Process Mining Applied to Postal Distribution

Victor Martinez[1], Laura Lanzarini[1,2], and Franco Ronchetti[1,2,3]

[1] *Facultad de informática, Universidad Nacional de La Plata*
[2] *Instituto de investigación en Informática LIDI (UNLP-CIC)*
[3] *Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC-PBA)*

martinezvictor@hotmail.com, {laural, fronchetti}@lidi.info.unlp.edu.ar

Abstract. Process mining is a technique that allows analyzing business processes through event logs. In this article, different process mining techniques are used to analyze data based on the postal distribution of products in the Argentine Republic between the years 2017 and 2019. The results obtained allow stating that 85% of the shipments made conform exactly to the model. The analysis of the situations with a low level of adjustment to the discovered process constituted a tool for quick identification of some recurring problems in the distribution, facilitating the analysis of the deviations that occurred. In the future, we expect to incorporate these techniques to build early notifications that warn about the existence of excessive deviations from the process

Keywords: Process Mining, Data Mining, Postal Distribution, Postal Processes, Business Process Management

1 Introduction

Currently most information systems record the activities carried out, whether they are business management systems (ERP, CRM, WMS, BI, etc.) or developments specific to each company. In general, the information recorded includes what was done, who did it, and when it was done, among other data. From this stored data, information that describes the process can be obtained and then used to identify improvement opportunities or solves problems.

Process Mining techniques are used to analyze these data and find patterns of behavior. The tasks with which a process begins, the sequence followed, and the tasks run to end the process can be identified. This allows "discovering" the process that is being carried out for a certain activity. With process mining, you can answer questions such as: What really happened? Why did it happen? What could happen in the future? How can process control be improved? How can the process be redesigned to improve performance? [1]. One of its main advantages is that it allows working directly on real data and obtaining the true behavior of the process, which, in some cases, is not the one originally designed.

In this article, process mining techniques will be applied to postal distribution in the Argentine Republic to analyze its operation and find operational deviations, bottlenecks and other problems that negatively impact service quality.

In postal distribution, a record is kept of all activities, from the entry of the product to its delivery to the customer. These activities include receiving the shipment, entering a warehouse, internal transfers or delivery attempts, among others. All must be done in a specific order and within a given time frame. Occasionally, deviations occur. These may be task redundancy or inconsistency (tasks are repeated or not performed in the corresponding order), excessive time to completion, or others.

In the literature, there are works that relate the postal business, big data and process mining. Such is the case of [2], which uses data mining in a big data environment in the China Post. Due to the nature of the postal business, in that article, clustering techniques were used to group customers based on behavior, consumption habits and focus of interest, generating a more accurate and effective postal marketing strategy with very satisfactory results. Another example is [3], where process mining is applied in logistics to look for similarities and differences between different delivery processes in a changing context of manufacturing and logistics. In that work, different processes are compared using clustering techniques to achieve an automated documentation of processes in a changing context. In [4], a methodology is presented that serves as a guide for the execution of process mining projects that describes the different stages. In addition, its actual application to the IBM purchasing process is shown as a case study.

In this article, different process mining techniques will be used to analyze data based on the postal distribution of products in the Argentine Republic between the years 2017 and 2019. The authors of this work wish to state that, at the date of generation of this document, they are not aware of the existence of any similar works that implement process mining to postal distribution in the Argentine Republic.

2 Process Mining

The starting point of process mining is the event log. The process to be analyzed is assumed to require the recording of a series of sequential events pertaining to the activity and that are related to a specific case. Even though additional data can be stored for each event, recording the date of the event (day and time), the case identifier, and the type of event is mandatory.

The three basic types of process mining [5] are:

- *Discovery*: Discovery techniques take a process log as a starting point and generate a model without any additional information. A relevant example is the algorithm Alpha [6], which takes log data and generates a Petri net that explains the behavior reflected in the log. Like any technique that extracts knowledge from data, the quality of the discovered process will depend on the degree of representation of the events surveyed in relation to how the process operates. Process parts that are not represented by events cannot be discovered.

- *Compliance Verification*: It consists of comparing an existing model (it may be the one previously discovered or a different one) with the actual sequence of events to identify deviations and verify how the process works. Applications capable of generating graphical representations and animations are usually used to observe actual behavior and see to what degree it follows to the originally defined process. Its main advantage is that it shows reality, it is not a simulation, meaning that a much more accurate analysis can be carried out.
- *Improvement*: This type of process mining is aimed at extending or improving the existing process through the underlying information in the sequence of events. Unlike the Compliance Verification type, where data are compared to the model, the goal here is to modify the process.

The degree of abstraction to be used during the analysis should be regulated considering the following aspects: fitting, accuracy, generalization and simplicity [1]. There is a relationship of compromise between these that has to be taken into account to achieve good results. Fitting refers to the ability of the model to explain the observed behavior. Accuracy refers to the accuracy with which the process is executed. In this sense, it is important that the model is not overfitted to input data because this would result in a lack of generalization, preventing the desired level of abstraction from being achieved. Simplicity is also affected by overfitting, as it is achieved by adding more detail to the process description.

Finally, it should be noted that the result obtained from the process mining analysis is highly linked to the quality of the input data. In fact, it is a known fact that there is always a certain amount of noise in the data, which can be due to incomplete tracing, intervals that have not been correctly recorded, or data duplication. This information can distort or falsify the result of the analysis [7].

In general, input data has to be verified and preprocessed to remove as much noise as possible.

3 Discovering the Postal Distribution Process

In this section, the discovery of the process will be carried out using data from the postal distribution of products in the Argentine Republic between the years 2017 and 2019.

The postal distribution process encompasses different types of products. In all cases, the process consists of receiving the product from the sender (either by physical means or a digital channel) and attempting delivering it to the recipient; actual delivery may or may not be successful. Product non-delivery does not mean that the process has failed, since there may be reasons to account for the situation.

The process records at least the following activities: receipt and identification, shipment for distribution, one or more delivery attempts, waiting at the distribution center, and return (if delivery was not possible). In all cases, each steps carried out must be recorded with a unique shipping identifier, which allows knowing the current status of the shipment and providing that information to the customer.

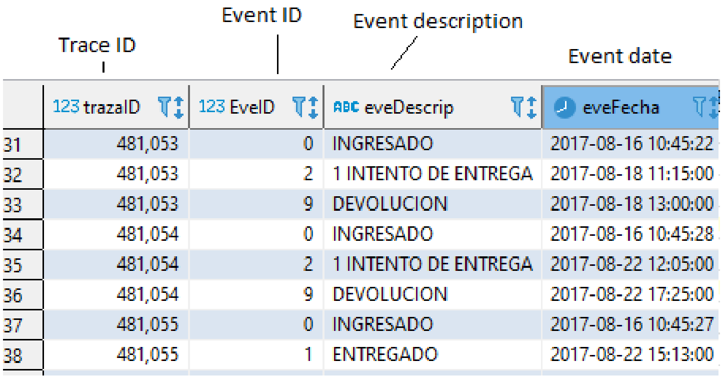
Data Extraction

The first step in discovering the model consists in collecting and preprocessing the data to be used. In particular, for this case study, product shipments with two home delivery attempts were used. The current procedure establishes that those products that cannot be delivered are kept for a given time so that the recipients can pick them up; after that time, the packages are returned to their corresponding senders.

A trace is defined as a shipment. Each movement that is recorded for that shipment is an event. A trace is considered to be completed when the shipment has an entry and an end on record, with successful delivery or not.

As a result of the data collection, a table or sample of around 33,000 traces and a total of 78,000 events was generated. For each event, the minimum required fields for the analysis were recorded; namely, trace ID, event ID, event date and event description (Figure 1). Each trace can have one or more associated events. Trace identifier and identifier of each of the events is needed to build the trace history.

This allowed having available the different steps or events that occurred throughout the shipment of each product.



	123 trazaID	123 EveID	ABC eveDescrip	eveFecha
31	481,053	0	INGRESADO	2017-08-16 10:45:22
32	481,053	2	1 INTENTO DE ENTREGA	2017-08-18 11:15:00
33	481,053	9	DEVOLUCION	2017-08-18 13:00:00
34	481,054	0	INGRESADO	2017-08-16 10:45:28
35	481,054	2	1 INTENTO DE ENTREGA	2017-08-22 12:05:00
36	481,054	9	DEVOLUCION	2017-08-22 17:25:00
37	481,055	0	INGRESADO	2017-08-16 10:45:27
38	481,055	1	ENTREGADO	2017-08-22 15:13:00

Fig. 1. Example of events extracted for analysis

Then, the data were then transformed using the XES format [8]. XES is a grammar for a label-based language whose objective is to provide information systems designers with a unified and extensible methodology to capture system behaviors through event logs and flows [8]. Thus, data management is streamlined and can be processed by different tools more efficiently.

To facilitate the discovery of the correct process, all incomplete traces were eliminated. An incomplete trace is a trace that does not have either a start event (reception) due to a loading error or a final event (delivery, return, recipient non-existent, deceased, refused shipment, or moved). The latter could occur due to a loading error or because the stipulated time to finish the process has not yet elapsed.

Using a filter of simple heuristic rules, complete traces were identified, and only those that had valid initial and final states remained. As a result of this filtering process, approximately 16,000 traces with 43,000 events were obtained.

3.1 Process Model

To generate the process model, a classic process mining algorithm was used, the Alpha algorithm, first proposed by van der Aalst, Weijters and Maruster[6]. The objective of this algorithm is to reconstruct causality from a set of sequences of events. It builds Petri nets with special properties (workflow nets) from event logs (such as those that an ERP system might collect). Each transition in the network corresponds to an observed task.

As regards our case study of postal distribution, even though it is possible to build the model directly from the 16,000 previously mentioned traces, some cases considered to be the most representative ones were manually selected to find the ideal process that the traces should comply with, thus simplifying this discovery stage of the process.

Figure 2 illustrates the Petri net corresponding to the discovered process that the traces must fulfill.

The process found is as follows: All traces must begin with an entry event and then go out to distribution. If a trace can be delivered, the event is logged and the process ends. If it cannot be delivered, in case of a final event (recipient deceased, missing address data, no address, unknown, etc.), the reason is recorded and the process ends. If it is not a final event (for example, the address could be reached but there was no one home) a first delivery attempt is recorded. A new visit is made at a later date. If it could be delivered, the event is logged and the process ends. If on the second visit the trace cannot be delivered, the shipment is kept at the office for a time, waiting for the recipient to come to pick it up, at which time the delivery at the office is recorded. After the waiting time has elapsed, if the recipient has not come to pick up the shipment, a return event is recorded. Those are the possible scenarios contemplated by the process.

3.2 Model Verification

The process model discovered in the previous stage was generated from a subset of previously selected traces. In this section, the compliance check carried out will be described. As previously explained, our goal is to establish how well sample traces comply with the process. Steps fulfilled, steps missed, and any additional steps not reflected in the discovered process will be considered. To do this, each of the traces obtained will be compared against the discovered found

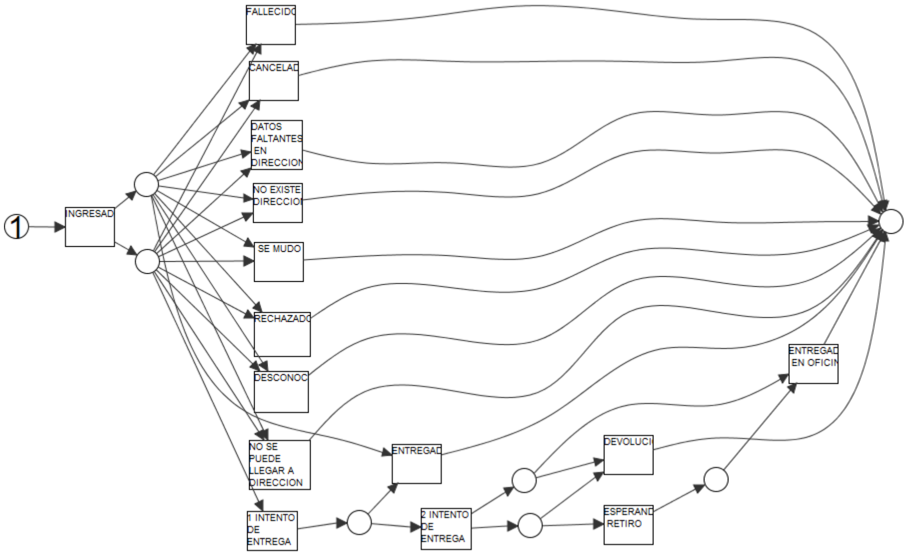


Fig. 2. Process discovered that all traces must fulfill

in the previous section. The data sample obtained in the previous section will be used, which consists of 16,811 traces with a total of 43,888 events.

The result is shown in Figure 3, with the most common events highlighted in a dark color and the most frequent runs represented with a thicker line. It can be seen that most of the traces end with the delivery, either on the first or the second attempts.

Sample statistics indicate that in 80% of the cases, the parcel is delivered either in the first or the second attempt, and that the remaining 20% is evenly distributed.

Based on these observations, it can be stated that 85% of the traces perfectly fit the process discovered, with an average number of 2.6 events in each trace.

However, some traces do not exactly follow this process, either because steps are skipped, they are not carried out in the correct order, or there are additional steps not included in the model. These traces have a compliance value that will be lower as the deviation from the model increases. Those whose compliance is below 50% are the object of analysis in this section, since this value is considered to be a significant operational deviation.

Based on these results, our analysis is focused on two paths – on the one hand, traces that have an excessive number of movements, and on the other, those that do not fit the model, either because there are missing events or because they do not follow the corresponding sequence.

In the first case, given that the number of movements per trace follows a normal distribution with a mean of 2.6 and a deviation of 0.95, it was considered appropriate to use the value of the mean plus three standard deviations

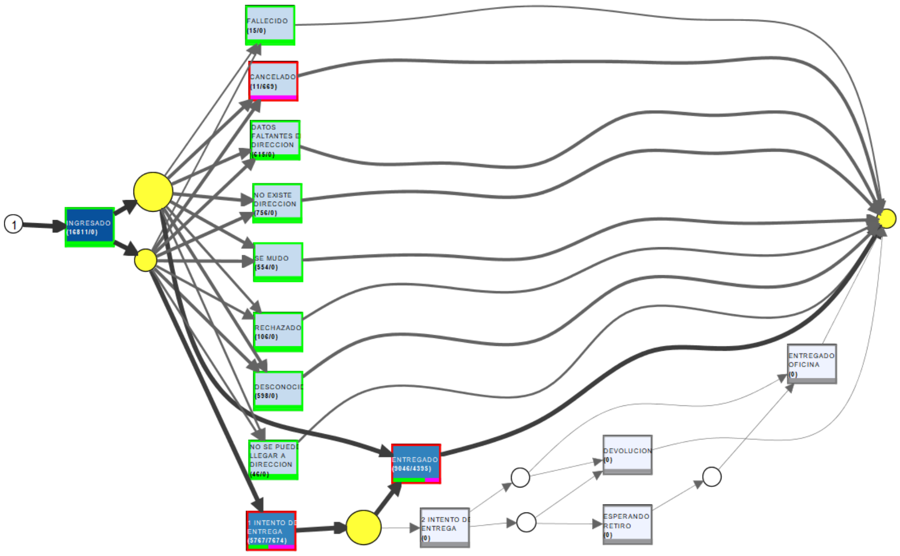


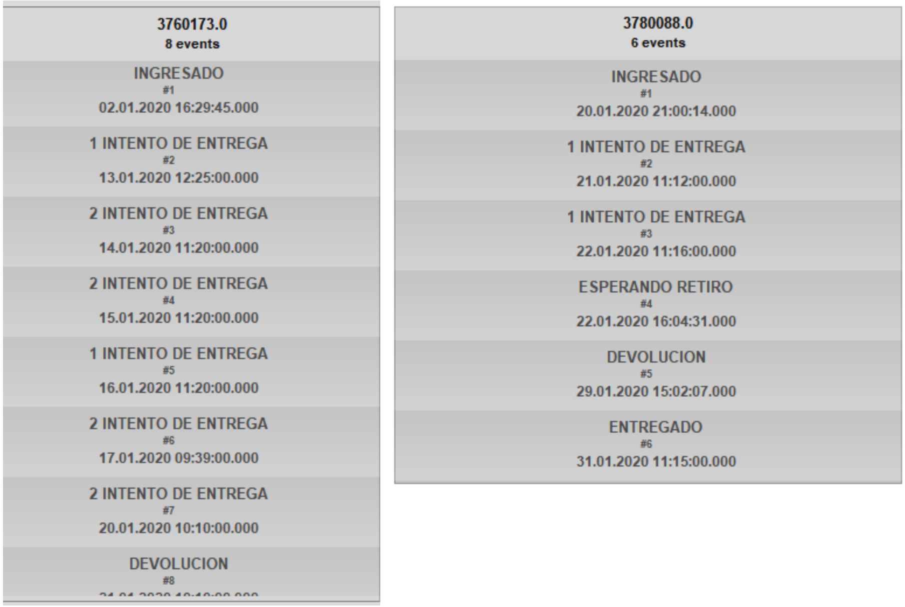
Fig. 3. Trace travel: most common events in dark and more frequent tours in thicker lines

as a representative value of an excessive number of movements. Based on this, we proceeded to filter and subsequently analyze the traces with more than 6 movements, identifying a total of 108 such cases.

The second case analyzed was that of the traces that did not exactly match the process discovered. On this occasion, the value of the compliance threshold was set at 0.5, meaning that traces with at least 50% of their movements misaligned with the process were considered to be highly deviated and required inspection. As previously stated, this can occur either because traces have events that are not part of the model or because the events in the trace do not follow the right sequence.

As a result of this, it was observed that these events were repetitions of events on different days or inconsistent records. Figure 4 illustrates both situations. The table on the left shows a case in which a first visit is recorded after a second visit; this situation can only be caused by a registration error. In the box to the right (same figure), an inconsistency is observed, since it is not possible to deliver a product that was previously returned.

Visual tools can also be used to generate animations that facilitate understanding these situations. In this particular case, with the set of traces whose compliance is below 50%, Inductive Visual Miner [9] was used to generate an animation that allows visualizing a timeline of the events for each trace. Figure 5 illustrates this animation. In this figure, each trace is symbolized with a token (or circle) that goes through the different stages of the process. This visual representation shows that some traces take longer than others and that sometimes, some erroneous behaviors occur. For example, in Figure 5, backward movements



3760173.0
8 events

INGRESADO
#1
02.01.2020 16:29:45.000

1 INTENTO DE ENTREGA
#2
13.01.2020 12:25:00.000

2 INTENTO DE ENTREGA
#3
14.01.2020 11:20:00.000

2 INTENTO DE ENTREGA
#4
15.01.2020 11:20:00.000

1 INTENTO DE ENTREGA
#5
16.01.2020 11:20:00.000

2 INTENTO DE ENTREGA
#6
17.01.2020 09:39:00.000

2 INTENTO DE ENTREGA
#7
20.01.2020 10:10:00.000

DEVOLUCION
#8
21.01.2020 10:10:00.000

3780088.0
6 events

INGRESADO
#1
20.01.2020 21:00:14.000

1 INTENTO DE ENTREGA
#2
21.01.2020 11:12:00.000

1 INTENTO DE ENTREGA
#3
22.01.2020 11:16:00.000

ESPERANDO RETIRO
#4
22.01.2020 16:04:31.000

DEVOLUCION
#5
29.01.2020 15:02:07.000

ENTREGADO
#6
31.01.2020 11:15:00.000

Fig. 4. Repeated events and inconsistencies

are observed in the traces; in particular, circles have been used to highlight 25 traces that go from a second attempt to an initial attempt, and 20 that go back on themselves.

Exporting traces that do not fit the model allows carrying out a detailed analysis of the reasons for deviations and thus identifying areas of improvement in the distribution process.

4 Conclusion and future lines of research

In this article, different process mining techniques have been used to analyze data based on the postal distribution of products in the Argentine Republic between the years 2017 and 2019. At the date of generation of this document, no similar works have been found that implement process mining to postal distribution in the Argentine Republic.

Through the model generated from select traces, the process that is actually carried out was discovered. The impact that the selection criteria of these traces has on the model obtained should be noted. The first models, generated from the entire set of traces, had excessive detail and this made representation and interpretation more difficult. Currently, work is being done to identify the most frequent traces from the initial model and then automatically filter the most representative ones.

For its part, compliance verification has allowed identifying anomalous situations of interest. Cases that did not comply with the model were detected,

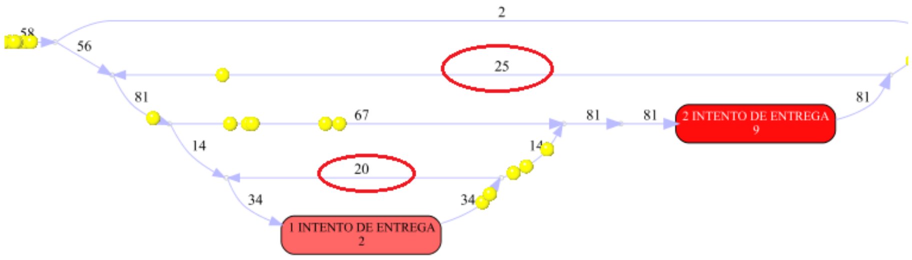


Fig. 5. Traces that do not comply with the model, in a circle the amount that goes backwards from a state instead of going forward (see direction of the arrow)

as well as cases that did follow it but outside the expected times or with task redundancy. Further analysis is required to determine if these cases represent manual load errors, and to look for a solution.

A fitting threshold was used to establish the minimum degree of distortion that a trace should meet so as not to affect the process. This factor should be analyzed in more detail to determine its value based on the case study at hand.

As a future line of work, we will continue working with process mining techniques not only to model situations that have already occurred, but also to be able to insert early warnings into the system when there are excessive deviations from the model.

References

1. Process Mining: Data Science in Action, Wil van der Aalst,978-3-662-49850-7,2016,Springer
2. Research of Postal Data mining system based on big data, Xia Hu1; Yanfeng Jin1; Fan Wang, 3rd International Conference on Mechatronics, Robotics and Automation, 2015, 10.2991/icmra-15.2015.124, https://www.researchgate.net/publication/300483008_Research_of_Postal_Data_mining_system_based_on_big_data
3. Context Aware Process Mining in Logistics, Mitchell M. Tseng; Hung-Yin Tsai; Yue Wang, 2017,The 50th CIRP Conference on Manufacturing Systems, https://www.sciencedirect.com/science/article/pii/S2212827117303311
4. PM2: a Process Mining Project Methodology,Maikel L. van Eck; Xixi Lu; Sander J.J. Leemans; Wil M.P. van der Aalst,Eindhoven University of Technology, The Netherlands, http://www.processmining.org/_media/blogs/pub2015/pm2_processminingprojectmethodology.pdf
5. Wil van der Aalst, The Process Mining Manifesto by the IEEE Task Force, 2012, https://www.tf-pm.org/resources/manifesto
6. Workflow mining: discovering process models from event logs, W. van der Aalst; T. Weijters; L. Maruster, 1041-4347, 2004, IEEE, https://ieeexplore.ieee.org/document/1316839

7. Process mining in flexible environments, Christian Walter Gunther,978-90-386-1964-4,2009,Technische Universiteit Eindhoven, https://research.tue.nl/en/publications/process-mining-in-flexible-environments
8. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams, IEEE Std 1849-2016, 2016, DOI 10.1109/IEEESTD.2016.7740858
9. inductive visual miner, Sander J.J. Leemans, 2017, http://leemans.ch/leemansCH/publications/ivm.pdf
10. Reinventing the Postal Sector in an Electronic Age, Michael A. Crew; Paul R. Kleindorfer,978-1849803601,2011,Edward Elgar Publishing