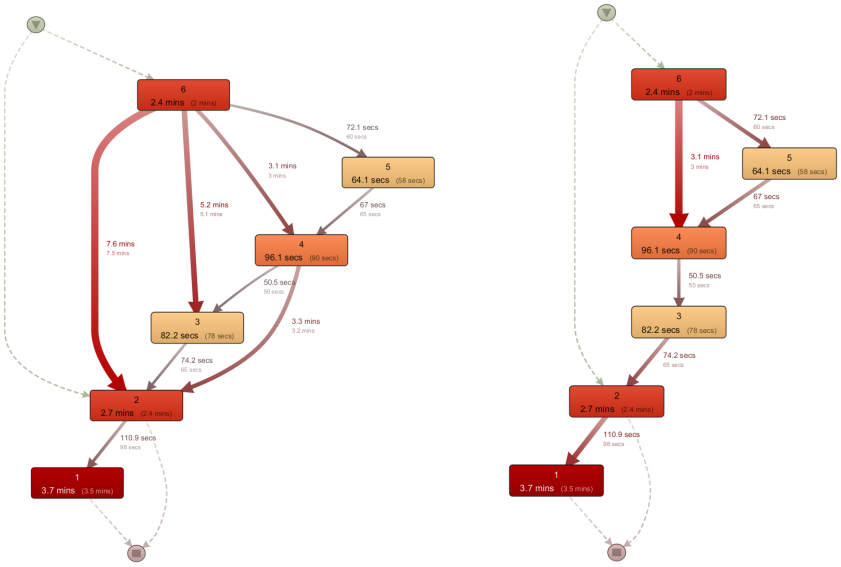
Figura 5.26: Modelo de proceso para el flujo de ómnibus con dirección suroeste sobre tramos de 8 de Octubre. Se muestra el porcentaje de cobertura de casos.

ninguna de las líneas que sale del corredor vuelve a ingresar posteriormente.

Dado que los tramos en los que se dividió el corredor tienen diferencia importante de longitud, se realizó un cálculo aproximado de la velocidad entre los distintos tramos que se presenta en la Tabla 5.4. En lugar de separar tramos y transiciones, se agrupó el tramo junto con la transición al siguiente tramo, como se indica en la columna "Trayecto".

El largo total del trayecto es de 6,2 km, la duración media total es de 22,3 minutos y la velocidad media general del corredor es de 16,6 km/h.

En la Tabla 5.4 se observa cierta simetría en las velocidades que no era



(a) **(b)**

Figura 5.27: Modelo de proceso para el flujo de ómnibus con dirección suroeste sobre tramos de 8 de Octubre con características de performance. Se muestra la media y mediana de las duraciones (a) completo (b) transiciones más frecuentes.

Tabla 5.4: Velocidad aproximada por tramo de 8 octubre, en sentido suroeste.

Trayecto	Vel. media aprox. (km/h)
$6 \rightarrow 5$	20,53
$5 \rightarrow 4$	16,71
$4 \rightarrow 3$	11,24
$3 \rightarrow 2$	11,17
$2 \rightarrow 1$	15,11
$1 \rightarrow$ U1	19,71

evidente en el modelo. Los tramos en los que los ómnibus se desplazan más rápido (aprox. 20 km/h) son el primero y último. La velocidad entre los tramos 5 a 4 y 2 a 1 es muy cercana a la velocidad promedio general (16,6 km/h). La mayor lentitud (aprox. 11 km/h) se observa entre los tramos 4 y 2, poco más de la mitad de la velocidad de los más rápidos.

Algo para destacar es que procesando los eventos se descubre una clara bifurcación de los recorridos con relación a las paradas, en los tramos 2 al 4 para ambos sentidos. Esto puede observarse en la Figura 5.28, con una clara bifurcación de los recorridos a partir de la parada 4865. Por un lado, se si-

gue la secuencia 1: 3190-3192-3194-3196-3198-3200 y por el otro la secuencia 2: 3189-3191-3193-3195-3197-3199. Esto se explica porque diferentes líneas utilizan diferentes paradas durante el recorrido, siendo claramente más utilizada la primera secuencia, coincidiendo esto último con que 6 de las 8 paradas con más ascensos correspondan a esta secuencia.

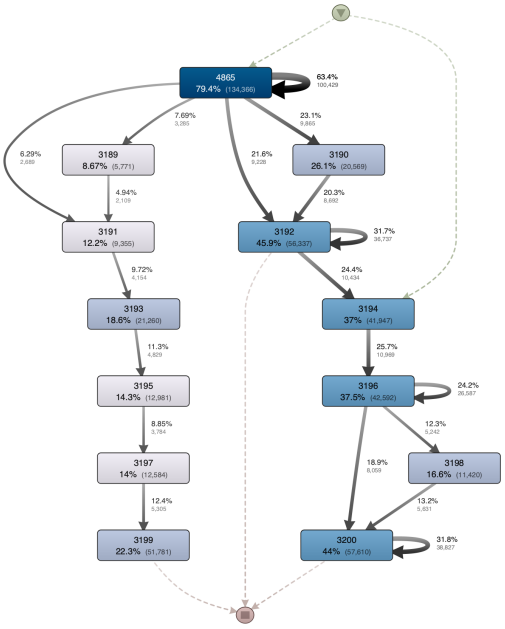


Figura 5.28: Modelo de proceso de los tramos 2 al 4 de Av. 8 de Octubre, con paradas como actividades, en sentido noreste. Con nivel de detalle de actividades al 90 % y transiciones al 10 %.

Tabla 5.5: Trazas principales en los tramos 2 al 4 en Av. 8 de Octubre, en sentido noreste.

Traza	Ocurrencias	Proporción
4865-3192	2.163	5,06 %
4865	1.914	4,48 %
4865-3190-3192-3194-3196-3200	1.828	4,30 %
4865-3190-3192	1.803	4,22 %
4865-3192-3194-3196-3200	1.626	3,81 %
4865-3190-3192-3194-3196-3198-3200	1.438	3,37 %
4865-3192-3194-3196-3198-3200	1.041	2,44 %
4865-3189-3191-3193-3195-3197-3199	913	2,14 %
4865-3191-3193-3195-3197-3199	879	2,06 %

La Tabla 5.5 presenta las principales trazas observadas en estos tramos, por ejemplo, la traza donde todos los casos paran únicamente en las paradas 4865, 3190 y 3192, y en ese orden, se da en un 4,22 % de los casos. Se aprecia que en un 27,68 % de los casos corresponde a viajes que realizan la secuencia 1. En este caso, no existen recorridos que paren en todas las paradas del tramo, por lo que el máximo de paradas consecutivas que pueden darse en este caso son 7 paradas, en ambas secuencias. Para la secuencia 1 esto se da en un 3,37 % de los casos (1.438 viajes), y para la secuencia 2 en un 2,14 % (913 viajes).

Con relación a la pregunta 3, en la Figura 5.23 y Tabla 5.3 puede verse que la mayor cantidad de ascensos en el sentido suroeste se da en el tramo 2, así como en el sentido noreste la Tabla 5.6 refleja que el tramo más concurrido (ignorando al intercambiador) es el 1. Puede destacarse también que las paradas más frecuentes en ambos sentidos están en la esquina con Bv. Batlle y Ordóñez, cerca del hospital CASMU.

Como conclusiones generales del corredor, se pueden destacar las diferencias

que existen respecto a la cantidad de ascensos totales y por tramo en cada sentido, esto se refleja en la Tabla 5.6. Se observa una mayor cantidad de ascensos totales en el sentido noreste, sobre todo en el tramo 7 (intercambiador), donde los ascensos en el sentido noreste triplican al sentido suroeste, esto puede explicarse fácilmente por una mayor cantidad de líneas y paradas en el intercambiador en el sentido noreste en comparación con el suroeste. Aun así, también se observan mayores ascensos en varios tramos, salvo los tramos 5 y 6, en el resto el noreste registra, y con diferencia, bastantes más ascensos.

Tabla 5.6: Comparación de ascensos totales y por tramos en cada sentido del corredor 8 de Octubre.

Tramo	**Suroeste**	**Noreste**	**Diferencia (Noreste - Suroeste)**
1	181.515	274.793	+93.278 (+51,38 %)
2	193.304	226.390	+33.086 (+17,11 %)
3	94.279	118.780	+24.501 (+25,98 %)
4	90.111	133.395	+43.284 (+48,03)
5	57.471	38.197	-19.274 (-33,53 %)
6	131.808	48.517	-83.291 (-63,19 %)
7	100.208	300.622	+200.414 (+199,99 %)
Total	**848.696**	**1.140.694**	**+291.998 (+34,40 %)**

Se puede señalar la clara relevancia de los tramos 1 y 2 (desde 18 de Julio hasta Comercio) en ambos sentidos, donde en el suroeste representan la mayor cantidad de ascensos aun incluyendo el tramo 7, y en el noreste lo son si se excluye al tramo 7.

En el sentido suroeste, 3 de cada 5 ascensos se dan en los tramos 1, 2 y 6. Por otro lado, una cantidad considerable de viajes tienen sus primeros ascensos en los tramos 4, 6 y 7, así como también la mayoría tienen sus últimos ascensos en los tramos 1, 2 (estos esperables) y 6. Por otro lado, en el sentido noreste, poco más de la mitad de los ascensos se dan en los tramos 1 y 7, y 4 de cada 5 ascensos se dan en los tramos 1, 2, 4 y 7. Luego de eso, los tramos 3 y 4 tienen cantidades de ascensos similares, mientras los tramos 5 y 6 registran muy pocos ascensos, lo cual quizás es un poco esperable al estar cerca del final del corredor, y también cerca del tramo con más ascensos, el intercambiador. Por otro lado, la gran mayoría de viajes tienen sus primeros ascensos en el tramo 1, y también en cantidades no despreciables en los tramos 2 y 7. En cuanto a los últimos ascensos de los viajes, como era de esperar, la gran mayoría están en el tramo 7, destacando también los tramos 2 y 4 en ese sentido.

En cuanto a líneas y variantes, puede decirse que las líneas 103 y 110 son las de mayor cantidad de ascensos en ambos sentidos, aunque mientras que la línea 103 predomina en casi todos los tramos (excepto intercambiador en sentido suroeste) la línea 110 no lo hace en ninguno. Sin embargo, mantiene un alto nivel de ascensos de forma pareja entre todos los tramos.

En cuanto a los ascensos por tipo de día (hábil, sábado, domingo/feriado), en la Figura 5.29 se observa que, como era de esperar, la mayor cantidad de

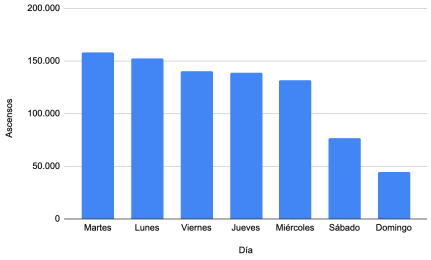


Figura 5.29: Cantidad de ascensos por día de semana en Av. 8 de Octubre, sentido suroeste.

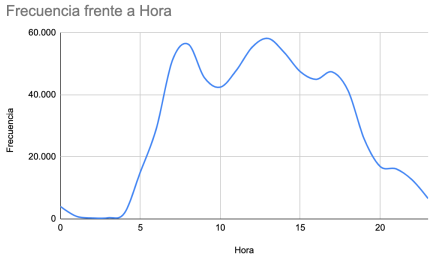


Figura 5.30: Ascensos por hora en el corredor de Av. 8 de Octubre en días hábiles, sentido suroeste.

ascensos se dan en días hábiles, alcanzando un 85,6 % en el sentido suroeste, y 83,8 % en el noreste. La distribución de ascensos es seguida en ambos sentidos por los días sábado, con una proporción en el entorno del 10 %, y los domingos con alrededor de 6 %. Se encuentra la particularidad también que en ambos sentidos el día con mayor ascensos es el martes, con alrededor del 18,6 %, y el menor los miércoles, con cerca de 15,3 %.

Respecto a la pregunta 4, la Figura 5.30 muestra que, teniendo en cuenta la cantidad de ascensos en el tiempo, para los días hábiles se tiene que hacia el suroeste hay dos picos de ascensos, uno a las 8 horas y otro entre las 12 y 14 horas, teniendo alrededor de las 13 horas un promedio de 56 ascensos por minuto a lo largo de todo el corredor. Hacia el noreste el pico es entre las 17 y 18 horas, registrando en ese intervalo un promedio de 84 ascensos por minuto a lo largo de todo el corredor (Figura 5.31).

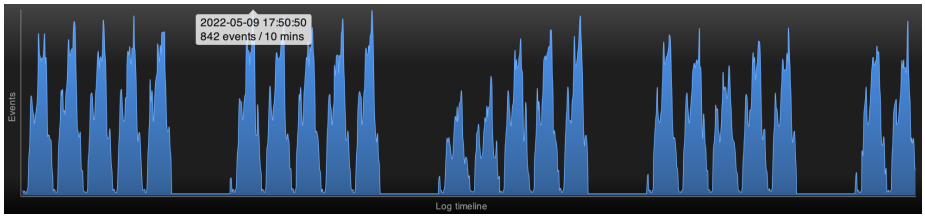


Figura 5.31: Cantidad de ascensos ocurridos en todos los tramos en días hábiles, sentido noreste.

Dado que las intersecciones del corredor con diferentes calles por lo general tienen dos paradas (una en cada sentido), se utiliza la notación *"(X NE, Y SO)"* para indicar que en dicha intersección se encuentran la parada con *codigo_parada_origen=X* en sentido noreste (NE) y la parada con *codigo_parada_origen=Y* en sentido suroeste (SO).

En el tramo 1, en ambos sentidos la línea que más ascensos registra es la 103, y si bien en el noreste lo hace con una amplia diferencia, seguido de la línea 110, en el suroeste está un poco más distribuida junto con las líneas 546,

110 y 102. Se observa que en estos tramos las paradas con más ascensos son las que se encuentran en las intersecciones con L.A de Herrera (3186 NE, 3230 SO) y Garibaldi (4789 NE, 4207 SO), en el sentido suroeste también destaca la parada 3231 frente al hospital militar. Hacia el noreste puede destacarse que el 24 % de los viajes solo utilizaron las paradas 3186 y 4729, mientras que hacia el suroeste un 28,47 % solo pararon en las paradas 3230, 3231 y 4207. En cuanto a la participación de las empresas, en ambos sentidos quien registra más ascensos es CUTCSA, seguido de forma pareja por COMESA y COETC, con la particularidad de que en el sentido suroeste COMESA registra 4 veces más ascensos que la línea UCOT. Sin embargo, la línea de la empresa COMESA tiene en promedio 1 minuto más de permanencia que la de la de UCOT, a pesar de que esta última tiene una parada más en el recorrido en este tramo.

Con respecto a los tramos 2 al 4, se destaca que nuevamente la línea con más ascensos en ambos sentidos es la 103, donde en el sentido noreste es seguida por las líneas 105 y 109, y en el sentido suroeste por las líneas 300 y 546. En ambos sentidos se repiten las paradas con más ascensos, siendo las más importantes las cercanas a la intersección con Bv. Batlle y Ordóñez (4865 NE, 3227 SE), teniendo una amplia importancia en el sentido noreste, y estando frente al hospital CASMU en el sentido suroeste. También se destacan en ambos sentidos las paradas en las intersecciones con Pan de Azúcar (3199 y 3200 NE, 3217 SO) y Comercio (3192 NE, 3223 SO). En estos tramos algo para destacar es que procesando los eventos se descubre una clara bifurcación de los recorridos con relación a las paradas, como se aprecia en la figura 5.28. En el sentido suroeste, la secuencia menos empleada es recorrida por las líneas 115, 174, 300, 316 y 404, mientras que su equivalente en el sentido noreste es recorrida por las líneas 105, 109, 111, 112, 113 y 546. En cuanto la distribución de cantidad de ascensos entre empresas, no hay grandes cambios en el sentido noreste, mientras que en el sentido suroeste se destaca que la empresa CUTCSA, a pesar de tener mucha mayor cantidad de ascensos, aun así tiene menores tiempos medios de permanencia. Luego, entre las otras empresas, se destaca la empresa COETC con menores tiempos de permanencia respecto al resto.

Respecto a los tramos 5 y 6, puede decirse que nuevamente la línea con más ascensos es la 103 en ambos sentidos. No obstante, y de forma similar a lo que ocurre en el tramo 1, en el sentido noreste la distribución está muy pareja con la línea 102, mientras que en el suroeste hay más paridad junto con las líneas 102, 100, 110 y 316. En el sentido noreste, las paradas con más ascensos se distribuyen de forma pareja entre la 3207 (Gdor. Vigodet), 3204 (Ramón Castriz - Corrales), 3206 (Habana) y 3202 (José Villagrán). En el sentido suroeste, las paradas con más ascensos son 3209 (frente al Intercambiador), 2546 (Gdor. Vigodet) y 3215 (20 de febrero). En cuanto a las trazas de recorridos, en el sentido noreste un 42,31 % corresponde a viajes que pararon en una única parada, y en total un 62,87 % responden a combinaciones de máximo 2 paradas, destacando las paradas 3202, 3204 y 3207. Respecto a la distribución de ascensos entre empresas, se destaca que la empresa COMESA no tiene ascensos en estos tramos, mientras que el orden de cantidad de ascensos de las otras 3 empresas en estos tramos se mantiene igual que el corredor en general, primero CUTCSA, luego

UCOT y finalmente COETC.

En cuanto al análisis por empresa, puede verse que el común denominador es que la mayor cantidad de ascensos, para todas las empresas y en ambos sentidos, se da en los extremos del corredor, siendo esto más claro aún en las empresas UCOT y COETC. También se observa que los ascensos en la empresa CUTCSA son los que se encuentran mejor distribuidos, esto puede deberse en parte a la mayor cantidad/variedad de líneas, y que también en ciertos puntos del corredor (en los tramos intermedios, sobre todo, y en el sentido suroeste) los ómnibus van "todos para el mismo lado", favoreciendo esa distribución más pareja. De todo el análisis anterior por tramos, queda en evidencia la amplia mayoría de ascensos hechos en la empresa CUTCSA, seguido casi siempre de las empresas UCOT y COETC, en ese orden, y observando también que en el tramo 1 la empresa COMESA cobra una gran relevancia a pesar de tener una única línea, la 546. De igual manera, en este mismo tramo en sentido suroeste, esta empresa registra tiempos de permanencia mayores al resto, a pesar de tener 1 parada menos en el recorrido, aunque esto puede ser atribuido a una mayor cantidad de ascensos. En el mismo sentido suroeste, pero en los tramos 2 al 4, es de destacar los mejores tiempos de la empresa CUTCSA, a pesar de que como se mencionó, tiene con diferencia la mayor cantidad de ascensos.

5.2.3. Multitramos en 8 de Octubre

En la presente sección se analizan los viajes multitramos en 8 de Octubre para responder las preguntas 5 y 6 planteadas en la sección 5.1. Se utilizó el log de multitramos de 8 de Octubre (LOG_MULTI8).

Definición: un viaje multitramo es un viaje realizado por una misma persona que tiene varios tramos, separados por ascensos en paradas no necesariamente diferentes. En los datos abiertos, un viaje se considera multitramo si el atributo *id_viaje* aparece al menos 2 veces en el mes.

Si bien este análisis puede ser obtenido con consultas a base de datos, se hizo uso de la herramienta de minería de procesos, ya que lo que se quiere analizar son: los trasbordos de cada usuario del STM, que es un proceso y, por lo tanto, es posible sacar ventaja de la herramienta. Alguna de las ventajas son:

- Filtrado rápido: una vez cargado los datos en la herramienta, rápidamente se pueden realizar filtros que son de relevancia (e.g obtener viajes con combinaciones tengan como último ascenso en 8 de Octubre).
- Visualización de variantes principales.
- Visualización de variantes alternativas.
- Exploración de los datos y obtención de nuevo conocimiento.

En la Tabla 5.7 se muestran datos estadísticos de los multitramos de 8 de Octubre.

En base a la definición de viaje multitramo, se toma el campo *id_viaje* como ID de Caso, el campo *en_corredor* como actividad, y el campo *fecha_evento* como marca de tiempo. Considerando si las paradas del multitramo se encuentran o

Tabla 5.7: Datos estadísticos de los multitramos de 8 de Octubre.

Multitramos 8 de Octubre	
# Ascensos	1.535.053
# Viajes Multitramo	726.834

no en 8 de Octubre, entonces las variantes principales son las que muestra la Tabla 5.8.

Tabla 5.8: Variantes principales de los multitramos de 8 de Octubre.

Variantes Principales	
Variante	**Cantidad de Viajes**
F - 8 Octubre	465.354
8 Octubre - F	149.980
8 Octubre - 8 Octubre	42.185
Otras (más de 2 paradas)	69.315

La Tabla 5.8 se lee de la siguiente manera:

- 465.354 viajes comienzan con una parada **fuera** de 8 de Octubre y la siguiente es una parada **dentro** de 8 de Octubre.
- 149.980 viajes comienzan con una parada **dentro** de 8 Octubre y la siguiente es una parada **fuera** de 8 de Octubre.
- 42.185 viajes comienzan con una parada **dentro** de 8 Octubre y la siguiente también es una parada **dentro** de 8 de Octubre.
- 69.315 viajes tuvieron más de 2 paradas donde al menos 1 de ellas se encuentra en el corredor 8 de Octubre.

Respecto a la pregunta 5, y utilizando el campo del barrio de la parada como actividad, es posible analizar los multitramos por barrio para saber entre qué barrios se realiza la mayor cantidad de transbordos (viajes multitramos), como muestra la Tabla 5.9.

Tabla 5.9: Variantes principales (por barrio) de los mulitramos de 8 de Octubre.

Variantes Principales (Por Barrio)	
Variante	**Cantidad de Viajes**
UNIÓN - UNIÓN	32.590
FLOR DE MAROÑAS - UNIÓN	19.292
PTA. RIELES, BELLA ITALIA - UNIÓN	17.900
JARDINES DEL HIPÓDROMO - UNIÓN	16.747
BUCEO - UNIÓN	16.640

De la Tabla 5.9, se muestran simplemente las primeras 5 variantes principales. Hay muchas otras, en particular, hay 15.969 variantes distintas. Sin embargo,

la mayoría tienen pocos casos (la gran mayoría solo 1). Esto se debe a que hay muchos viajes que son "únicos" en cuanto a los barrios que presentan ascensos a lo largo del viaje. Un ejemplo de esto es la siguiente variante "única":

VILLA ESPAÑOLA - UNIÓN - CORDÓN - LA BLANQUEADA

Para esta variante solo hay un viaje con ascensos en esa combinación de barrios. En particular, el *tipo_viaje* es de 2 HORAS, y corresponde a un USUARIO CORRIENTE (*grupo_usuario*).

Existen muchas variantes "únicas", las cuales se pueden ver en la Figura 5.32 que muestra la cantidad de viajes acumulados por las variantes existentes.

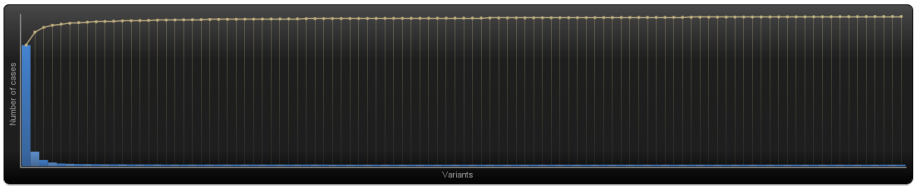


Figura 5.32: Distribución de variantes de multitramos en 8 de Octubre.

De la Figura 5.32 se concluye que, las primeras 159 variantes contienen el 80,72 % de los viajes multitramos. Además, si se consideran las primeras 318 variantes, se tiene el 89,48 % de los viajes multitramos. Las variantes restantes tienen muy pocos casos. En particular, a partir de la variante 319 se tienen como máximo 221 casos por variante. Se concluye que, con 318 variantes se cubre gran parte de los viajes multitramos, y además, ninguna de las variantes restantes influye demasiado en la cantidad de viajes totales. O lo que es lo mismo, hay 76.462 (10.52 %) viajes multitramos que se repiten pocas (o una) veces durante el mes.

Este análisis se puede detallar más considerando la parada, línea o variante del recorrido. Sin embargo, hay que tener en consideración que, en general, cuanto más granularidad se utilice, la cantidad de variantes de casos será mayor.

Conteo de ascensos agrupado por barrio

Como se planteó en la pregunta 6, interesa conocer las zonas que más se utilizan como entrada y/o salida de 8 de Octubre. Para esto, pueden obtenerse las paradas más utilizadas, y luego agruparlas por barrio. En este caso no se utiliza Disco, ya que se dificulta filtrar aquellas paradas de salida de tramo si no se tiene un atributo en el log que indique que es una parada de salida de 8 Octubre (mismo caso para entrada a 8 de Octubre).

Para este análisis se ejecutó un script que recorre el log (viaje a viaje), y cada vez que encuentra una parada de 8 de Octubre, la agrega al conteo (diferenciando si es de entrada y/o salida) y obtiene datos de la parada anterior/siguiente. Además, cuenta también cuáles fueron las líneas que se utilizaron para entrar a 8 Octubre.

Se puede diferenciar entre, conteos de ascensos previos a un ascenso en 8 de Octubre, y conteos de ascensos seguidos de un ascenso en 8 de Octubre.

Para el conteo de ascensos **previos** a un ascenso en 8 de Octubre, es posible contar valores absolutos (Figura 5.33), y se puede contar la cantidad de ascensos por barrio por cada parada (Figura 5.34)

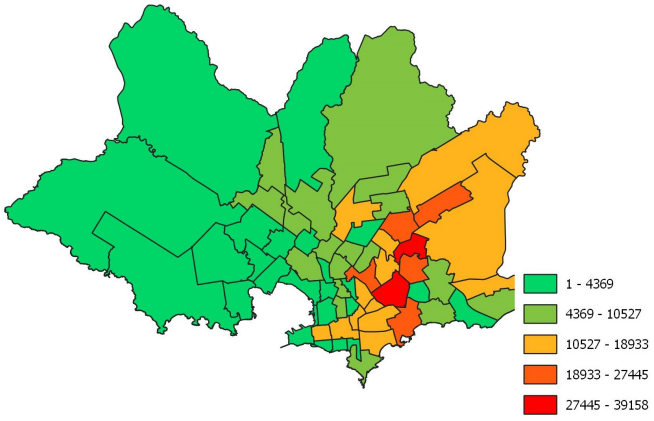


Figura 5.33: Cantidad de ascensos por barrio, previo a una entrada en 8 de Octubre

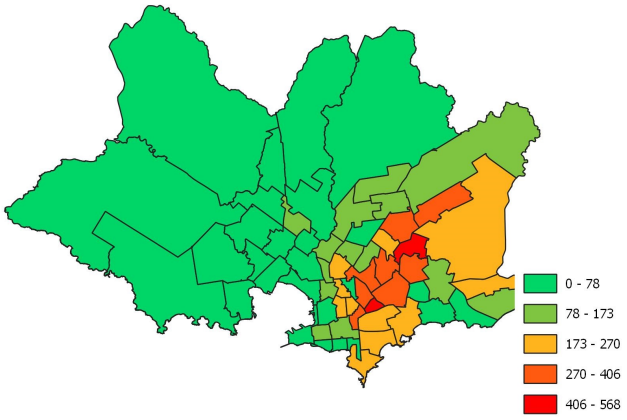


Figura 5.34: Cantidad de ascensos por barrio por cada parada, previo a una entrada de 8 de Octubre

Para el conteo de ascensos **seguidos** a un ascenso en 8 de Octubre, se pueden contar contar valores absolutos (Figura 5.35), y se puede contar la cantidad de ascensos por barrio por cada parada (Figura 5.36)

La Figura 5.36 no es idéntica a la Figura 5.35, pero los barrios más prominentes siguen siendo los que se encuentran en la línea que atraviesa 18 de Julio y 8 de Octubre, con algunas excepciones como Ciudad Vieja, Prado, Piedras

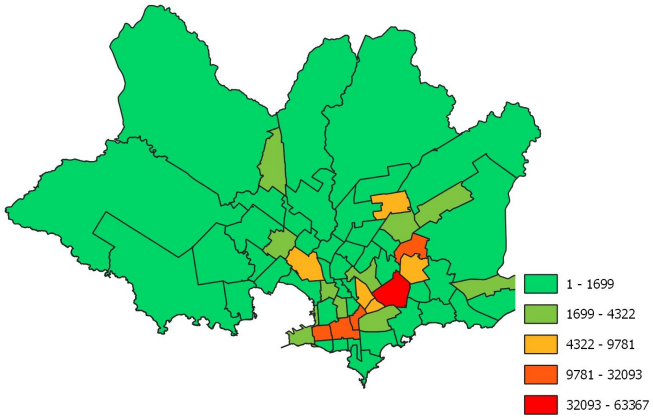


Figura 5.35: Cantidad de ascensos por barrio, seguido a una salida de 8 de Octubre

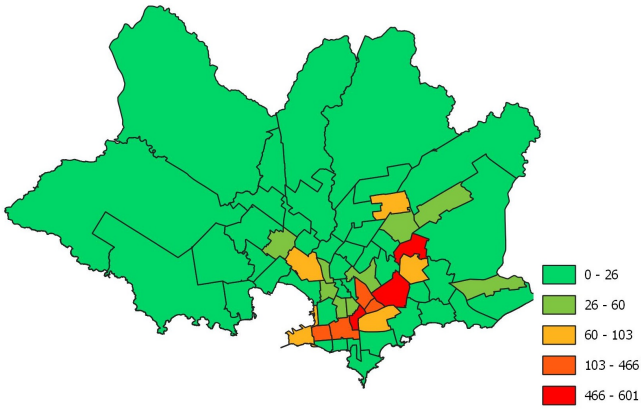


Figura 5.36: Cantidad de ascensos por barrio por cada parada, seguido a una salida de 8 de Octubre

Blancas y Parque Batlle, que tienen mayor importancia si se tiene en cuenta la cantidad de paradas en esos barrios.

Conclusión

En cuanto al análisis propio de los datos obtenidos de los multitramos, se puede destacar que los multitramos generalmente comienzan fuera del corredor 8 de Octubre y luego tienen un ascenso en el corredor. Sin embargo, si se realiza una agregación de los ascensos previos por barrios, se observa que Unión es el barrio que tiene mayor cantidad de estos ascensos. La Tabla 5.9 daba una idea de esto, sin embargo, las visualizaciones en los mapas lo confirman.

En tanto, si se considera qué sucede luego de un ascenso en el corredor 8 de Octubre, se observa que los ascensos siguen en la misma "línea" que une 18 de

Julio y 8 de Octubre.

En cuanto a qué combinación de barrios se usan para realizar ascensos, puede verse que los multitramos están concentrados en pocas combinaciones de barrios, con 318 de estas combinaciones, se obtienen casi el 90 % de los viajes, y además, el restante 10 % de los viajes se repiten pocas (o una) veces durante el mes.

Desde el punto de vista de la relación entre minería de procesos y el análisis de multitramos, se desprende que ciertos aspectos del proceso pueden ser analizados utilizando esta disciplina: analizar variantes principales (e.g combinación de barrios principales), cómo se distribuyen estas variantes, contar datos absolutos de los multitramos (e.g cantidad de ascensos o cantidad de viajes, líneas más utilizadas, etc). Para los casos que se quieran contar ascensos que dependan de otros ascensos (e.g un ascensos que están después que un ascenso en 8 de Octubre), es necesario el uso de otras técnicas como por ejemplo un script.

5.3. Reducir la brecha entre minería de procesos y GIS

Como se mencionó en la Introducción, la minería de procesos es una disciplina que está inmersa en el área de la Ciencia de Datos, además de la Ciencia de Procesos. Por lo tanto, muchas veces puede tenerse la necesidad de realizar ciertos análisis complementarios de otras áreas, como clusterings, optimización, entre otras. Lo que se observó en el caso particular del análisis del transporte público, es que la minería de procesos carece de la dimensión geoespacial de los datos. En minería de procesos, una parada del sistema de transporte es simplemente un código que carece de información geoespacial. Aquí, es donde se observa que la minería de procesos puede ser complementada con el área de información geográfica. Para un analista puede resultar de especial interés ver información, datos y modelos en un mapa. En (Behkamal y cols., 2022) se analiza un caso particular de integración entre minería de procesos e información geoespacial, en el dominio de entrega de paquetes entre ciudades. Representar modelos en un mapa ayuda a los analistas a comprender mejor modelos complejos y permite utilizar información del contexto geográfico para el análisis de los mismos (Behkamal y cols., 2022). Es por esto, que se utilizó la minería de procesos y sistemas de información geográfica, de manera complementaria.

Por ejemplo, la Figura 5.37 muestra un modelo descubierto de un recorrido de un ómnibus, y la Figura 5.38, es la representación del mismo modelo en un sistema de información geográfico. El proceso del modelo es el recorrido de una variante específica, con las paradas del recorrido como actividades. Los metadatos de performance del modelo fueron exportados y proyectados en un GIS. De esta manera se puede ver la duración entre parada y parada del recorrido en un mapa.

En el ejemplo anterior se proyectó en el mapa la vista de performance del modelo. También se puede proyectar la información del flujo del modelo en el mapa, como se ve en la Figura 5.39.

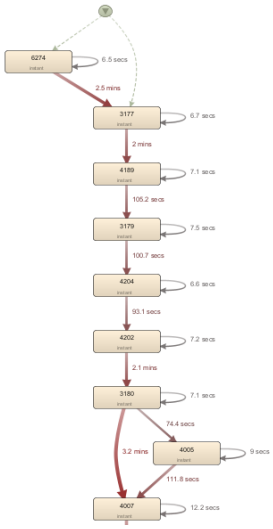


Figura 5.37: Modelo descubierto (recortado) de un recorrido de ómnibus.

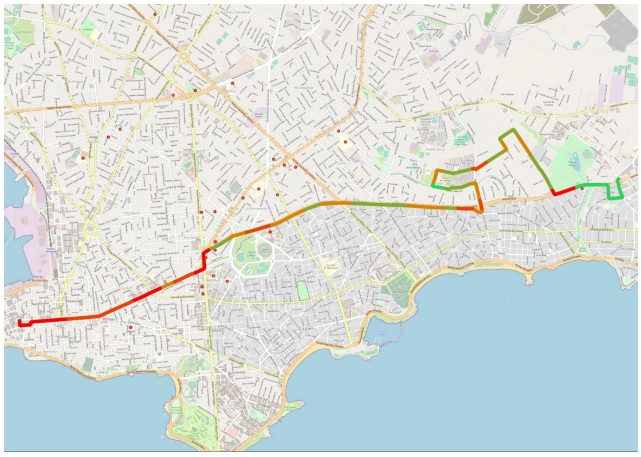


Figura 5.38: Vista de performance del modelo en GIS.

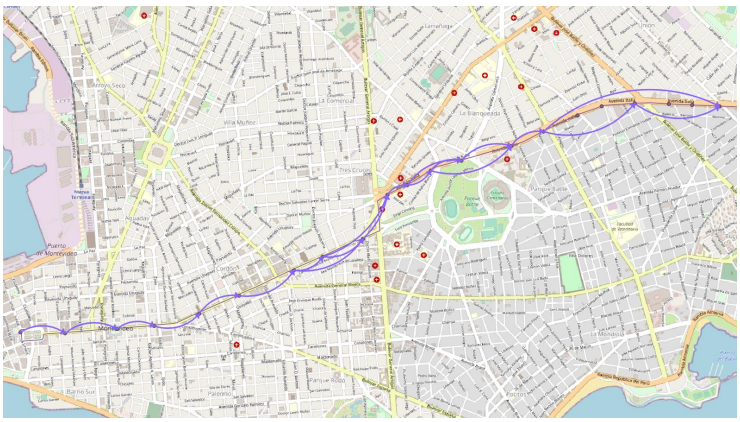


Figura 5.39: Vista de flujo de modelo en GIS.

La misma información que se ve a través de la herramienta de minería de procesos se puede visualizar en un mapa, lo que da un mayor contexto a la visualización. La Figura 5.40 muestra con mayor detalle una parte del mapa y en la Figura 5.41 se visualiza la contraparte de la Figura 5.40 desde una herramienta de minería de procesos.

Si se comparan las dos vistas del modelo (figuras 5.40 y 5.41), es notoria la diferencia. El mapa da un contexto mucho mayor, rápidamente se conoce la ubicación exacta de las actividades (paradas en este caso), lo que permite relacionar el modelo generado con información no incluida en este.

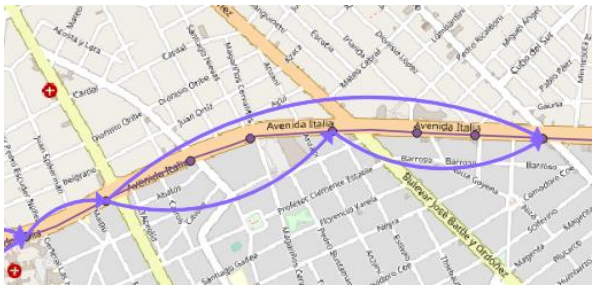


Figura 5.40: Visualización de bifurcación de flujo en GIS.

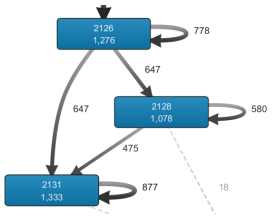


Figura 5.41: Visualización de bifurcación de flujo en Disco.

Por ejemplo, en este caso se podría investigar por qué en varias ocasiones no hay ascensos en la parada 2128. A modo de ejemplo, algunas causas podrían ser las siguientes:

- Infraestructura de las paradas.
- Seguridad de la zona.
- Ascensos de usuarios particulares de la zona (por ejemplo escolares).

5.3.1. Transformación del modelo a capa GIS

Para realizar la transformación del modelo a una capa GIS, se hizo uso de scripts, que toman como entrada el modelo generado (que es posible exportar desde Disco) y generan capas para visualizar en mapas.

En la Figura 5.42 se presenta un esquema del proceso de transformación del modelo hacia la visualización del mismo en un sistema geográfico.

Cabe destacar que el script fue implementado para la herramienta QGIS[2] y tiene que ser configurado para el proceso particular y ejecutado manualmente por el usuario.

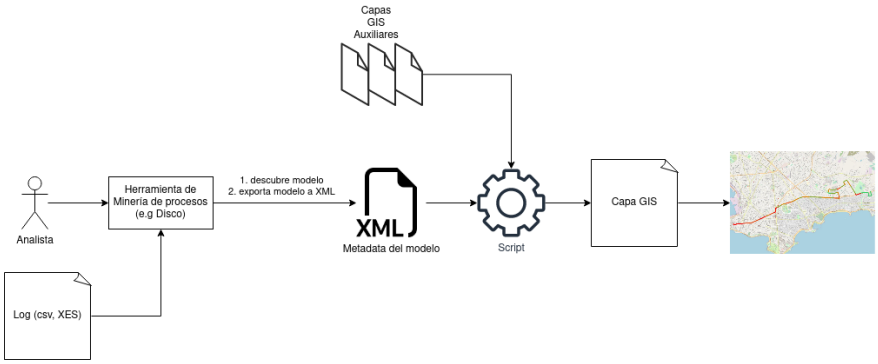


Figura 5.42: Esquema del proceso de transformación de modelo a visualización en mapa.

[2]QGIS. https://www.qgis.org/es/site/

5.3.2. Modificación de scripts a algoritmos de procesamiento en QGIS

Como la idea inicial del pasaje de modelo de Disco a QGIS fue realizar una prueba de concepto, la facilidad de uso no fue de relevancia en esta etapa. Para que no sea necesario modificar los scripts, se pueden adaptar estos a lo que se llama en QGIS, algoritmos de procesamiento. Esto permite que se ejecuten los scripts desde una interfaz amigable para el usuario, permitiendo elegir archivos desde el explorador de archivos (e.g elección del archivo XML del modelo) y cambiar variables asociadas al algoritmo (e.g sobre qué recorrido es el algoritmo, layers de entrada y salida, entre otros). Se considera que esta adaptación no es una tarea compleja, sin embargo, se priorizó realizar otros análisis y tareas en lugar de continuar su desarrollo.

5.3.3. Extensión

El poder utilizar los scripts como algoritmos de procesamiento en QGIS tiene como consecuencia una mayor facilidad de uso de los scripts. Siguiendo en esta línea se plantea si es siquiera necesario utilizar la interfaz de Disco para generar los modelos. Si la generación de modelos puede ser realizada de forma automática, se abre un abanico de posibilidades, entre las que se encuentran:

- Facilidad de uso.
- Menor tiempo entre generar un modelo y representarlo en un mapa.
- Poder generar el modelo directamente en el mapa, sin necesidad de hacer un post procesamiento del modelo.
- Generar modelos con datos en tiempo real y representar estos en el mapa.

Generar modelos con poca o nula intervención del usuario es sin duda algo a favor. Para esto, las herramientas de minería de procesos deben proveer APIs (Application Programming Interface) para poder generar los modelos. Este punto no lo cumple la herramienta Disco, por lo que no es posible utilizarla en este sentido. Aquí es donde destacan las herramientas open source como ProM, donde cada algoritmo que incorpora la herramienta (e.g algoritmo de descubrimiento de modelos) es un plugin que puede ser descargado y ejecutado de forma individual. Así, se pueden generar los modelos automáticamente y utilizarlos para otros procesamientos, en este caso para visualizarlos en un mapa.

5.4. Otros análisis de movilidad

Un análisis interesante a realizar contando con el dato de tarjeta de pasajero (disponible en VROSTM_FT) es la inferencia de destinos de los viajes que realizan los pasajeros (parada en que se baja el pasajero luego de un ascenso). Conocer los destinos posibilita entre otras cosas la generación de la matriz origen-destino, que describe la movilidad de una ciudad, indicando la cantidad de viajes entre pares de ubicaciones relevantes. También permite calcular la cantidad de gente a bordo de cada ómnibus en cada momento.

5.4.1. Inferencia de destinos

Un problema encontrado en los datos de ascensos públicos es que no tienen identificación de la tarjeta, no pudiendo identificar a las personas y, por lo tanto, limitando otros tipos de análisis. Avanzado el proyecto se logró conseguir estos datos para mayo de 2022, con lo cual se pudo realizar el estudio de inferencia de destinos y la generación de la matriz origen-destino. Para la inferencia de destinos de los viajes existen varios modelos. En (Li y cols., 2018) se presenta una lista de distintos trabajos publicados relacionados con la estimación de destinos en sistemas de transporte público. De los modelos existentes se decidió utilizar el de encadenamiento de viajes (trip chaining), propuesto originalmente en (Barry y cols., 2002) y específicamente del modo en que se detalla en (Massobrio, 2018), el cual trabaja sobre datos del sistema de transporte público de Montevideo, iguales a los que se poseen pero de mayo del 2015 en lugar de mayo del 2022.

Método de encadenamiento de viajes

El método de encadenamiento de viajes se basa fundamentalmente en dos hipótesis. La primera es que en general el origen de un viaje es próximo al destino del anterior y la segunda es que al final del día las personas por lo general retornan a su primer origen. Es decir, que para estimar la parada destino de un viaje de una persona, se buscará de entre las paradas que recorre el ómnibus (posteriores a la de ascenso), la más cercana a la parada de origen del siguiente viaje de la persona (o del primer viaje del día en caso de estar infiriendo el destino del último viaje del día).

Un parámetro importante es la distancia máxima permitida entre descenso y siguiente ascenso. Un valor muy grande es más proclive a identificar destinos erróneamente y uno muy pequeño identificará menos destinos. En este caso se utilizó 1000 metros, que es también el valor utilizado en (Massobrio, 2018) donde fue seleccionada por ser la mediana de las distancias utilizadas en trabajos revisados, así como la más frecuente.

Se utiliza las 3 am como inicio del día por ser la hora con menor número de transacciones, solamente 0.07 % del total de mayo. Este es el mismo criterio utilizado en (Massobrio, 2018) donde el horario con menos transacciones son las 4 am.

Obtención de la matriz origen-destino

Además de encontrar el destino de viajes particulares, interesa identificar el destino de viajes de más de un tramo, es decir, viajes en los que se producen transbordos hasta llegar a destino. Con esta información se puede generar una matriz origen-destino entre paradas, que luego se puede agrupar en municipios, códigos postales, barrios u otras divisiones geográficas de interés.

Para identificar estos viajes de múltiples tramos se decidió aplicar el mismo criterio aplicado en (Massobrio, 2018). En el conjunto de datos VROSTM_FT se tiene para cada transacción un *id_viaje* que es el mismo para viajes de más de un tramo realizados con la misma tarjeta, por ejemplo, 3 viajes realizados con

un mismo "boleto de 2 horas" compartirán el *id_viaje*. Se identifican entonces los pares origen-destino basándose en este registro.

Problemas de método y algunas posibles mejoras

El método presenta problemas para casos en los que se caminan largas distancias o se utilizan métodos de transporte alternativos y no es capaz de inferir destinos de personas que tienen un solo ascenso en el día. Existen técnicas que se pueden emplear para mitigar estos problemas como las que se exponen en (M. A. Munizaga y Palma, 2012) o también en (M. Munizaga, Devillaine, Navarrete, y Silva, 2014), también se podrían utilizar métodos alternativos para intentar inferir los destinos que el método de encadenamiento de viajes no pudo inferir.

El método también puede tener problemas en la inferencia de pares origen-destino, dado que algunas personas con un mismo boleto pueden desplazarse hasta un destino para realizar una o varias "actividades cortas" y luego volver o ir hacia otro sitio dentro de la validez del mismo boleto. En un caso como el anterior, lo ideal sería identificar 2 pares origen destino pero el método utilizado identificará solo uno dado que se utilizó el mismo boleto y por lo tanto los ascensos tendrán el mismo *id_viaje*. Esto también podría intentar mejorarse con criterios de tiempo máximo hasta nuevo ascenso como se propone en (M. Munizaga y cols., 2014). También se puede estimar el tiempo de caminata y verificar que la misma línea del ómnibus en el que hace el siguiente viaje no haya pasado una o más veces desde la hora en que se calcula que la persona llegó a la parada.

Detalles de implementación

El código fue hecho basándose en el pseudocódigo presentado en (Massobrio, 2018). La única diferencia importante es que en (Massobrio, 2018) si un destino no se puede inferir, entonces la cadena de viajes se considera rota y se pasa al siguiente pasajero, mientras que en esta implementación simplemente se pasa al siguiente viaje del mismo pasajero. Esto es porque aunque la inferencia de un destino falle se puede seguir intentando inferir el resto y en cuanto a la identificación de pares origen-destino tampoco cambia nada mientras que el identificador de viaje sea el mismo, el primer viaje para ese identificador tenga ordinal de tramo 1 y se haya podido inferir el destino del último viaje con ese identificador.

Al guardar pares origen-destino, se guarda también la hora del primer ascenso, esto es útil si se quiere analizar la matriz origen-destino en ciertos días y/o rangos horarios. Al sumar un par origen-destino, el algoritmo toma en cuenta la cantidad de pasajeros que viajan con el mismo boleto.

Porcentaje de inferencia

Para poder aplicar el método, las transacciones deben ser con tarjeta y una tarjeta debe tener al menos dos transacciones en un día para poder intentar in-

ferir los destinos de esa persona dicho día. Del total de transacciones de mayo de 2022, un 88 % son con tarjeta y de estas un 92 % cumplen la segunda condición.

Se logró inferir el destino del 88,6 % de los viajes y 12.244.408 pares origen-destino entre paradas.

Comparación con porcentaje de inferencia de 2015

En (Massobrio, 2018), para mayo de 2015 se tiene un total de 18.885.711 registros con tarjeta y con al menos dos transacciones por día y tarjeta. De estos, el método de encadenamiento de viajes logra inferir el destino del 81,6 %. En este trabajo, para mayo de 2022 se tienen un total de 20.310.118 registros con tarjeta y al menos dos transacciones por día y tarjeta. De estos, el método logra inferir el destino del 88,6 %. La diferencia en el porcentaje de inferencia de destinos podría en parte deberse a que el algoritmo presentado en (Massobrio, 2018) deja de procesar los viajes de un día para un pasajero luego del primer fallo en identificación de alguno de sus destinos.

Por otro lado, en (Massobrio, 2018) se infieren 9.465.314 pares origen-destino y en el trabajo actual se infieren 12.244.408. Esta diferencia se debe en parte por la mayor cantidad de transacciones de 2022 y nuevamente podría deberse a lo mencionado para inferencia de destinos puntuales.

Análisis de viajes origen-destino entre municipios

En la Figura 5.43 se muestra la división de Montevideo en sus 8 municipios. En la Tabla 5.10 se muestra el top 10 de combinaciones origen-destino entre municipios. Se puede ver que las primeras dos abarcan bastante más cantidad que el resto. Estas corresponden a viajes origen y destino dentro del mismo municipio, siendo la más frecuente dentro del municipio A y luego en el B. Se observa también que 7 de las 10 combinaciones más frecuentes son de este estilo, dentro del mismo municipio. La única no presente es la combinación {E, E} que está en la posición 13. Para todos los municipios se cumple que el destino más frecuente es dentro del mismo municipio, a excepción del CH, en el cual supera por muy poco la cantidad de viajes hacia el municipio B.

Otro factor destacable es que la mayoría de las combinaciones de la tabla son entre municipios adyacentes, a excepción de 4 combinaciones (A-B, B-A, B-E, D-B) que son todas desde o hacia el municipio B.

En la Tabla 5.11 se presenta un resumen con el porcentaje de viajes desde un municipio a sí mismo, hacia otros y desde otros hacia él. En esta tabla se observa como el municipio B es el más frecuente como origen y como destino. Los municipios CH y C también tienen gran cantidad de viajes hacia y desde otros municipios. El municipio A los sigue con un poco menos de viajes, pero como se vio es el municipio con más cantidad de viajes dentro de él mismo. En general se observa cómo los viajes desde un municipio hacia otros y desde otros hacia él son bastante parejos.

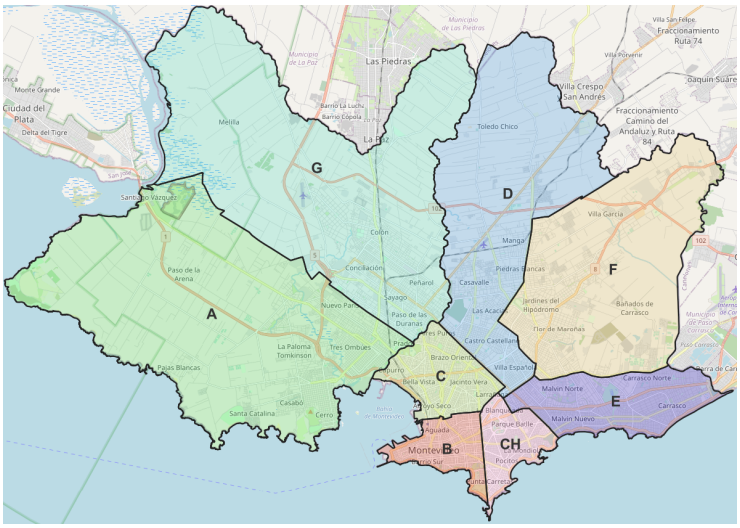


Figura 5.43: Municipios de Montevideo.

Tabla 5.10: Top 10 combinaciones más frecuentes de viajes origen-destino entre municipios.

Origen	Destino	Frecuencia	Proporción
A	A	836.068	6,62 %
B	B	724.209	5,74 %
G	G	446.223	3,54 %
B	CH	445.359	3,53 %
F	F	429.425	3,40 %
D	D	408.848	3,24 %
CH	B	389.189	3,08 %
CH	CH	381.267	3,02 %
C	C	361.669	2,87 %
C	B	339.951	2,69 %

5.4.2. Matriz origen-destino

En (Massobrio, 2018) se realiza un mapa de calor de la matriz origen-destino de viajes entre municipios de Montevideo para mayo de 2015 y se compara estos resultados con los presentados en una encuesta de movilidad de 2016 (Mauttone y Hernández, 2017). Se decidió efectuar una comparación entre la matriz origen-destino inferida para mayo de 2022 con la presentada en (Massobrio, 2018) para mayo de 2015.

En la Figura 5.44 se muestra el mapa de calor correspondiente a la matriz origen-destino de mayo de 2022 y en la Figura 5.45 se muestra el resultado presentado en (Massobrio, 2018) correspondiente a la matriz origen-destino de mayo de 2015.

Tabla 5.11: Porcentaje total de viajes desde cada municipio a sí mismo, hacia otros y desde otros hacia él.

Municipio	A sí mismo	Hacia otros	Desde otros
A	6,62	7,04	6,93
B	5,74	15,12	13,95
C	2,87	9,87	9,37
CH	3,02	10,34	11,20
D	3,24	6,90	8,39
E	2,38	6,72	6,20
F	3,40	6,53	6,18
G	3,54	5,31	5,25

Las escalas utilizadas permiten una comparación de proporciones en lugar de valores puntuales de viajes de cada combinación. Se observa a simple vista como los mapas de calor son muy similares, es decir, que con una diferencia de 7 años no se perciben grandes cambios en la distribución de viajes.

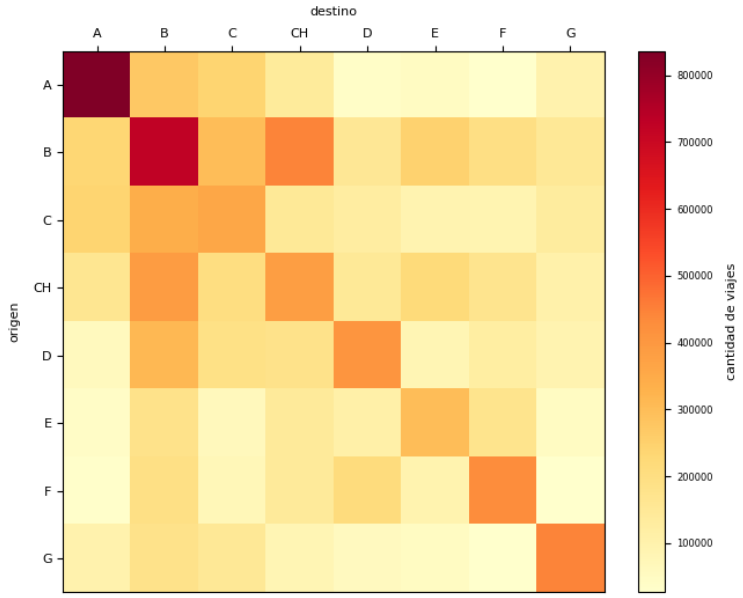


Figura 5.44: Mapa de calor de la matriz origen-destino entre municipios de Montevideo en mayo de 2022.

Para analizar las diferencias de manera más fácil se realiza una gráfica similar a la de los mapas de calor pero con la diferencia de proporción. Más específicamente, se generan dos matrices, una para el 2015 y otra para el 2022, en las cuales cada celda tiene el porcentaje de viajes entre el origen y destino correspondientes sobre el total de viajes inferidos. Luego se computa una nueva

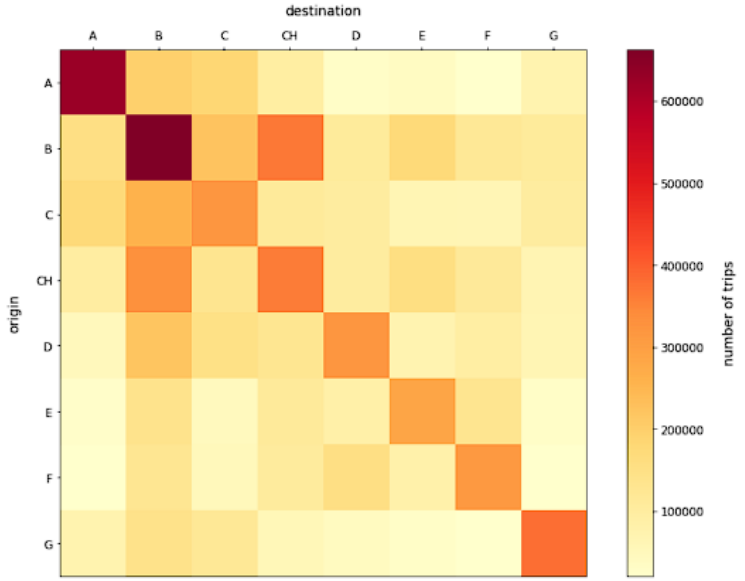


Figura 5.45: Mapa de calor de la matriz origen-destino entre municipios de Montevideo en mayo de 2015 presentada en (Massobrio, 2018).

matriz como la diferencia entre la matriz de 2022 y la de 2015. En la Figura 5.46 se puede ver el resultado.

Observando la Figura 5.46 se puede ver que los cambios son mínimos, notar como la escala va desde -1.1 % hasta 1.1 %. El valor mínimo (o mayor decremento) es -1,09 y corresponde a la diferencia en proporción de viajes dentro del municipio B (origen y destino B). El mayor incremento se da en los viajes entre el municipio B y el F con 0.34 %. En la diagonal se puede ver la variación de los viajes desde un municipio a sí mismo (viajes internos). Se observa como en general hay un decremento leve en la proporción de viajes internos, salvo en 2 de los 8 municipios (A y F) en los cuales hay un pequeño incremento.

Si en lugar de observar pares origen-destino puntuales se observan las columnas, entonces se puede ver el incremento o decremento general de viajes con destino el municipio correspondiente a la columna y al observar las filas se tiene el incremento o decremento de viajes con origen el municipio correspondiente a la fila.

El municipio A es el que tiene mayor incremento de viajes hacia él, con la mayoría de los viajes provenientes de los municipios B, CH, sí mismo y C. Le sigue F con más viajes provenientes de B principalmente, y también desde sí mismo y CH. A estos les sigue el municipio D con un incremento de viajes mayoritariamente desde CH, B y F. Estos municipios (A, F y D) no tienen prácticamente decremento de viajes desde ningún municipio.

El municipio B es el que presenta mayor decremento global de viajes internos. El decremento de viajes se da principalmente por menos viajes internos y menos

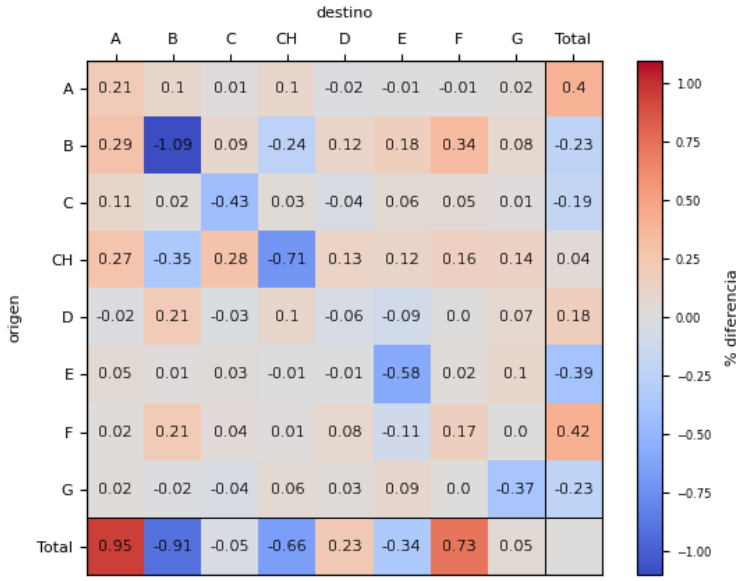


Figura 5.46: Mapa de calor de la diferencia en proporción de viajes origen-destino entre los resultados de 2022 y 2015.

viajes desde CH, pero también tiene una incremento considerable de viajes desde los municipios D, F y A. Le sigue el municipio CH, que como se puede ver, también tiene una reducción, principalmente en los viajes internos y menos viajes desde B, si bien tiene un leve incremento en los viajes desde A y D. A estos les sigue el municipio E, que tiene principalmente una reducción de viajes internos, así como reducción leve de pequeña de viajes desde D y F.

La variación general de viajes hacia los municipios C y G es casi nula, en C por ejemplo se puede ver cómo hay menos viajes internos, pero llegan más viajes principalmente desde CH y B. En G la situación es similar, se ven menos viajes internos pero más viajes desde CH, B y D mayoritariamente.

Se realizó además un mapa de calor similar, pero tomando en cuenta solo los fines de semana, al igual que se hizo en (Massobrio, 2018). La comparación entre estos se puede ver en el documento anexo "Inferencia de destinos". En este documento anexo también se puede ver el análisis para otras agrupaciones geográficas más pequeñas (códigos postales y barrios). El documento también tiene mayor detalle en varias secciones, especialmente en los detalles de implementación y en el apéndice del documento se pueden ver mapas de porcentaje de viajes entre municipios de Montevideo que brindan otra visión a los datos de la matriz origen-destino.

Capítulo 6

Conclusiones y trabajo a futuro

En esta sección se presentan las principales conclusiones y aportes de este proyecto, así como el posible trabajo a futuro a realizar sobre el mismo.

6.1. Conclusiones

El objetivo principal de este proyecto fue explorar la aplicación de técnicas de minería de procesos destinadas a la búsqueda de soluciones a problemas de interés de la IM en relación con la movilidad urbana. Para cumplir con este objetivo se realizó parte del análisis de corredores y de multitramos utilizando minería de procesos. En particular, se utilizó la etapa de descubrimiento de modelos para poder obtener una visualización simplificada de los corredores y multitramos, usando datos de ascensos, de modo de obtener conocimiento de flujos y performance.

Se desarrolló un análisis de los corredores Av. Italia y 8 de Octubre desde varias dimensiones, teniendo en cuenta el sentido de viaje, tramo del corredor, paradas, tipo de día (hábil, sábado, domingo), horas del día, empresa del ómnibus, líneas, variantes y otros. Se propuso un método de agregación de los eventos según los tramos del corredor, se presentaron criterios para la definición de los tramos, su utilidad y los inconvenientes de esta división para los datos de ascenso en paradas (sin información de GPS).

De estos análisis de corredores se concluye que la aplicación de minería de procesos puede ser muy útil para visualizar comportamientos en trayectos específicos, ya sea descubriendo patrones de recorridos más recurrentes, complementando información de performance existente, comparando el desfase de tiempo entre buses, o proporcionando otra manera de explicar comportamientos ya observados. Por ejemplo, con estos análisis se muestra que es posible construir (descubrir) el recorrido de cada línea de ómnibus aun sin contar con el mismo formalmente, y en caso de contar con el mismo, se puede usar la herramienta

para contrastar los resultados. De estos análisis también se demuestra que es posible estimar la velocidad media de los ómnibus en cada tramo, haciendo uso de las herramientas de minería de procesos.

Los ascensos multitramos también pudieron ser analizados con minería de procesos. En particular, para 8 de Octubre, se presentó un método de identificación de multitramos utilizando minería de procesos y se analizaron los multitramos, donde se distinguen cuáles son las combinaciones de barrios frecuentes, donde además también se concluyó que la mayoría comienzan fuera del corredor.

Los modelos generados carecen de información geográfica, por lo que se implementó un prototipo para convertir estos en capas geográficas que pueden ser visualizadas en un GIS. Si bien puede ser costoso en tiempo visualizar los modelos en un GIS debido a la fuerte interacción manual que se requiere, esta provee de un contexto geográfico que puede ser útil para un analista. Además, como se menciona en la sección 6.2, existe la posibilidad de que la interacción manual pueda ser automatizada con distintas herramientas.

Por otro lado, se presentó un algoritmo para la identificación de ómnibus para el caso en que los datos disponibles no tienen esta información, como es el caso de los datos disponibles públicamente. Se evaluó la confiabilidad del algoritmo contrastando contra datos reales y también se propuso un método para evaluar la confiabilidad de los resultados cuando no se poseen datos reales. Se presentaron algunos problemas y posibles mejoras del algoritmo.

Por último, se desarrolló un algoritmo para identificación de destinos de viajes y generación de matrices origen-destino basándose en el trabajo de (Massobrio, 2018). Se compararon las matrices origen-destino presentadas en dicho trabajo contra las generadas para 2022, analizando la variación en los desplazamientos de los usuarios entre municipios, observando como la variación fue mínima, teniendo en cuenta además que hay 7 años de diferencia entre ambos estudios. Se detallaron varios problemas del método y posibles mejoras propuestas en la literatura.

6.2. Trabajo a futuro

Diferentes análisis e investigaciones han quedado por fuera del alcance del proyecto, con proyección a ser retomadas en futuros trabajos. Estos han surgido en paralelo a la realización del proyecto, habiendo quedado por fuera del alcance del mismo, en mayor parte por limitaciones de tiempo.

En cuanto al análisis de performance realizado para los recorridos, se plantea la posibilidad de mejoras con base en datos de GPS de los ómnibus. Esto permite tener más exactitud para analizar tiempos de demora y cuellos de botella. Además, con los datos de GPS y herramientas de minería de procesos como Disco, es posible generar una visualización (simplificada) y con animaciones del flujo real de los ómnibus.

El prototipo realizado para convertir los modelos de recorridos generados a mapas, pretende mostrar la viabilidad de implementación, aunque aún queda mucho trabajo, en especial en su generalización a otros procesos. El mismo

utiliza las paradas de los recorridos como actividades del proceso, lo que plantea el desafío de generalizar a diferentes procesos, es decir, considerar otros atributos como actividades del proceso.

En línea con lo anterior, se podría considerar la automatización de visualizaciones. Si se determina cuáles son los mapas/visualizaciones que son de mayor utilidad, sería interesante analizar la posibilidad de generar estos de forma automática. Para esto es necesario analizar cuáles son las herramientas que proveen APIs para ejecutar la etapa de descubrimiento de modelos de forma automática para luego implementar la transformación a visualizaciones en mapas. También surge la posibilidad de generar visualizaciones a medida que se generan los logs, de manera de proveer información de los procesos en tiempo real.

Referencias

Agard, B., Morency, C., y Trépanier, M. (2006). MINING PUBLIC TRANSPORT USER BEHAVIOUR FROM SMART CARD DATA. *IFAC Proceedings Volumes*, *39*(3), 399–404.

Alsger, A. A., Mesbah, M., Ferreira, L., y Safi, H. (2015, enero). Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record: Journal of the Transportation Research Board*, *2535*(1), 88–96.

Augusto, A., Conforti, R., Dumas, M., Rosa, M. L., Maggi, F. M., Marrella, A., ... Soo, A. (2019, abril). Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, *31*(4), 686–705.

Bădică, A., Bădică, C., Buligiu, I., y Ciora, L.-I. (2022). Exploring the usability of process mining in smart city. *IFAC-PapersOnLine*, *55*(11), 42–47.

Barry, J. J., Newhouser, R., Rahbee, A., y Sayeda, S. (2002, enero). Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, *1817*(1), 183–187.

Behkamal, B., Pourmasoumi, A., Rastaghi, M. A., Kahani, M., Motahari-Nezhad, H. R., Allahbakhsh, M., y Najafi, I. (2022). Geo-enabled business process modeling. *18th International Conference on Business Process Management (BPM 2020), Sevilla, Spain, September 13 - 18, 2020*.

Berlingerio, M., Calabrese, F., Lorenzo, G. D., Nair, R., Pinelli, F., y Sbodio, M. L. (2013). AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. En *Advanced information systems engineering* (pp. 663–666). Springer Berlin Heidelberg.

Delgado, A., y Calegari, D. (2022). *Curso de Taller de Gestión y Tecnologías de Procesos de Negocio (TPM).* https://eva.fing.edu.uy/course/view.php?id=1572. (Accessed: 2023-02-06)

Diamantini, C., Genga, L., Marozzo, F., Potena, D., y Trunfio, P. (2017, agosto). Discovering mobility patterns of instagram users through process mining techniques. En *2017 IEEE international conference on information reuse and integration (IRI)* (pp. 485–492). IEEE.

Farzin, J. M. (2008, enero). Constructing an automated bus origin–destination matrix using farecard and global positioning system data in são paulo,

brazil. *Transportation Research Record: Journal of the Transportation Research Board*, *2072*(1), 30–37.

Foell, S., Kortuem, G., Rawassizadeh, R., Phithakkitnukoon, S., Veloso, M., y Bento, C. (2013, septiembre). Mining temporal patterns of transport behaviour for predicting future transport usage. En *Proceedings of the 2013 ACM conference on pervasive and ubiquitous computing adjunct publication* (pp. 1239–1248). ACM.

Günther, C. W., y van der Aalst, W. M. P. (2007). Fuzzy mining – adaptive process simplification based on multi-perspective metrics. En *Lecture notes in computer science* (pp. 328–343). Springer Berlin Heidelberg.

Jans, M., van der Werf, J. M., Lybaert, N., y Vanhoof, K. (2011, septiembre). A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*, *38*(10), 13351–13359.

Kurauchi, F., y Schmoecker, J.-D. (Eds.). (2021). *Public transport planning with smart card data*. London, England: CRC Press.

Li, T., Sun, D., Jing, P., y Yang, K. (2018, enero). Smart card data mining of public transport destination: A literature review. *Information*, *9*(1), 18.

Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M. P., y Bakker, P. J. M. (2008). Application of process mining in healthcare – a case study in a dutch hospital. En *Biomedical engineering systems and technologies* (pp. 425–438). Springer Berlin Heidelberg.

Massobrio, R. (2018). Urban mobility data analysis in Montevideo, Uruguay [Manual de software informático]. Montevideo, Uruguay.

Mauttone, A., y Hernández, D. (2017). Encuesta de movilidad del área metropolitana de Montevideo. principales resultados e indicadores (report) [Manual de software informático]. http://scioteca.caf.com/handle/123456789/1078. (Accessed: 2023-02-18)

Munizaga, M., Devillaine, F., Navarrete, C., y Silva, D. (2014, julio). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, *44*, 70–79.

Munizaga, M. A., y Palma, C. (2012, octubre). 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.

Nassir, N., Khani, A., Lee, S. G., Noh, H., y Hickman, M. (2011, enero). Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record: Journal of the Transportation Research Board*, *2263*(1), 140–150.

Pelletier, M.-P., Trépanier, M., y Morency, C. (2011, agosto). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568.

Sussman, J. M. (2005). *Perspectives on intelligent transportation systems (its)*. Springer Science+Business Media.

van der Aalst, W. M. P. (2016). *Process mining* (2.ª ed.). Berlin, Germany: Springer.

van der Aalst, W. M. P. (2019). A practitioner's guide to process mining: Limitations of the directly-follows graph. *Procedia Computer Science*, *164*, 321–328.

van Eck, M. L., Lu, X., Leemans, S. J. J., y van der Aalst, W. M. P. (2015). PM^2: A process mining project methodology. En *Advanced information systems engineering* (pp. 297–313). Springer International Publishing.

Wang, W. L., Lo, S. M., y Liu, S. B. (2015, septiembre). Aggregated metro trip patterns in urban areas of hong kong: Evidence from automatic fare collection records. *Journal of Urban Planning and Development*, *141*(3).

Zhao, J., Rahbee, A., y Wilson, N. H. M. (2007, julio). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5), 376–387.

Zhao, K., Tarkoma, S., Liu, S., y Vo, H. (2016, diciembre). Urban human mobility data mining: An overview. En *2016 IEEE international conference on big data (big data)* (pp. 1911–1920). IEEE.

Glosario

Línea: Nombre público con el que se conoce a un conjunto de recorridos de una empresa de Transporte Colectivo. Por ejemplo: 2, 103, 306, 405, 538, D11.

Sub-línea: Cada uno de los trayectos que tiene una línea, que implique desplazamiento por calles diferentes, bajo un mismo nombre público. Por ejemplo: la línea 409 tiene 2 sub-líneas, una es la que termina en Aviación Civil y otra la que termina en el Hospital Saint Bois. Para considerar sub-líneas diferentes, debe haber diferencia en las calles que se transitan. Cuando un ómnibus termina "cortado", sin llegar a su destino, o cuando sale de un punto diferente del origen, pero sin que haya variación en las calles que se recorren, no se considera como sub-línea diferente, sino una "variante" (se define más abajo).

Sentido: Cada una de las dos posibilidades de recorrer una sub-línea. Por ejemplo: la sub-línea del 103 que va al km. 21 tiene dos sentidos: Aduana – Km. 21 y Km.21- Aduana.

Variante: Cada instancia del recorrido de una sub-línea, con un origen, destino y sentido determinado. Sentido de una variante resultará DESCENDENTE cuando la latitud del origen (tomada en valor absoluto –sin signo) sea MENOR que la latitud del Destino. De lo contrario, si es MAYOR será ASCENDENTE.

Anexo IB1

Pseudocódigo - algoritmo identificación buses

El Algoritmo 1 presenta el pseudocódigo del algoritmo utilizado para la identificación de buses.

Algoritmo 1 Pseudocódigo del algoritmo de identificación de buses.

```
Entrada
LVV:  Lista de viajes de una variante ordenados por fecha.
    • [id_viaje, fecha, parada, ordinal_tramo, etc...]

LRHV: Lista de recorrido (orden de paradas) y horarios estipulados de pasadas por parada de una
variante según tipo de día (hábil, sábado, domingo/feriado).
    • [tipo_dia, frecuencia, parada, ordinal, hora, etc...]

Salida
LRES: Lista de viajes asignando a cada viaje una frecuencia (bus) y datos extra:
    • [id_viaje, frecuencia, parada, ordinal_parada, fecha, desfase, etc...]
    • desfase: Diferencia entre el horario estipulado de la frecuencia identificada por la parada y
      la hora en que se registra el viaje (ascenso al bus) del pasajero.

Pseudocódigo
lst_frecuencias = lista de frecuencias de la variante ordenadas por fecha de salida para el tipo de
día de la fecha analizada // obtenido de LRHV
lst_viajes_procesados = [] // lista de viajes con frecuencia asignada
dmax = 5   // máximo desfase permitido en cada iteración
dinc = 5   // incremento de desfase máximo permitido
dfin = 45  // utilizado como condición de parada, termina si dmax llega a dfin

while (cant_viajes_asignados < cant_viajes_total and dmax <= dfin)
    • while (not lst_frecuencias.emtpy())
        ○ f = lst_frecuencias.next()
        ○ ord_actual = 0
        ○ for (viaje v no procesado de LVV)
            ■ p = v.parada
            ■ o = p.ordinal // ordinal de la parada en el recorrido de la variante v
            ■ h = p.hora // hora estipulada de pasada de la frecuencia f por la parada p
            ■ d = v.hora - h // cálculo de desfase
            ■ if ((abs(d) <= dmax) and (o >= ord_actual))
                • if (no existe en LRES un viaje para f que quede inconsistente con v)
                  // coherencia entre ordinal de parada y fecha-hora del ascenso
                    ○ LRES.Add([v, f, p, o, d])
                    ○ lst_viajes_procesados.add(v)
                    ○ ord_actual = o
        ○ lst_frecuencias.remove(f)
    • dmax += dinc
```

Anexo IB2

Análisis de ejecución del algoritmo de inferencia de ómnibus sin los datos reales

Para probar y analizar las distintas versiones del algoritmo de identificación de ómnibus previo a la obtención de los datos de frecuencias de variante se tuvo en cuenta la gran cantidad de datos que existían y la complejidad de ejecutar los algoritmos sobre estos. Los distintos algoritmos se ejecutaron sobre un subconjunto muy reducido de los datos:

- para que el tiempo de ejecución del algoritmo sea breve y con esto poder ejecutar y analizar más veces.
- debido a que el algoritmo no depende del subconjunto de datos utilizado.
- para que sea más fácil analizar los datos.
- para tener certeza que los datos de los horarios de las frecuencias y sus paradas sean los correctos para los datos analizados.

Es por esto que se tomó la decisión de analizar los datos de un día completo (comenzando antes de la primera frecuencia y terminando luego del fin de la última frecuencia) de una variante en particular. Con esta simplificación el algoritmo no pierde demasiada generalidad y se obtienen las ventajas antes mencionadas. Además, al estar trabajando con una variante en particular, es más fácil el análisis, ya que uno comienza a tomar contacto permanentemente con esa variante y puede identificar problemas en el algoritmo y/o datos con mayor facilidad.

A medida que se obtenían los resultados del algoritmo se identificaba problemas en los datos. Uno de los problemas más frecuentes era que había ciertos eventos (subida de un pasajero) que no eran asignadas a ningún ómnibus. Toma mucho tiempo analizar el por qué ciertas veces el algoritmo no hace la asignación a un bus para todos los eventos no asignados, se analizaron algunos de estos y se volvió a mejorar el algoritmo. La versión del algoritmo que dio mejores resultados, no estaba exento de este problema, seguían habiendo eventos que no eran asignados a un ómnibus en particular, sin embargo, la cantidad de estos casos era menor y había mejoras con respecto a las otras versiones. En particular, tomando las consideraciones antes mencionadas (fijando variante y día) el

algoritmo que tuvo mejores resultados, tuvo un porcentaje del 2,24 % promedio de eventos sin asignar con un máximo de 3,8 % y un mínimo de 1,2 %. Sabiendo que el volumen de datos analizados por el algoritmo es grande, este porcentaje no hará un cambio considerable en el modelo final, por lo que se toma como un buen punto de partida para comenzar a realizar descubrimiento de modelos.

Otro problema que se tuvo y una de las causas por el cual se tomó un subconjunto tan reducido de los datos, era que los datos de los horarios de las frecuencias y las paradas que se tenían hasta el momento eran los más actualizados, de marzo 2022. Esto hacía que si se quería probar el algoritmo, por ejemplo, en agosto 2021, era probable que los horarios de las frecuencias y paradas en ese entonces hubiera sido distinto y, por tanto, el resultado del algoritmo no sea el esperado.

Se mencionó antes, que para analizar el algoritmo, se tomaba en cuenta la cantidad de eventos sin asignar a un ómnibus. Esto es una medida de calidad del algoritmo, mientras baja este valor se puede determinar que el algoritmo mejora. Sin embargo, esta métrica podría no haber sido la única. Para tener certeza de la precisión del algoritmo se tienen que tomar en cuenta otros valores. Por ejemplo, otras métricas que se podrían haber considerado para analizar globalmente el algoritmo, son las siguientes:

- CONFIABILIDAD = 1 - (cantidad de eventos que fueron asignadas a más de una frecuencia/cantidad de eventos).
- M = cantidad media de veces que fue asignado un evento a un ómnibus

Por ejemplo, si CONFIABILIDAD es 0.5 el algoritmo la mitad de las veces no sabría determinar a qué frecuencia realmente corresponde un evento. Esta métrica mide que tan confiable es el algoritmo. En relación con la anterior, si M es cercano a 1 indica que el algoritmo en general hace un match 1 a 1 de eventos y ómnibus.

Se podrían considerar otras métricas, y tomando aquellas que tienen más sentido para el problema, se podría evaluar el algoritmo en su totalidad. A partir de esto, se podría tener un valor general de calidad del algoritmo (o contabilizar cada métrica por separado), haciendo más sencillo determinar cuando el algoritmo es correcto y de este modo no seguir intentando mejorarlo.