Process Mining for improving urban mobility in smart cities: challenges and application with open data

Andrea Delgado, Daniel Calegari
Instituto de Computación, Facultad de Ingeniería,
Universidad de la República, Uruguay
{adelgado, dcalegar}@fing.edu.uy

Abstract

Urban mobility presents various challenges to favor urban development. These challenges have been traditionally analyzed using transport network optimization and simulation techniques. Nevertheless, it is possible to think of process mining as a complementary approach allowing, among other things, to discover behavioral transportation models, obtain execution measures and detect bottlenecks. The objective of this article is to analyze how suitable PM is for the analysis of urban mobility problems. We use open data from the Metropolitan Transportation System (STM) of Montevideo, Uruguay, which, among other things, provides the ability to record up-to-date information on its transportation network and trips of its citizens. We apply process mining to process discovery, both from buses and their users, and carry out various analyses linking such data with time information, costs, types of users, and city areas.

Keywords: urban mobility, process mining, open data

1. Introduction

Urban mobility is crucial to the functioning of our society, allowing us to access housing, jobs, and urban services. Transport modeling and planning (de Dios Ortúzar and Willumsen, 2011) and studying citizens' needs allow for addressing several challenges, such as traffic congestion, public transport crowding, pedestrian difficulties, and atmospheric pollution.

Smart cities (Deakin and Waer, 2011) use technology to improve urban services. In such context, smart mobility (Benevolo et al., 2016) addresses urban traffic and transportation, e.g., by using Intelligent Transportation Systems (ITS) (Sussman, 2005) that incorporates technologies generating real-time data that can be processed to extract valuable information.

In 2010, the Municipality of Montevideo, Uruguay, defined a Mobility Plan[1] transforming the transportation system to raise the quality of life and economic and social development of the city. It defines the Metropolitan Transportation System (Sistema de Transporte Metropolitano, STM). The STM defines preferential corridors for bus lines with a reduction in the number of transversal crossings and coordination of traffic lights; terminals, and interchanges; as well as an ITS with on-board GPS unit control systems on each bus, smart cards for citizens, and traffic cameras all over the city. The Mobility Management Center[2]) manages and controls traffic and transportation in the city, using the real-time data provided by the STM.

Urban mobility has been traditionally studied by employing transport network optimization, simulation (de Dios Ortúzar and Willumsen, 2011), and data analysis (Massobrio and Nesmachnow, 2019) techniques. They provide a solid background for decision making, e.g., calculating the number of trips emanating from a zone and providing an Origin Destination matrix calculating the trips made from an Origin to a Destination during a particular period. Nevertheless, many of these techniques only offer static information and visualization.

Process Mining (PM) (van der Aalst, 2016) provides means for analyzing runtime events from information systems to discover corresponding business process (BPs) models (process discovery). Also, it allows verifying the compliance of the enacted BPs concerning the expected one (conformance checking) and analyzing key execution measures, e.g., resources used and time duration. A process perspective provides dynamic information and visualization of a given context, e.g., how a specific process case evolves. In this sense, we can think about essaying the application of PM techniques for analyzing urban mobility in which some aspects are conceptualized as processes, e.g., the bus line's route.

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

[2] Mobility Management Center. https://montevideo.gub.uy/centro-de-gestion-de-movilidad

URI: https://hdl.handle.net/10125/102846
978-0-9981331-6-4

As far as we know, very few works apply PM to urban mobility, e.g., for discovering user mobility patterns from georeferenced information of Instagram profiles (Diamantini et al., 2017). Also, professional initiatives describe the application of PM to various mobility problems (Mobility as a Service (MaaS), 2016) without delving into details. Thus, a deeper analysis of the PM application could contribute to opening new research opportunities.

In this work, we perform a preliminary experience trying to answer the following research question: how suitable is PM for analyzing urban mobility? We address a set of questions defined by the Municipality of Montevideo in the context of a joint research project concerning their transportation system. We first analyze theoretically how PM could be used to answer such questions. We then perform an initial PM experience focusing on some of the questions. For this, we use open data provided by the STM within the governmental open data catalog[3]. In the long-term, we want to assess to which extent PM is a complementary approach for improving expert analysis devised for decision making.

The rest of the paper is structured as follows. In Section 2, we summarize related work. In Section 3, we introduce the context and the available open data. In Section 4, we define analysis goals and theoretically analyze how PM could tackle them. In Section 5, we describe the PM experience. Finally, in Section 7, we conclude and provide an outline of future work.

2. Related work

The mobility of people and goods has been studied for a long time (de Dios Ortúzar and Willumsen, 2011). Strategies have been enhanced by the availability of real-time data provided by the transportation systems (Benevolo et al., 2016), e.g., employing Intelligent Transportation Systems (ITS) (Sussman, 2005).

One of the main research topics within urban areas is the discovery of mobility patterns, i.e., understanding the spatio-temporal characteristics of human mobility, which is valuable for traffic management and urban planning. In general, this problem was addressed from the perspective of data analysis, collecting trip data from different sources, e.g., citizens' cellphones (Lorenzo et al., 2016), a bike sharing system (Purnama et al., 2015), and taxi GPS data (Liu et al., 2021). The main difference with studying mobility patterns in public transportation is that it has fixed routes, so mobility is restricted to existing lines and not open spaces. However, urban mobility needs to consider every transportation system, and public transportation benefits from this information to redefine its lines according to explicit citizens' needs.

When focusing on public transportation, we found data analysis works exploiting smart card data from citizens. In (Cui et al., 2015), the authors collect data from the automated fare collection system of Beijing, China, to analyze the evolution of extreme transit behaviors of travelers, i.e., travelers with unusual behavior categorized as early birds, night owls, tireless itinerants, and recurring itinerants. In (Han et al., 2010), the authors use data from smart cards in Changchun, China, to calculate the passenger flow time distribution of different groups of passengers, among other aspects. In (Branda et al., 2020), the authors use data collected from a bus ticketing platform in Italy to discover the factors influencing travelers in booking and purchasing a bus ticket. In (Massobrio and Nesmachnow, 2019), the authors combine several sources of urban data from the public transportation system of Montevideo, Uruguay, to describe mobility patterns in the city.

We conducted a literature review to refine the search for related work focused on PM and mobility. We searched within electronic databases: ACM, Google Scholar, IEEE, ScienceDirect, SCOPUS, and Springer. We use the following search string: `'process mining' AND ('urban mobility' OR mobility OR transport OR transit).`

Works describe PM in the domain of mobility, e.g., to the optimization of customer experience operative processes in Uber Rowlson, 2020), and the assistance to non-professionals to perform visual trend analysis Burkhardt et al., 2020. However, very few works focus on applying PM to mobility problems. In (Diamantini et al., 2017), the authors discover user mobility patterns using georeferenced social media data (Instagram posts) published by visitors of an exposition. In (Yousfi and Weske, 2019), the authors use smartphone sensors to collect users' data and apply PM to discover commute patterns and determine the users' significant locations, routes, commuting times, and travel modes. In (Mobility as a Service (MaaS), 2016), the authors describe the application of PM to various problems, such as identifying optimal combinations using different transportation systems, simulating changes in travel patterns, and optimizing bus routes. However, there are no details on how they can be addressed.

3. Research context & STM open data

The Municipality of Montevideo made an open call to seek answers to mobility questions defined by business experts, among others:

[3] Open Data Catalog. https://catalogodatos.gub.uy/organization/intendencia-montevideo

1. What is the route of a bus line?
2. How is the mobility of people within the STM?
3. How is the mobility of people post-COVID?
4. Are there delays in the main corridors of the city of Montevideo?
5. How does a traffic detour affect public transport?
6. What is the impact of weather conditions on public transportation?
7. What would be the impact of the exclusion of private vehicles on 18 de Julio street?
8. What would be the impact on traffic of pedestrianization in the old city quartier?

We proposed to study bus routes and passengers' behavior from the PM perspective using open data. The following sections provide initial insights on how PM could answer these questions. Nevertheless, concrete solutions to mobility problems probably also require using traditional techniques and a deeper analysis of business experts to define changes in the transportation system, which is out of the scope of this paper.

3.1. Domain description

In the STM, a line is a public name by which a set of routes of a transportation company is known, e.g., 2, 103, 306, 405, 538, D11. The sub-lines are each one of the routes that a line has, which implies traveling through different streets under the same public name, e.g., line 409 has two sub-lines, one that ends at "Civil Aviation" stop and another one that ends at the "Saint Bois Hospital" stop. Different sub-lines imply that there is a difference in the streets traveled. When a bus ends "cutting" the route without reaching its destination, or when it leaves from a different point of origin, but without variation in the streets that are traveled, it is not considered a different sub-line but a variant (defined below). The direction of the sub-line is each one of the two possibilities of traveling a sub-line, e.g., the 103 sub-line that travels the route Customs - km. 21 of the City, has two directions: i) Customs – Km. 21, and ii) Km.21- Customs. The direction will be descending when the latitude of the origin is less than the latitude of the destination. Otherwise, it will be ascending.

A variant is each instance of the route of a sub-line, with a specific origin, destination, and direction, e.g., line 103 with its sub-line at km. 21 with Customs origin and destination km. 21 (which determines the direction). There are three types of variants: the maximal variant, which is the longest route; the non-maximal variant, which is a partial route of a sub-line; and the circular variant that operates continuously, i.e., when it reaches the destination of the first variant, it begins the route of the second variant, and the passenger continues the trip without noticing this change (and with the same ticket). Every variant (line, sub-line, direction, origin, and destination) has a set of frequencies that determine a departure time on a given day (working days, Saturdays, Sundays, and holidays). In addition, the stops in the route of each variant are known, as well as the stop of origin and destination, all identified by a unique code. The theoretical schedule by which a bus travels through each stop is also known for each frequency. Apart from the bus data, the information on STM card payments of passengers is also known. There are different types of passengers, e.g., common users, retirees, and students. Every time a payment is recorded, there is a registry with the kind of passenger, the stop at which it boards, the variant and frequency to which the bus belongs, and the time of departure, among other information.

3.2. Available datasets

Based on a preliminary analysis of the problems of interest and the available open data from the Metropolitan Transportation System (Sistema de Transporte Metropolitano, STM), we identified as relevant the following datasets. As an example of the data included, we show some items for the first two datasets, which are the ones we used in the process mining application we present in Section 5.

[UBS] Urban bus schedules, by stop[4]. It contains the bus schedules for urban transportation. These are estimated theoretical schedules in which a bus line will pass through a particular stop along its route. These data are obtained by estimating each stop's passing times according to the transport units' average speed, predefined schedule, and distance between stops.

- type of day: 1 - working days (Monday to Friday), 2- Saturdays, and 3 - Sundays and holidays.
- cod_variant: identifies the variant of the line.
- frequency: identifies the frequency for the variant for the type of day, i.e., the specific hours.
- cod_loc_stop: shows the code of each stop
- ordinal: shows the number of stops within the bus line route, i.e., the first (1), the second (2), etc.

[4] Urban bus schedules, by stop. https://catalogodatos.gub.uy/dataset/intendencia-montevideo-horarios-omnibus-urbanos-por-parada-stm

- hour: shows the estimated hour for the bus to arrive at the stop
- day_before: indicates if the frequency started the day before (for late night lines)

[TSB] Trips made on STM buses[5]. It contains all the trips made on the urban collective transport lines in Montevideo by each operating company, line, variant, day and time. Upgrades at all stops in the system, by type of user, payment method, and sections of each trip. The information provided comes from all the records processed by the STM trip validation machines.

- id_trip: identifies the trip within the system; in combined tickets, the id_trip is repeated in all buses the user rides while the ticket is valid.
- line_code: shows the code of the bus line
- cod_variant: identifies the variant of the line
- frequency: identifies the frequency for the variant
- with_card: shows if the STM card was used in the payment of the trip
- date_event: timestamp of date and hour of the trip
- trip_type: combined (1 hour, 2 hours), student, retired, common user.
- origin_stop_code: shows the code of the stop at which the user got on the bus

[BLOD] Bus lines, origin, and destination[6]. It contains geographic information that, among other data, includes the lines' description, origin, and destination.

[CTSC] Buses: stops and checkpoints[7]. It contains geographic information of stops and control points.

[USHG] Usage Statistics: How To Go[8]. It contains usage statistics for the "Cómo ir" application, including routing queries (from one point in the City to another) and schedule queries (departing or arriving at a specific time) made anonymously by citizens.

[ODS] Origin-Destination Survey[9]. It contains the data of the origin-destination surveys carried out in 2009 and 2016.

In addition to the above datasets, there are others available that may be of interest for a more extensive urban mobility analysis but are not as closely related to the basic questions of interest, e.g., transit routes, limits of Zonal Community Centers, population by census area, vehicle counting in the main avenues, average vehicle speed on the main avenues, issued traffic tickets, and traffic accidents. Moreover, they are web applications providing static and dynamic real-time data, e.g., a real-time map of buses[10], as the one depicted in Figure 1, and a query for checking bus schedules [11].

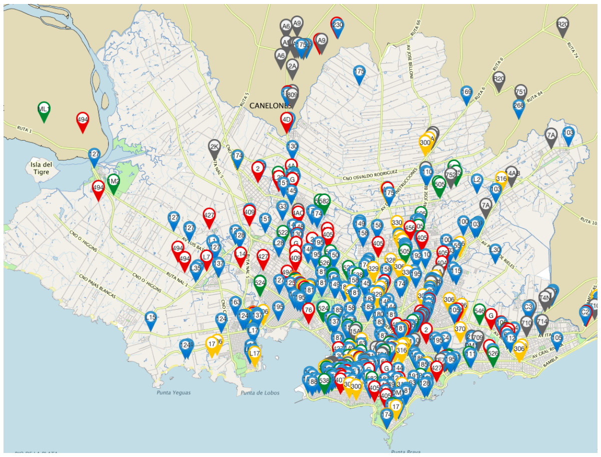

Figure 1. Real-time information of buses (Sunday 06-12-22, 11AM)

4. Theoretical analysis

We theoretically analyze how the available open data could be used from the PM perspective and how to address the questions described in Section 3.

4.1. General analysis

Focusing on bus routes and passengers' behavior, we can perform different actions linked to the three main aspects of PM.

Process discovery

Using bus frequencies (UBS dataset), we could derive a reference model defining the path of each

[5] Trips made on STM buses. https://catalogodatos.gub.uy/dataset/intendencia-montevideo-viajes-realizados-en-los-omnibus-del-stm

[6] Bus lines, origin, and destination. https://catalogodatos.gub.uy/dataset/intendencia-montevideo-lineas-omnibus-origen-y-destino

[7] Buses: stops and checkpoints https://catalogodatos.gub.uy/dataset/intendencia-montevideo-transporte-colectivo-paradas-y-puntos-de-control

[8] Usage Statistics: How To Go. https://catalogodatos.gub.uy/dataset/intendencia-montevideo-estadisticas-de-uso-comoir

[9] Origin-Destination Survey. https://catalogodatos.gub.uy/dataset/intendencia-montevideo-encuesta-origen-destino-montevideo

[10] Bus Map https://www.montevideo.gub.uy/buses/mapaBuses.html

[11] Bus schedule. https://www.montevideo.gub.uy/app/stm/horarios/

bus line. This reference model represents each stop as a node of the model and the connections between stops (through the bus lines) as transitions between those nodes. In addition, different execution traces are available on this model: the bus routes and the potential and actual trips of the users.

Using users' data (TSB, USHG, and ODS datasets), it could be possible to discover the same model or a subset of it to the extent that specific paths are not reflected, as well as combinations made between different lines.

Using usage statistics (USHG dataset), it could be possible to derive a model of queries made by users to see the type of usual route queried.

From the survey (ODS dataset), it could be possible to derive a model of travel, which could be compared with the actual trips of both bus lines and users.

If zones (or sets of lines of interest) can be identified, it could be possible to reduce the model to trips made only between those zones (or line routes) by zooming in on a portion of the routes. The reduction of reference models simplifies the analysis of waiting times and blockages in peak hours, among others.

Taking a specific stop (or a set), it could be possible to discover a model of departures/arrivals and analyze the times and load of the stop, recognizing as events the arrival and departure of lines from that stop.

Conformance checking

Using buses and users data (UBS, TSB, and USHG datasets), it could be possible to identify deviations against the reference model. With models discovered using data from the survey (ODS dataset) or usage statistics (USHG dataset), the check can be made with the actual data of the card (TSB dataset) to see if the trips coincide with what is expected.

It could also be possible to use the data on the effective trips of a bus and the users' trips (TSB dataset) to replay the traces on the model of stops, that is, graphically visualize the events that occur on the model (e.g., the trips of the buses over time).

Moreover, it could also be possible to compare the changes in pre-covid and post-covid mobility patterns, both in the number of trips and in the form of them, e.g., routes made by the same people and whether bus combinations are maintained or changed.

Model extension

It could be possible to display waiting and service times, which can be viewed by replaying traces for specific zones or stops. Bottlenecks could also be detected if analyzed for specific stops or sets of stops. Also, detecting more usual combination stops between lines could be possible. Finally, daily, weekly and seasonal patterns could be spotted since, in any of the compliance check cases, the time slots can be reduced to specific time slots or special dates, and the checks could be done per line, user, or group of users.

4.2. Goals analysis

By taking the above points as a reference, it is possible to analyze the questions expressed by business experts, described in Section 3.

1. What is the route of a bus line? As mentioned in Section 4.1, it could be possible to discover the reference model of a bus line that allows visualizing the line's route according to the defined stops.

2. How is the mobility of people within the STM? As also mentioned in Section 4.1, it could be possible to exploit users' data to discover a travel model for a line filtering data by days (business days, Saturdays, Sundays, and holidays), seasons, type of user, company, etc. After filtering and discovering the model, it could be possible to analyze load metrics. For example, it could be possible to answer: At which stops do people not get on? At which stops do more people get on? At what times does the bus have more and fewer passengers? What happens to those lines in a particular area? In the same way, it is possible to answer performance questions such as What is the total duration of the frequencies (on average, maximum, minimum)? Are there frequencies with delays?

3. Are there delays in the main corridors of the city of Montevideo? The STM defines transportation corridors[12] which have exclusive lanes for buses. By identifying and filtering bus lines that traverse a transportation corridor between two stops for zooming in on the trips between them, it could be possible to visualize the load of the traffic network. As mentioned before, this could allow calculating average times, worst case, etc.

4. How is the mobility of people post-COVID? Based on the discovery of a pre-covid and post-covid model (and during covid), it could be possible to compare usage metrics concerning lines, variants, and frequencies. Also, combined with data mining techniques, it could be possible to identify clusters of behaviors by bus lines, areas, or types of users.

[12] Main corridors. https://montevideo.gub.uy/areas-tematicas/movilidad/transito/corredores

5. How does a traffic detour affect public transport? In the case of identifying a period and area in which a deviation occurred, it could be possible to perform a conformance check between the trips made during that period and the reference model to detect deviations. It could even be possible to simulate alternative routes based on the existing connections between stops, assuming cuts in the traffic network.

6. What is the impact of weather conditions on public transportation? As mentioned before, it could be possible to visualize the load of the traffic network and analyze many performance metrics by reducing the study to dates where there is a meteorological event or between seasons.

7. What would be the impact of the exclusion of private vehicles on 18 de Julio street? Including information on the current circulation of cars on 18 de julio (the main street in the city center), the impact of moving those trips to the buses that circulate through the center could be analyzed similarly to detours.

8. What would be the impact on traffic of pedestrianization in the old city quartier? It could be done similarly to the study of detours, focused on the mobility of buses, since we are still not considering information on the traffic of private cars.

5. Process Mining application

We now present a PM experience focusing on exploring from a practical perspective some of the business expert questions expressed in Section 3.

5.1. Data analysis

We selected a sample of bus lines to analyze, covering different geographical zones and neighborhoods of the city, not only focusing on the concentration of buses traveling towards the East coast, the city center, and the old town, as shown in Figure 1. We selected seventeen (17) bus lines based on characteristics such as the zones of the city it covers (i.e., going from the city center, old city, and East coast to the West, North, and East as destination), neighborhoods it traverses, length of the bus line, and traveling through at least one of the transport corridors of the city (e.g., Av. 18 de Julio, Av. 8 de Octubre, Br. Batlle y Ordoñez). In Figure 2 we show the coverage of the lines we selected. Table 1 presents descriptive data of the selected lines.

Two of the selected lines are the top two in STM trips of all lines in the May 2022 file, the 103 and G lines, which together account for 1.5 million trips. Altogether, the selected lines cover around 30% of the total trips registered for the month.

As mining/analysis objectives, we selected some of the questions expressed by business experts, described in Section 3, that could be directly tackled using the open data already published in the STM catalog. In particular, we focus on the files identified as USB and TSB, which we complement with data in BLOD and CTSC allow us to answer questions 1 to 3.

Regarding the quality of the data, we found some issues, such as the lack of reference to the specific frequency of the bus in the STM trips (USB file), which prevents us from performing a direct analysis on cases over the variants of the buses. As this data is registered but not published, the municipality provided us with a new file for the STM trips for May 2022, which is the one we analyzed here. The file contains 25 million records, including all trips from the STM system registered within the 136 bus lines. Some frequencies are set to zero, which can be taken as any day. We added them in the analysis by using the day of the trip.

5.2. PM analysis

We inspected the event log using Disco's process mining tool with an academic license. We analyzed the cases in the log and the variants identified by the tool, i.e., different paths over the control flow. In the following, we present a summary of the analysis of the selected bus lines, exemplifying some results.

To illustrate how to proceed with discovering the reference model and the STM trips model, we selected line 125, which goes in one direction from the old city to the West, ending in Cerro beach, and returning in the other direction. This line has 12 variants from which we select the maximal in the first direction (i.e., from the old city to Cerro beach) identified as variant 667. It has 61 stops, and in May 2022, it made 1477 bus trips, i.e., frequencies for every day of the month in specific buses, registering a total of 82.606 trips from people.

Regarding the reference model of the bus lines (question 1), we can obtain it from the TSB file that contains the theoretical schedules. We select variants for each line, setting the variant and frequency codes as *case ID*, the stop code as *activity ID*, the hour of the stop as *timestamp*, and the rest of the data fields as *attributes* of the activity. Since variants have different directions corresponding to the sub-line, we obtain two sequential sections in each direction. As mentioned, we selected the maximal variant of each direction to serve as a reference model, and we present the discussion for

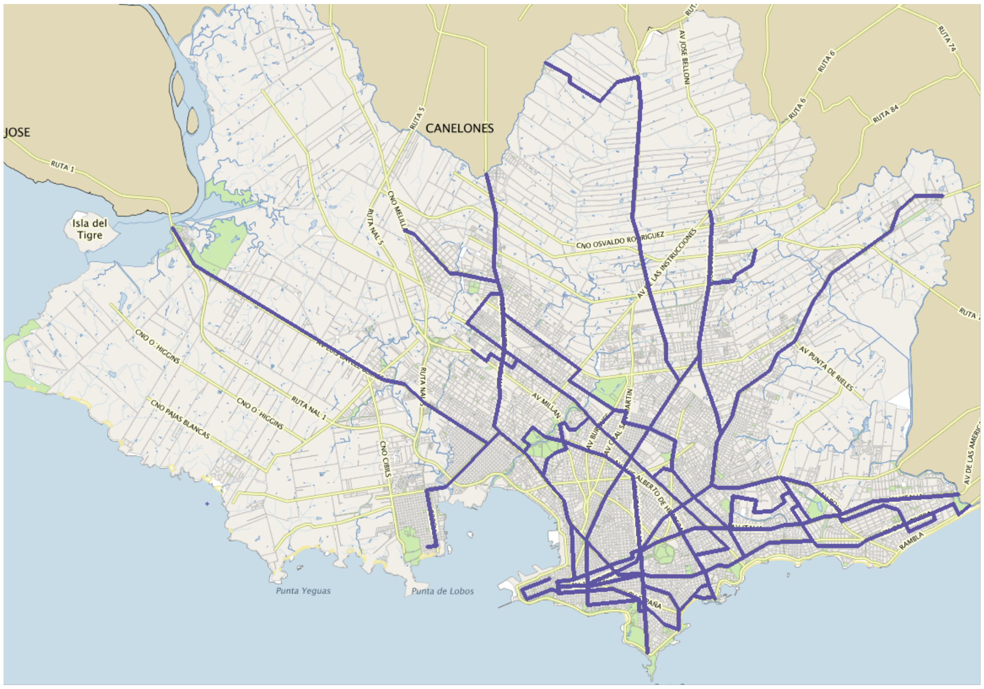

Figure 2. Lines coverage for the PM initiative

Table 1. Lines selected for the analysis, variants, frequencies, and STM trips

Line	Variants	Frequencies	STM trips
125	12	400	173868
127	53	524	323825
130	16	454	361570
G	14	459	682263
175	14	531	428140
505	8	321	382326
103	39	827	864030
105	29	494	350705
21	12	464	357193
370	13	307	483357
142	14	350	200025
144	7	436	292116
145	17	463	576901
522	5	248	259291
174	26	397	485190
181	9	785	571657
300	14	390	552313

the 667 variant here. The reference model must contain the same stops in the same order as the ones defined by the variant in the STM system, which is accessible from the website for a scheduled consultation. In Figure 3, we present the bus line 125 route with stops for variant 667, which is the maximal variant in one direction. Figure 3 (a) shows the bus line 125 route with stops for variant 667, with origin stop 4041 in the old city, and destination stop 1122 in the Cerro beach, with 61 stops covering the route from the origin to the destination. Figure 3 (b) shows an excerpt of the reference model obtained with process mining in disco for variant 667, with the first 11 stops of the route, starting with the origin one. As expected, the reference model is sequential since

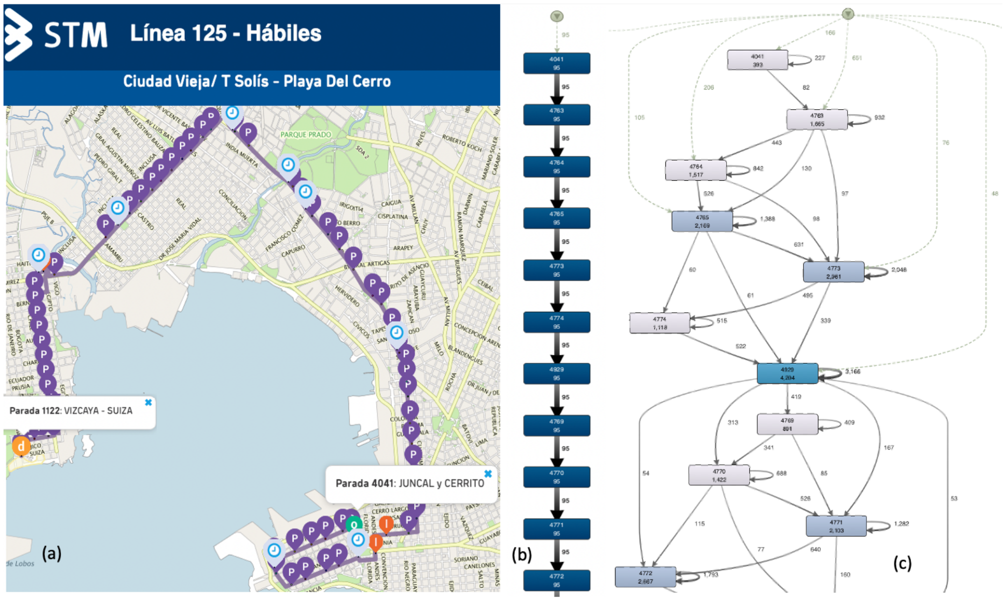
STM
Línea 125 - Hábiles
Ciudad Vieja/ T Solís - Playa Del Cerro
Parada 1122: VIZCAYA - SUIZA
Parada 4041: JUNCAL y CERRITO
(a)
(b)
(c)

Figure 3. Bus line 125 route with stops for the 667 variant: a) in the STM website Bus schedule, b) excerpt of the reference model, and c) excerpt of the STM trips within its frequencies

stops are traveled by frequency of the variant within the defined route and schedules.

For the STM trips analysis, i.e., the actual use of the bus lines by users in the city, we select the variant, frequency, and day of the trip as *case ID*, the stop code as *activity ID*, the timestamps of each trip of each stop as the *timestamp* for the activity, and the rest of the data fields as *attributes* of the activity. Figure 3 (c) shows the STM trips within the frequencies of variant 667. It is the actual bus trips of people within each frequency for each day in a physical bus. In this file, each person's trip is recorded with interest data, including at which stop the person boards the bus, as presented in section 3. It is shown in the model as loops within each stop, with the total number of people that boarded at that stop in all bus frequencies in May 2022. The model also shows cases where in some frequencies, no people are boarding the bus at the first two stops. Hence, the first stop registered is the one with the first boarding, depicted with the dotted lines that go from the init marking to other stops that are not the 4041 (e.g., 4763).

In both models, the intensity of the blue color in the stops (activities) represents the frequency of the activity within the cases, i.e., in how many cases the activity is executed (the more intense the blue, the more times it appears). In the model of STM trips in (c), the color represents how many bus trips have people boarding at that stop (i.e., records of people trips). In the reference model in (b), as expected, all stops are of intense blue since all frequencies have a schedule for the 61 stops of the variant. It can be seen in (c) that the origin stop 4041 has a shallow frequency, other stops such as 4765 and 4773 have a little more and 4929 has even more. Of all stops, 4925 is the one that has more frequency, a 10.83% which corresponds to the Cerro terminal, where there is a combination with other buses. The stop where almost no people board corresponds to the four last stops, which makes sense since the trip is finishing and people are getting off the bus.

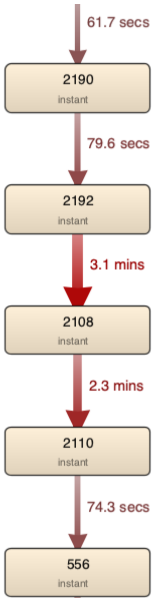

Figure 4. Bus line 300 worst performance

Regarding other questions raised in Section 3 for question 2 (people mobility), the median duration for the bus trip is 49.3 minutes, and cutting some very short trips with few stops, is around an hour, which is consistent with reality. The maximum duration registered is one hour and 20 minutes, and there are a few less than 10 minutes with people boarding only in a few stops that correspond to early morning hours. The type of trip most used is the 1-hour trip, which allows people to make a combination of buses within one hour, representing 46.15% of the registered trips, followed by student trips of type G and A, representing 14% and 9%, respectively.

As an example of analysis of question 3 (delays in corridors), we chose maximal variant 8398 of line 300, with origin "Instrucciones" and destination "Cementerio Central", with 69 stops covering the route from the origin to the destination. In May 2022, it made 2982 bus trips, registering 271.291 trips from people. This variant has the property of traversing five corridors (Av. Belloni, Av. 8 de Octubre, Av. L.A. de Herrera, Av. Italia and Br. Artigas). When looking for performance metrics on paths, we found that the mean duration from going from one stop to another is about 30 to 90 seconds on average. However, there are some regions where this mean duration increases. As shown in Figure 4, the worst results are between stops 2192 and 2110. In particular, stop 2108 is in the middle of a corridor change (from Av. Italia to Br. Artigas), and corresponds to the main bus terminal (Tres Cruces), i.e., one of the most crowded areas in the city. The second worst result on this variant is also a corridor change from Av. Belloni to Av. 8 de Octubre, corresponding to a bus interchange terminal, so the bus makes a special stop, possibly with a slightly longer waiting time than other stops.

6. Limitations and evaluation

In this work, we aimed to answer the research question presented in Section 1: how suitable is PM for analyzing urban mobility? We gave both a theoretical and practical analysis, showing an application of process mining to open data of the Municipality of Montevideo, also guided by business experts' questions regarding the city mobility of interest for the Municipality. Although it was the first experience, we have obtained promising results in that PM allows providing valuable information adding a process perspective to existing data analysis.

We showed how using the open datasets provided in the STM catalog, process models of the bus lines routes and trips can be discovered and analyzed, providing a different view of the data based on processes. We showed that the bus lines route could be constructed from the bus line schedules, presenting which data fields are used as key elements of the event log for PM to allow discovery of the reference process model. We also presented the data fields correspondence from the STM actual trips dataset, which allowed us to discover and show the specific behavior of the bus line trips for each variant and frequency of each day, based on how people use the buses in the city. For this, we selected some bus lines to try to cover different city zones.

In this way, information regarding, for example, the use of the stops by users of the bus lines, the difference of usage between working days and weekends or holidays, the performance of the trips, from start to end and within specific paths or stops of interests, among several others, can be seen directly over the process model discovered. It provides a valuable and fast way of presenting the data to business experts using existing PM tools, which give already implemented discovery algorithms. These tools are also able to summarize data showing, for example, the frequency of the trips by type of user (students, combined trips, common users, etc.), stops that are more used (activities frequency), among others, which are also provided as is by direct inspection of the input event log.

During the project, we also detected some weaknesses in the published open data for using it with the PM approach, e.g., there was no connection between an STM trip and the concrete bus in which it was used. Since event logs need to identify key elements such as case ID, activity ID, and activity timestamp, without a way to make the correspondence of existing data fields to that elements, PM cannot be directly applied. Due to this, some datasets need more work to be able to use as event logs with the required information. However, this also happens for PM applications in other domains since data can rarely be used without some manipulation. Since the necessary information was available within the STM databases, we worked with the Municipality of Montevideo to redefine the published data. The new dataset's analysis was straightforward, as presented in Section 5. We continue collaborating since some problems require information that is not published, e.g., deviation data.

7. Conclusions

This paper explored how suitable PM is for analyzing urban mobility problems. We reviewed related work finding that there are few works in the area, none delving into the transport network. We then tackle specific questions of interest defined by the Municipality of Montevideo in the context of a joint research project concerning their Metropolitan Transportation

System (STM). We analyzed them theoretically and then performed an initial PM experience using open data.

We answered fundamental questions about buses and users' behavior and have a theoretical path to answer more advanced ones. Since not much work is available in this sense, it opens new research opportunities about how PM could complement traditional transport network optimization, simulation, and data analysis techniques in the context of urban mobility. A deeper comparison with these techniques could provide information about the limitations of PM in this domain.

We are working on validating the findings with experts from the Municipality of Montevideo. This feedback could provide insights into how PM results could be better understood in this context. For example, we are integrating performance analysis with geographical information within the Montevideo map to improve results visualization.

Acknowledgements

Supported by project ”Minería de procesos para el análisis de movilidad urbana” funded by program IM-Udelar ”Ing. Oscar J. Maggiolo” from Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, and Intendencia de Montevideo, Uruguay.

References

Benevolo, C., Dameri, R. P., & D'Auria, B. (2016). Smart mobility in smart city. In T. Torre, A. M. Braccini, & R. Spinelli (Eds.), *Empowering organizations* (pp. 13–28). Springer.

Branda, F., Marozzo, F., & Talia, D. (2020). Discovering travelers' purchasing behavior from public transport data. *Machine Learning, Optimization, and Data Science - 6th Intl. Conf., LOD*, *12565*, 725–736.

Burkhardt, D., Nazemi, K., & Ginters, E. (2020). Innovations in mobility and logistics: Assistance of complex analytical processes in visual trend analytics. *2020 61st Intl. Scientific Conf. on Information Technology and Management Science of Riga Technical University (ITMS)*, 1–6.

Cui, Z., Long, Y., Ke, R., & Wang, Y. (2015). Characterizing evolution of extreme public transit behavior using smart card data. *IEEE First Intl. Smart Cities Conf., ISC2*, 1–6.

Deakin, M., & Waer, H. A. (2011). From intelligent to smart cities. *Intelligent Buildings International*, *3*(3), 133–139.

de Dios Ortúzar, J., & Willumsen, L. (2011). *Modelling transport, 4th ed.* Wiley.

Diamantini, C., Genga, L., Marozzo, F., Potena, D., & Trunfio, P. (2017). Discovering mobility patterns of instagram users through process mining techniques. *2017 IEEE Intl. Conf. on Information Reuse and Integration (IRI)*, 485–492.

Han, X., Mao, J., & Jin, N. (2010). Research on data mining of public transit ic card and application. *2010 Intl. Conf. on Intelligent Computation Technology and Automation*, *2*, 1134–1137.

Liu, Q., Zheng, X., Stanley, H. E., Xiao, F., & Liu, W. (2021). A spatio-temporal co-clustering framework for discovering mobility patterns: A study of manhattan taxi data. *IEEE Access*, *9*, 34338–34351.

Lorenzo, G. D., Sbodio, M. L., Calabrese, F., Berlingerio, M., Pinelli, F., & Nair, R. (2016). Allaboard: Visual exploration of cellphone mobility data to optimise public transport. *IEEE Trans. Vis. Comput. Graph.*, *22*(2), 1036–1050.

Massobrio, R., & Nesmachnow, S. (2019). Urban data analysis for the public transportation system of montevideo, uruguay. *Smart Cities - Second Ibero-American Congress, ICSC-Cities 2019, Revised Selected Papers*, *1152*, 199–214.

Mobility as a Service (MaaS). (2016). Mobility Process Mining.

Purnama, I. B. I., Bergmann, N. W., Jurdak, R., & Zhao, K. (2015). Characterising and predicting urban mobility dynamics by mining bike sharing system data. *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, 159–167.

Rowlson, M. (2020). Uber: Process mining to optimize customer experience and business performance. In L. Reinkemeyer (Ed.), *Process mining in action: Principles, use cases and outlook* (pp. 59–63). Springer.

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

van der Aalst, W. M. P. (2016). *Process mining - data science in action, 2nd ed.* Springer.

Yousfi, A., & Weske, M. (2019). Discovering commute patterns via process mining. *Knowl. Inf. Syst.*, *60*(2), 691–713.