A Process Mining Analysis of Woolworth's GPS Data

Suriadi Suriadi, Moe Wynn, Petia Wohed,
Arthur ter Hofstede, and Jan Recker

Information Systems School
Science and Engineering Faculty
Queensland University of Technology
Brisbane, QLD Australia

December 12, 2012

Executive Summary

A research project between Woolworths and the Queensland University of Technology under the scheme of the Woolworths Chair of Retail Innovation was conducted between August 2012 to November 2012. This project involved the application of the novel process mining techniques to analyse the trailers' GPS data collected by the Woolworths' trailer tracking system. The aim of this project was to gain *objective, evidence-based* insights (as suggested by the GPS data) into the characteristics of, and variations in, the routes taken to deliver goods between a number of Woolworths' distribution centres (DCs) and Woolworths' stores. The data used in this project were collected from both the trailers' GPS events and the corresponding goods delivery scheduling plans over a period of 19 months (Jan 2011 to July 2012).

   This report details the analyses that have been conducted by researchers at the Queensland University of Technology with respect to the Woolworths' GPS data, in collaboration with the stakeholders from Woolworths. The research questions that drove the direction of the project are explained. The treatments applied to the data for the purpose of process mining analysis are also detailed. Finally, the results from our process mining analysis are described. Our analysis results suggest that process mining techniques are capable of addressing a number of questions that Woolworths had with regards to their logistic process, resulting in the delivery of insights which may be beneficial to Woolworths (e.g. an understanding of the correlation between certain delivery routes and their corresponding delivery time).

Contents

List of Figures

List of Tables

1 Introduction

Woolworths has implemented a system that tracks the trailers (vehicles) used in the delivery of goods/supplies from Wooworths' distribution centres (DCs) to Woolworths stores in the region of Sydney, New South Wales. In particular, data related to the planning of the delivery of goods to various stores and data from the corresponding trailers' GPS units were collected. While these data have already been used by the tracking system to generate various types of report (e.g. vehicles' idle time, and vehicles' current location), we see the potential for new insights to be gained by subjecting these data to further analysis using techniques from a relatively new discipline, called *process mining* [14, 13, 7].

Process mining has been applied to address a variety of problems related to ones' business processes in various domains, including government agencies [12], finance [5, 4, 6], health [8, 10], and manufacturing [11]. Process mining allows one to gain insights about one's processes by using existing *data* as the starting point for analysis, thus enabling an *evidence-based* approach to understanding the behaviours of one's process, including its problems and, more importantly, improvement opportunities. Given the relatively wide range of domains in which *process mining* techniques have been applied and the objective nature of these techniques, it is interesting to see the extent to which *process mining* techniques can be used to address a number of questions that Woolworths had with regards to their logistic process.

Thus, a research engagement between the Queensland University of Technology (QUT) and Woolworths, under the Woolworths Chair of Retail Innovation scheme, was established. The main goal of this research project is to answer a set of questions that Woolworths would like to address through the application of process mining techniques. The questions that Woolworths had with regards to the delivery of their goods from the DCs to the stores include:

- **Q1**: what are the optimal delivery routes between any two locations to minimize the total distance travelled?

- **Q2**: to what extend do the expected delivery window is met in practice?

- **Q3**: are there variations in terms of the routes chosen between different drivers, and if so, what are their impact on performance?

- **Q4**: are there any patterns of behaviour that can be used to predict the occurrence of *unauthorized* openings of trailers' doors?

We have attempted to address the above questions using well-known process mining techniques, including process model discovery [7], fuzzy animation [1], and performance analysis. Given the explorative and preliminary

4

nature of this project, an in-depth analysis of each question (particularly Q4) is out-of-scope; however, the usefulness of process mining techniques in addressing the above questions was explored, the results of which are detailed in this report.

Based on the results of our preliminary analysis of the data available from the trailers' GPS events and store delivery plans over a period of 19 months (Jan 2011 to July 2012), we have concluded that process mining techniques can be used to address many of the above questions (especially Q1 to Q3). To address Q4, we need to use insights gained from the process mining analysis as input to classical data mining analysis - an effort that is beyond the scope of this project.

This report is organized as follows: Section 2 provides a background about process mining. Section 3 details the pre-processing activities applied to the data before they can be used for process mining analysis. Section 4 summarizes the analyses undertaken and the corresponding results. The conclusion and future work are summarized in Section 5.

2 Background

Today's organizations rely heavily on IT systems to support their business operations. These systems record a vast amount of valuable data which can be studied and analyzed to reveal useful information to support business operations. However, this data is often left unexploited, resulting in a stark contrast between the perceived manners in which business operations are conducted (e.g. in the form of process models/flow charts) and their reality. This situation can easily lead one to make ill-informed decisions that may have detrimental consequences. Therefore, the abundance of business data recorded in today's IT systems should be exploited in order to support business decisions and to improve business processes. *Process mining* is a novel approach that facilitates such an exploitation of data.

Process mining consists of a set of analysis techniques which together can be used to understand the actual behaviour and performance of business processes by studying events recorded in IT systems (also known as event logs). The true benefit of process mining lies in the fact that it allows *evidence-based* business process improvement and re-design. The three main insights facilitated by process mining include process discovery, conformance, and enhancement (see Figure 1).

Process Discovery. Process mining allows the discovery of the actual manners in which various business processes were conducted and as such may provide valuable insight into differences between processes as they are expected to be executed and the way they actually are.
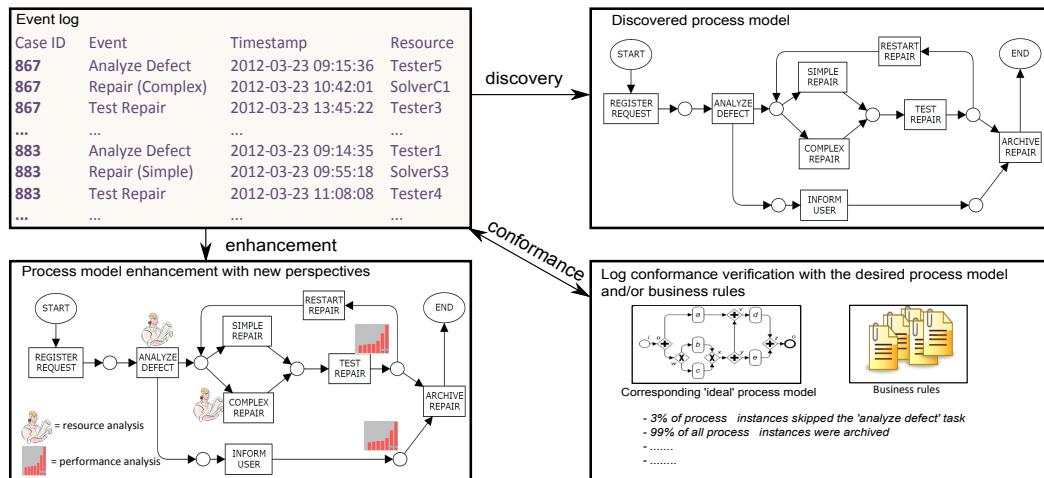
Figure 1: Key Capabilities of Process Mining

Conformance. If an organization already has a collection of process models (or business rules) for their respective processes, process mining can verify the extent to which executed processes actually conformed to the models and/or business rules.

Enhancement. Process mining also enables one to gain insights into other dimensions of business processes, including the relationships between resources (social network) and the performance of various business processes (e.g. bottleneck analysis).

Beneficial insights typically obtained from process mining include:

- the uncovering of new opportunities with regards to the way processes are executed,

- the identification of root causes explaining why certain process variants have better performance than others,

- the discovery of key indicators that can predict the behaviour of process instances in the near future,

- the confirmation/refutation of long-held beliefs about the behaviour of one's processes, and

- the identification of problem areas in existing business processes (e.g. process anomalies).

Two main ingredients required to start a process mining analysis are: (1) a set of questions to be addressed, and (2) the corresponding data that

case id	event id	timestamp	activity	resource	cost	...
			properties			
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	...

Figure 2: Example of an event log

contain the necessary information to answer the questions. These two ingredients were present in the project conducted with Woolworths, thus, allowing us to embark on this process mining analysis.

3 Data Extraction and Pre-processing

3.1 Data Quality

The quality of process mining analysis greatly depends on the quality of the data being analysed. The importance of having the right set of data in a process mining analysis is crucial and should not be underestimated. For example, attempting to discover the social network of resources working in a process will be impossible if the data set does not record any resource information at all.

Ideally, at a minimum, log data for process mining must contain the set of *activities/events* related to the process to be studied. For example, in Figure 2, the activities related to instances of an insurance claim request process recorded in the log are 'register request', 'examine thoroughly', 'check ticket', and so forth.

Additionally, each event recorded in the log must be linkable to a *case*. A case can be described as a sequence of events that together work towards achieving a particular end. For example, a case can be seen as a collection of steps needed to fulfil the lodgement of an insurance claim by a customer, or to approve an international travel application of a staff member. The log shown in Figure 2 contains activities related to two (2) cases, each identified with case identifier '1' and '2' respectively.

Finally, one must be able to order the set of activities/events recorded in the log according to the time when they occurred. This is normally

7

translated into the fact that each event recorded must have a *timestamp*.

Woolworths' Data   There are two main data sets obtained from Woolworths. The first data set contains information related to the planning of the goods delivery to various stores. The data were obtained from the logistic application called `TACTICS`. This data set contains information related to the date a delivery was to be performed, the start location, the end location, the stores to be visited on each delivery, the expected delivery window time, and other related information. Henceforth, we refer to the first data set as the 'TACTICS' data. A sample of this data set is provided in Table 1.

Date	Daily Runsheet TransportID	Store ID	Start/End Location	Trailer ID	Delivery Window	...
2011-01-07	1246300	Winston Hills	1954/1954	AL24192	2011-01-07 18:30 - 20:30	...
2011-01-07	1246300	Rouse Hill	1954/1954	AL24192	2011-01-07 13:00 - 17:00	...
2011-01-07	1246300	Kellyville	1954/1954	AL24192	2011-01-07 17:00 - 19:00	...
2011-01-07	1246300	Blacktown	1954/1954	AL24192	2011-01-07 18:00 - 22:00	...
...	...	...	...	...	...	...

Table 1: A snippet of data related to goods delivery plan

The second data set obtained contains data from the GPS system. It records various GPS-related events, such as geofence entry, geofence[1] exit, routine location polling data, ignition on/off, and engine going to sleep. Henceforth, we refer to the second data set as the 'GPS' data. Table 2 shows a snippet of the GPS-related data.

Date	Event	Trailer	Lattitude	Longitude	Address	...
2011-01-01 00:01:00	Geofence Exit	A241107	-33.7835	150.814	Sargents RD Minchinbury	...
2012-01-01 00:01:00	Door Open	AL241171	-33.8887	150.849	Balmoral Cecil Hills	...
...	...	...	...	...	...	...

Table 2: A snippet of data related to Trailer's GPS events.

These two data sets, in their raw form, are not ideal for process mining analysis: while the GPS data contain the necessary *activities* and *timestamp* information related to the delivery of goods between DCs and stores, the notion of *case* is missing in the data. In other words, each event in the GPS log *cannot be linked* to any particular *case*. More importantly, it was not even clear in the beginning what constitutes a case in the process of goods delivery from DCs to stores.

---

[1]A geofence refers to a particular geographic area that is bounded by certain perimeters, e.g. an area which is recognized as part of a Woolworth's store
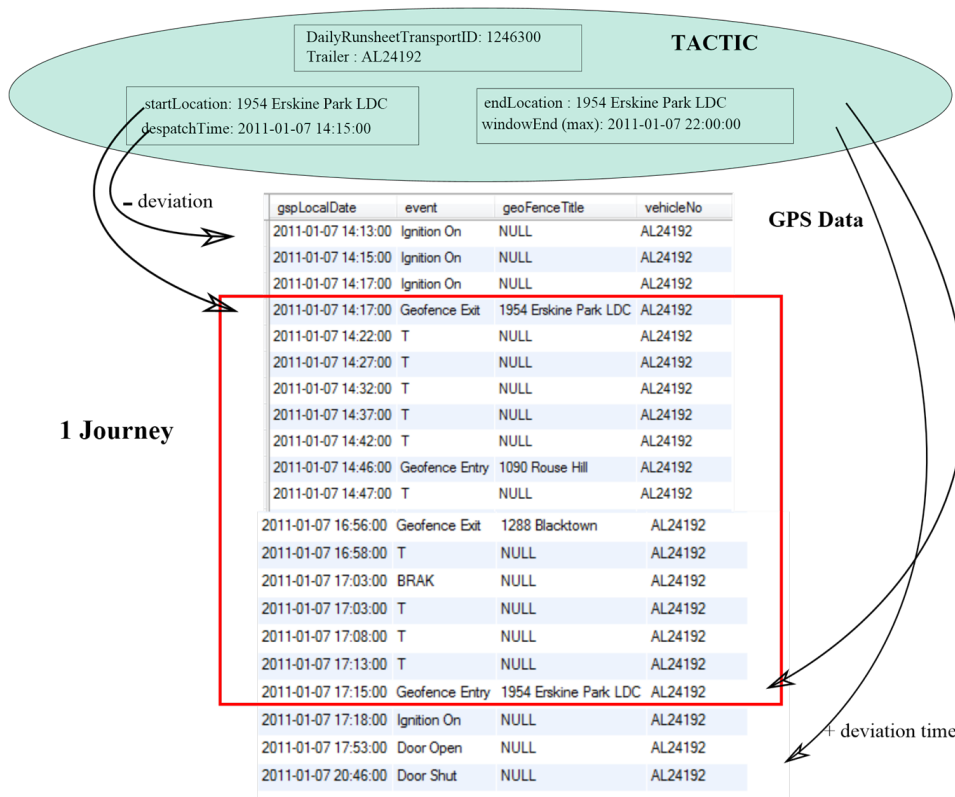
Figure 3: The definition of a 'case' for process mining analysis, which is defined as a 'journey' taken by a particular trailer.

3.2 Case Definition

By analysing the data, and through a discussion session with the stakeholders from Woolworths, we obtained the definition of a 'case'. A 'case' is defined as a *journey* taken by a particular trailer/vehicle from the time the trailer exits a particular geofence location (i.e. the start location for that particular journey), along all the stops at intermediary stores, to the time when the trailer enters the geofence entry point of the end location for that journey (which may be the same as the start location, but not necessarily so).

Figure 3 shows how we can map the definition of a 'case' or a 'journey' into the two sets of data we obtained. In general, we use the TACTICS data to obtain all the journeys that were recorded in the data. Each journey is uniquely identified by the 'dailyRunsheetTransportID' field.

As shown in Table 1, the same 'dailyRunsheetTransportID' can be repeated over a number of entries; however, each entry refers to a different store location. Thus, what Table 1 shows is that the journey, identified by the number '1246300', involved goods delivery to four stores: Rouse

9

Hill, Kellyville, Winston Hills, and Blacktown. Furthermore, this journey started and ended at the same location, identified by the number '1954'.[2] Finally, the earliest time when the trailer should arrive at the first stop (i.e. Rouse Hill) was '2011-01-07 13:00:00', and the latest (maximum) window time allowed for the trailer to arrive at the last store (i.e. Blacktown) was '2011-01-07 22:00:00'. By counting the number of *unique* 'dailyRunsheetTransportID' from the TACTICS data set, we thus obtained the total number of potential 'journeys' (or cases) that we can analyse from the data.

Figure 3 shows how we can obtain all the necessary GPS events for that particular journey by doing the necessary correlation between the TACTICS data and the GPS data. From the TACTICS data, for journey number '1246300', we can obtain the trailer ID involved (which was 'AL24192'), the *despatch time* for that journey (which was '2011-01-07 14:15:00'), and the *maximum/latest window* time the trailer should arrive at the last store for that journey (which was '2011-01-07 22:00:00'). Based on this information, we then search the GPS data for *all events* that:

- were started a number of hours *before* the *despatch time* to allow for any time deviation that may happen in practice, and

- were ended a number of hours *after* the *maximum window* time to allow for any time deviation as well, and

- were related to that particular trailer as stated in the 'TACTICS' data (i.e. 'AL24192').

Once we obtained all events that fulfilled the above three conditions, we then identified the *first* geofence exit event from the start location of that journey (in this case '1954 Erskine Park LDC'), and removed all events that happened before this geofence exit event. Similarly, we also identified the *first* geofence entry event to the end location for that journey (which is similar to the start location in this case), and removed all events that happened after this geofence entry event. The remaining events thus form a *journey* based on the definition we provided earlier (i.e those events inside the red box in Figure 3).

3.3   Data Correlation

The goal of the data correlation exercise performed on the TACTICS and GPS data is to be able to tag relevant events in the GPS data with their corresponding journey identifier (i.e. 'dailyRunsheetTransportID'). To ensure accurate results, basic characteristics about the data need to be established.

---

[2]This location number can be mapped to a particular place properly by correlating it with information from another auxilliary data set which contains all the location information.

To do so, we have imported all the data into a MySQL database and verified a number of key characteristics of the data, including the fact that:

- every journey (i.e. every 'dailyRunsheetTransportID') involves *exactly one* trailer/vehicle,

- every journey has *exactly one* despatch time, and

- every journey has *exactly one* start location and end location.

Data Completeness Issues. Upon further data analysis, we identified a number of issues with data completenes. For example, 56,313 journeys (out of the 153,426 journeys recorded in the TACTICS data) do not have start and end location identifiers that can be directly mapped to *known* locations. In other words, close to 37% of the journeys in the TACTICS data set started and ended in locations that were 'unknown'. Furthermore, certain journeys in the TACTICS data also do not have the necessary despatch time and/or maximum window end time information. Therefore, if we consider only those journeys with usable start/end location name, and proper recording of the despatch time and window end timestamp, there are a *maximum* of 96397 journeys that we can analyse.

There are 4,588 distinct trailer IDs recorded in the TACTICS data. However, there are only 200 distinct trailer IDs in the GPS data. In other words, our analysis can only track the movement of less than 5% of all trailers recorded in the TACTICS data. Of course, this is likely to reduce substantially the number of distinct journeys that we can correlate in the GPS data.

Data Integrity Issues. We also encountered a number of data integrity issues (which may be the result of incorrect recording or system glitches). For example, the journey with ID '1687289' contains one 'outlier' entry: the window start time for one of the stops has the timestamp that is almost *one year* apart from other stops. Furthermore, there were 8,860 journeys in which the despatch time was *later* than the latest window end timestamp. Similarly, there were around 2898 journeys in which the window end timestamp was *earlier* than the window start timestamp. These problematic entries were generally excluded in our analysis.

Data Correlation Results. Using the data correlation logic explained earlier in Figure 3, we ran the data correlation exercise between the TACTICS data and the GPS data. There are a total of 6,218,185 events recorded in the GPS data. Through this correlation exercise, we managed to identify 56,517 journeys from the events recorded in the GPS data. These journeys were derived from a total of 1,388,436 events. This means that there are

4,829,749 entries in the GPS data that we could not correlate/unaccounted for.

These unaccounted events in the GPS data could be expected (e.g. 'intermediary'/'in-between-journeys' events that we explicitly did not consider in our correlation logic). Equally probable is that our definition of a journey could have been too simplistic, not accounting for complications in the trailers' movement (e.g. a quick trip between two geofence locations to pick up certain goods that was not meant to be part of any particular journey). Further analysis is needed to polish the correlation logic.

The subsequent process mining analysis described in the remainder of this report only considered those journeys that we have successfully identified from our preliminary correlation exercise. Data related to these journeys were converted into the standard eXtensible Event Stream (XES) format [15] so that they could be used for process mining analysis.

4   Process Mining Analysis and Results

Using the data that we have processed into XES format, we conducted a number of preliminary process mining analyses to address some of the questions listed in Section 1. To get a better understanding about the data, basic performance anlayses were performed.

4.1   Basic Performance Analyses

Basic performance analyses were conducted using custom-built process mining tool, such as the DISCO tool,[3] which neatly summarized basic performance statistics related to the data being analyzed.

Out of the 56,517 journeys that we have managed to identify, around 55% of all journeys (or about 31,620 journeys) were completed between 50 minutes to over 1 day - the remaining cases were completed in less than 50 minutes. We did not consider those cases that were completed in less 50 minutes because we were not sure if it actually made sense to have a journey completed within such a short period of time.

Out of those 31,620 journeys, the average performance distribution is shown in Figure 4. The top diagram in Figure 4 shows the average journey duration. We can see that the case duration between various journeys follows more or less a 'normal' distribution, with the mean value of 2 hours and 42 minutes (the minimum value is 52 minutes, while the maximum value is 1 day and 13 hours).

Similarly, the mean number of GPS events recorded per journey is 30 (see the middle diagram in Figure 4). The minimum number of events recorded per journey is 3, while the maximum number of events per journey is 288.

---

[3]http://fluxicon.com/disco/

12

Figure 4: Basic Performacne Statistics of Goods Delivery Between DCs to Stores

The bottom diagram in Figure 4 shows an interesting result: up until end of May 2011, there were on average 200 *active journeys (or cases)* per day. However, since then, the number of active journeys per day drop to fewer than 100 per day. This could be due to a conscious decision to reduce the number of journeys recorded in the log, or could be due to other business decisions (e.g. change in the way journeys were to be planned, resulting in the overall reduction of the number of journeys conducted per day).

4.2   Finding Optimal Routes - Q1

The next analysis performed was to address Q1: finding optimal routes between any two locations to minimize total distance travelled. To do this, we needed to do some modification to the data that we have converted into the XES format. In particular, instead of treating the GPS events (e.g. geofence entry, geofence exit, and ignition on) as the *activities* in the log, we treated the 'address' field as the 'activity'. Thus, instead of obtaining a

sequence of GPS events, we now obtain a *sequence of locations* that a trailer passed through in a journey.

Next, to demonstrate the usefulness of process mining in finding optimal deliver routes, we took as an example the delivery between two *direct* endpoints (no intermediary stops): Erskine LDC to Nowra. By filtering the log to only contain those trips between these two endpoints, we obtained a clear picture about the various routes that trailers took in the delivery of goods between these endpoints.

In total, there were 190 total number of delivery between these endpoints over a period of 1 year. We divided these 190 trips into three classes: those that completed 'quickly' (between 2 to 2.5 hours), those that were 'slower' to complete (between 2.5 to 3.5 hours), and those that were really 'slow' (between 3.5 to 4 hours). The 'quick' class contains 115 trips, while the 'slower' and 'slow' classes contains 62 and 13 trips respectively.

Finally, we compared the two extreme classes ('quick' vs. 'slow') to identify the differences in the routes taken that caused the differences in the performance of the trips between these two classes. These differences were identified through the application of *proess discovery* and *fuzzy animation* techniques which (readily usable in the process mining tool, called the ProM Tool [2]).

There are a number of process discovery algorithms, including the Heuristic Miner [16], Fuzzy Miner [3], and many others. In this project, we applied the Fuzzy Miner in order to discover the variations in the routes taken between 'quick' and 'slow' trips. To facilitate easy comparison, we firstly created one 'common' process models that could capture all the possible routes taken by both the 'quick' and 'slow' trips. This model was derived by learning the data related to both 'quick' trips and 'slow' trips.

Once we obtained this one 'common' model, we then replayed the related data using the fuzzy animation [1] technique. During the animation, the more frequent a path between two activities (or in this case, between two locations) was traversed, the thicker the line became. Thus, at the end of the animation, we obtained a 'map' of most commonly-traversed routes: well-traversed/dominant routes are thicker than infrequently-traversed paths. By running the animation twice (first with the data related to 'quick' trips, and the second one with data related to 'slow' trips), and comparing the resulting maps, we can then visually identified the differences in the routes taken.

Figure 5 shows the resulting maps for both the 'quick' trips (left) and the 'slow' trips (right). Note how these two models share exactly the same structure, however, we can see that the dominant routes taken are very different between these two models because the thickness of the lines are different.

We have thus shown how we can apply process mining techniques to quickly and visually identify the differences in the routes taken between those 'quick' and 'slow' cases. We can then use the identified differences as
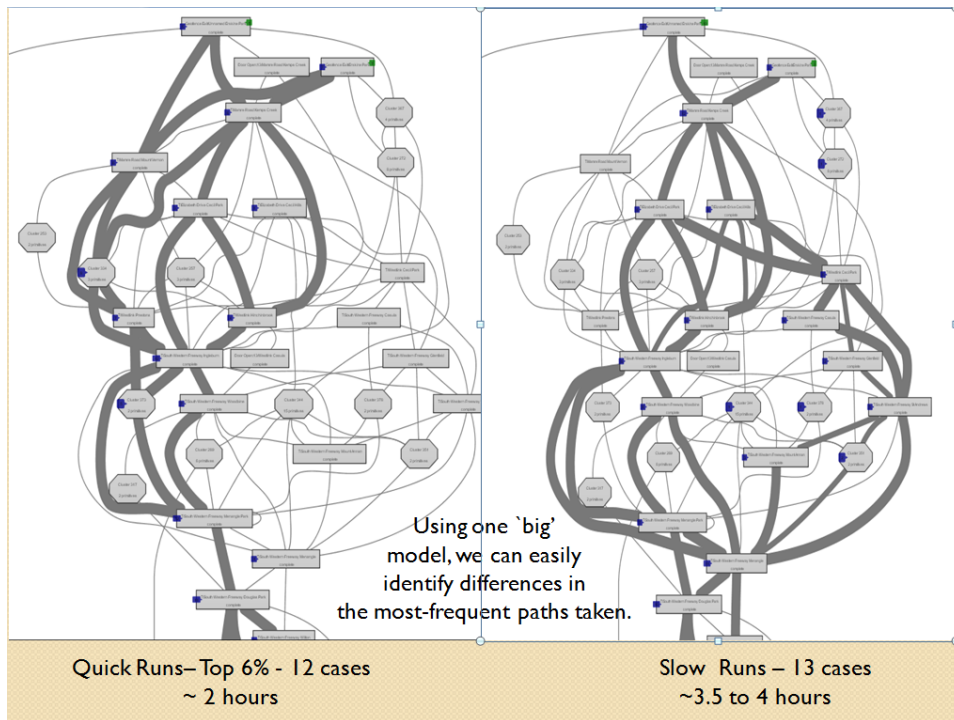
Figure 5: Comparison of routes taken between 'quick' trips and 'slow' trips

input to address Q1: we have identified 'optimal' delivery routes which, on average, took less time to complete. Of course, a more detailed analysis of the exact distance of these 'optimal' routes need to be taken to be sure that these optimal routes do translate into shorter distance travelled.

4.3 Accuracy of Delivery Window - Q2

To address Q2, we can use a custom-built tool, such as the DISCO tool to obtain the average delivery time distribution between any two endpoints. For example, by using the log that has been filtered to only contain trips between the Erskine LDC location and Nowra, we can use the DISCO tool to quickly show the delivery time distribution as shown in Figure 6.

From Figure 6, we can see that the majority of trips (61%) were completed between 2 to 2.5 hours. A minority of trip (around 6%) were completed between 3.5 to 4 hours, while in an extreme case, it took over 5 hours to compelte the trip. By using this information, we can then address Q2 by comparing the currently-used metrics in the estimation of delivery window arrival time in Woolworth's TACTICS application with those metrics shown in Figure 6.
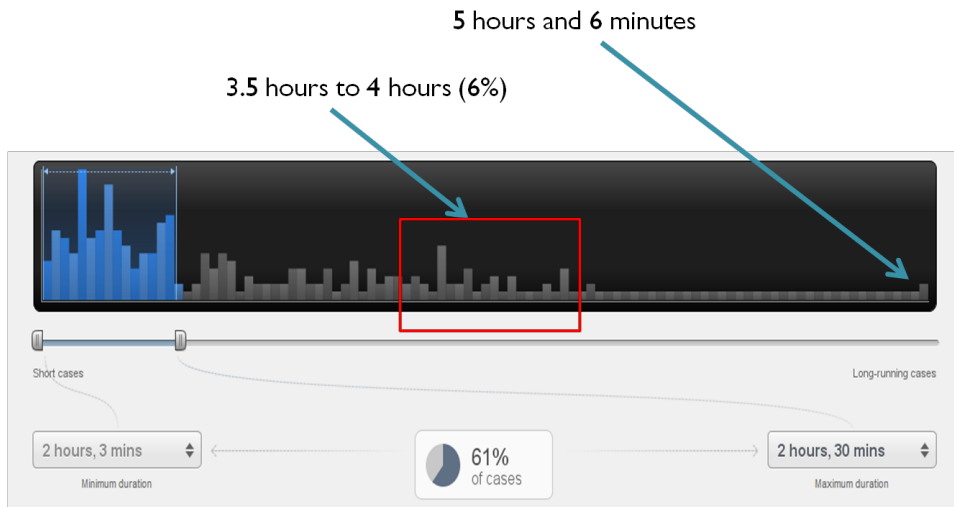
15

Figure 6: Delivery time distriibution between Erskin LDC to Nowra Store

4.4 Drivers' Behaviours - Q3

We can use the same combination of techniques to address Q3 as those used to address Q1: instead of comparing log data between 'quick' and 'slow' cases, we could split the log into those journeys driven by `DriverA` and those journeys driven by `DriverB`, between the two same endpoints (*direct* endpoints without intermediary stops). We can then discover a common process model based on these two logs, and used the same fuzzy animation to obtain the two maps that show the differences (if any) in the routes taken between the two drivers.

This analysis has not been conducted yet due to the inconsistent recording of the drivers' identifier in the log (i.e. the same driver can be reflected with two more different identifiers in the log).

4.5 Predicting the Occurrence of Certain Events - Q4

Addressing Q4 involves identifying patterns in the way a journey was conducted which gave rise to the occurrence of illegal door openings. Process mining alone cannot address this question fully; however, insights gained from process mining analyses can be used as input to classical data mining analysis, such as *classification* analysis [9], in order to identify the relevant predictors. For example, the choice of certain delivery routes by a particular driver (identified through process mining analysis) may be used as inputs to the well-known *classification* analysis to determine the precise causal relationships between the choice of a particular route and the occurrence of illegal door openings.

This project has not had the opportunity to go into such an in-depth

analysis. Nevertheless, a similar analyses with regards to Q1 has been conducted. In particular, we have done a preliminary analysis to identify relevant factors that may cause the occurrence of 'slow' trips. The results of our analysis show a discernible causal relationship among unauthorised door opening events, the use of a particular trailer, and the occurrence of 'slow' trips in the delivery of goods between Erskin LDC to Nowra:

- 'slow' trips contain a *higher proportion* of unauthorised door openings (about 7%), as opposed to 'quick' trips (about 3%), and

- about 50% of all unauthorised door openings occurred in a trailer with the identifier of 'AL241186'.

Of course, the exact causal relationships between the indicators listed above and the occurrence of 'slow' trip duration still need to be verified using the classical *classification* analysis from the field of data mining - an analysis that we have not had the chance to conduct in this project.

5   Conclusions

This report has summarised the preliminary process mining analyses that we have conducted on the Woolworths' GPS data. In particular, data related to the planning and delivery of goods from DCs to Woolworths' stores have been analysed using a number of process mining techniques. From our preliminary anlayses, we have shown the potential of process mining in addressing a number of questions that Woolworths' had with regards to their logistic process. Of course, deeper analyses to address each of the questions (listed in Section 1)is still needed. These in-depth analyses can be considered as part of the future work.

Acknowledgements

We would like to acknowledge the stakeholders from Woolworths, including Adrienne Renner, Chris Clarke, Jeff Forward, Michael Livingstone, and Peter Roebers, for their support and valuable insights throughout this research project.

Bibliography

[1] M. de Leoni, W.M.P. van der Aalst, and A.H.M ter Hofstede. Process mining and visual analytics: Breathing life into business process models. *BPM Center Report*, 2011.

[2] B.F. Dongen, A.K.A. Medeiros, H.M.W. Verbeek, A.J.M.M. Weijters, and W.M.P. van der Aalst. The ProM framework: A new era in process mining tool support. In *ICATPN*, volume 3536 of *LNCS*, pages 444–454. Springer, 2005.

[3] C. Gunther and W.M.P. van der Aalst. Fuzzy mining- adaptive process simplification based on multi-perspective metrics. In *BPM*. Springer, 2007.

[4] M. Jans, M. Alles, and M. Vasarhelyi. The case for process mining in auditing: Sources of value added and areas of application. *International Journal of Accounting Information Systems*, 2012.

[5] M. Jans, J.M. van der Werf, N. Lybaert, and K. Vanhoof. A business process mining application for internal transaction fraud mitigation. *Expert Syst. Appl.*, 38(10):13351–13359, September 2011.

[6] M.J. Jans. Process mining in auditing: From current limitations to future challenges. In *BPM Workshops*, pages 394–397. Springer, 2012.

[7] T. Jones. Understand how abpd can help improve business processes. *Gartner*, March 2012.

[8] R. Mans, M. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W.M.P. van der Aalst. Process mining techniques: an application to stroke care. In *Ehealth Beyond the Horizon- Get it There*, 2008.

[9] Dan H. Moore. Classification and regression trees, by leo breiman, jerome h. friedman, richard a. olshen, and charles j. stone. brooks/cole publishing, monterey, 1984,358 pages. *Cytometry*, 8(5):534–535, 1987.

[10] Á. Rebuge and D.R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Inf. Syst.*, 37(2):99–116, April 2012.

[11] A. Rozinat, I.S.M de Jong, C.W. Gunther, and W.M.P. van der Aalst. Process mining applied to the test process of wafer scanners in ASML. *IEEE Trans. on System., Man, and Cybernetics, Part C*, 39(4):474 – 479, 2009.

[12] W. M. van der Aalst, H. A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K.A. de Medeiros, M. Song, and H. Verbeek. Business process mining: An industrial application. *Information Systems*, 32:713–732, 2007.

[13] W.M.P. van der Aalst. *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[14] W.M.P. van der Aalst et al. Process mining manifesto. In Florian Daniel et al., editor, *BPM Workshops (1)*, volume 99 of *LNBIP*, pages 169–194. Springer, 2011.

[15] H.M.W. Verbeek, J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst. XES, XESame, and ProM 6. In *Information Systems Evolution*, volume 72 of *LNBIP*, pages 60–75. Springer, 2011.

[16] A.J.M.M Weijters, W.M.P. van der Aalst, and A.K.A. Medeiros. Process mining with the heuristic miner-algorithm. Technical report, Eindhoven University of Technology, 2006.