Distribution Analysis of Postal Mail in Argentina Using Process Mining

Victor Martinez[1,4][0000−0003−3581−5704], Laura Lanzarini[2][0000−0001−7027−7564], and Franco Ronchetti[2,3][0000−0003−3173−1327]

[1] School of Computer Science, National University of La Plata, Argentina
[2] Computer Science Research Institute LIDI (III-LIDI) (UNLP-CICPBA)
[3] Scientific Research Agency of the Province of Buenos Aires (CICPBA)
[4] Corresponding author

Abstract. Process mining combines a number of techniques that allow analyzing business processes solely through event logs. This article is a continuation of the research carried out in [1] to analyze data based on the postal distribution of products in the Argentine Republic between the years 2017 and 2020. The results obtained initially showed that 85% of the shipments made comply with the process correctly. Cases that did not fit within the model were also quickly identified, and recurring problems were found, which facilitates analysis for process improvement. The most common problems were traces that do not follow task order, excess movements or missing movements, and traces that comply with the process but take too long. In this article, a performance analysis was added to discover traces that, despite correctly following to the process, have operational deviations due to an excessive time to complete. These techniques are intended to be added to the process through early alerts that warn about the existence of such situations, which would help improve service quality.

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

1 Introduction

Currently, there is a wide variety of information systems that support the various areas in each company, both business-related and administrative. These tools include ad hoc developments for each company and management tools such as CRM, ERP, WMS, TMS, BI, etc.

Nowadays, storing data can be done easily and with a low cost, so most of these systems save information in text files or databases for auditing purposes or for error resolution. Among other things, it logs what was done, when it was done, who did it, if there were any errors, etc. Analyzing this data, insights can be obtained about errors or problems that have occurred, the business process carried out, and how to propose improvements or find solutions to various operational issues.

Process mining provides techniques that allow these logs to be analyzed to obtain insights.

Process mining can answer questions such as: How can process control be improved? What actually happened? Why did it happen? What could happen in the future? How can the process be improved to increase performance? [2].

By working directly with productive data, the real behavior that is carried out for the business and the complete process that is being carried out are obtained, which in some cases may differ from the one originally designed.

The case study in this work is the distribution of products by postal mail. It is a process in which different movements are logged, made up of events such as receiving a product at a branch, its entry into a distribution center, internal movements and the various delivery attempts. All tasks are clearly defined and must be performed in a certain order and within a given time interval. In daily operations, deviations occur due to delays in the execution of tasks or inconsistency (repeated tasks or tasks carried out in an incorrect order).

There are some jobs that link the postal business, process mining and big data, such as [3], where data mining is applied to China mail in a big data environment. Given the complexity of the postal business, clustering techniques were used to group customers based on consumption habits, main interests and behavior, so as to achieve a more effective and accurate marketing strategy. The results were very satisfactory. Another example is [4], where process mining is applied to a logistics and manufacturing chain to look for the similarities and differences between various delivery processes in a changing environment. To do this, different processes are compared using clustering techniques to automate process documentation. Finally, in [5], a methodology that can be used as a guide in process mining projects is shown, and the case study of its application in IBM is discussed as an example.

This work continues the research published in [1] in relation to the application of process mining techniques to postal distribution in Argentina to analyze its operation, identify deviations or issues in distribution, and propose enhancements that will help improve service quality. To add to this previous work, a performance analysis has been added that allows identifying a new type of operational deviations. The data analyzed correspond to postal mail distribution in Argentina between 2017 and 2020. It should be noted that, at the date of publication of this article, the authors are not aware of the existence of any other works with these characteristics for the postal business.

2 Process Mining

Process Mining operates on the log files of the information system to be analyzed. It begins with the extraction of the information of all the events of a certain activity and then, through an automatic analysis, the relevant business process is identified, as well as its activities and the sequence that has to be followed. This allows analyzing and answering questions such as: What happened? Why did it happen? What could happen in the future? How can control be improved?

Can performance be improved? [2].With the information obtained after model interpretation, different types of analysis can be carried out, such as who carries out the activity, whether the activity is complete and if it is carried out within a reasonable time interval. Process mining can be classified into three types [6]:

- *Discovery*: These are techniques capable of modeling the business process that is being carried out solely from the sequence of corresponding logs, which can be extracted from files, databases or some other media. Possibly the most popular algorithm to perform this task is the Alpha algorithm [7], which generates a Petri net with the discovered process.
  The quality of the generated model is directly proportional to the quality of the input data. This is a feature common to any inductive knowledge extraction technique. It is essential that the event sequences surveyed fully represent the process to be modeled; otherwise, there will be aspects that will remain undiscovered.
- *Compliance Verification*: It allows contrasting a real sequence of events against a process model (it can be the one previously discovered or a different one) to determine the compliance level and which are the occurrences that deviate from the process.
  Applications that graphically represent the results and perform animations can be used to liven up the presentation of the results. The resulting analysis can be very accurate, since it is based on real data rather than being a simulation.
- *Improvement*: Its goal is improving the existing process based on the analysis carried out. It differs from Compliance Verification in that it focuses on the process and not on trace execution.

There are four other aspects to take into account to obtain quality results in process mining. These are accuracy, fit, generalization, and simplicity [2].A balanced relationship between these four forces has to be maintained so that the discovered model is representative of the process and easy to understand.

Last but not least, it should be noted that input data usually carry a large amount of noise, which can be due to data duplication or incomplete traces that can distort analysis results [8].As in other mining techniques, pre-processing is usually carried out to improve input data quality by removing incomplete or corrupt data that could lead to errors in modeling.

3 Process Mining Applied to Postal Distribution

The discovery of the process will be carried out using data from the postal distribution in Argentina between 2017 and 2020.

Postal mail can be used to distribute various products, such as letters, telegrams, parcels or, less traditionally, e-commerce products.

In all cases, during the process, at least the following steps take place: registration and reception of the product to be distributed, internal routing through one or more sites or distribution centers, one or more delivery attempts, effective delivery or return to sender if delivery was not successful. Each step carried out is always logged, including information about the person responsible for the action, what they did, and when they did it. This record is associated with a unique shipping identifier that allows these events to be reported to the customer.

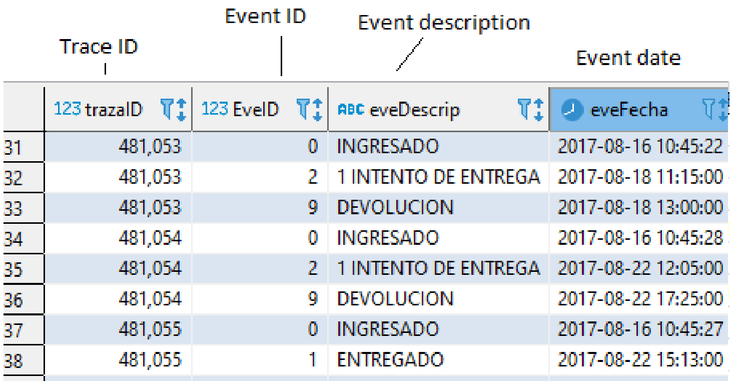
3.1 Data Extraction

To discover the model and then analyze the events, data will be extracted and pre-processed. For this particular case study, product shipments that required two delivery attempts were used. The procedure in this case is as follows: the product goes to distribution, if it cannot be delivered for any reason that is not final, a new delivery is attempted the next day. If it cannot be delivered once again, a period of time is given to the recipient to come and pick it up (a visit notice is left at each visit). Once the established period of time has passed, if the recipient has not collected the shipment, it is returned to the sender.

Each shipment is considered as a trace and each movement as an event. The 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 data collection, a sample containing around 33,000 traces with more than 77,000 events was generated (see Figure 1). Each trace has at least two associated events: a trace identifier and an identifier for each event. In particular, the necessary fields for analysis were recorded for all the events in the case study: trace identifier, event identifier within the trace, description, and event date.

With this information, each trace can be rebuilt and the corresponding model put together.



	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

Data are then converted to XES format [9].XES proposes a tag-based language that provides a unified and extensible methodology to log behaviors in information systems [9].Data structured in this way can be processed with a wide variety of tools in an efficient way.

To increase data consistency and facilitate the analysis, all incomplete traces — either because they did not have an initial status (entry) or a final status (delivered, returned, no address exists, died, or moved) — were removed. This may have been due to an error when loading the data or because the traces have not yet completed the process.

To remove incomplete traces, a simple heuristic rule filter was used, where the initial state and the valid final states were specified. The filter discards any trace that does not meet these requirements. After applying the filter, a sample of approximately 16,000 traces with 43,000 events was obtained.

3.2 Process Discovery

The Alpha algorithm was used for process discovery. The Alpha algorithm was proposed by van der Aalst, Weijters and Maruster [7], and is widely used in process mining. The algorithm rebuilds causality from a sequence of events and returns a Petri net where the business process used is reflected. Each transition in the network represents a task.

For this particular case study, a small representative sample is extracted that will be used to discover the process; then, the discovered process is contrasted with the rest of the traces to observe compliance.

Figure 2 shows the discovered Petri net that represents the process that the traces must follow.

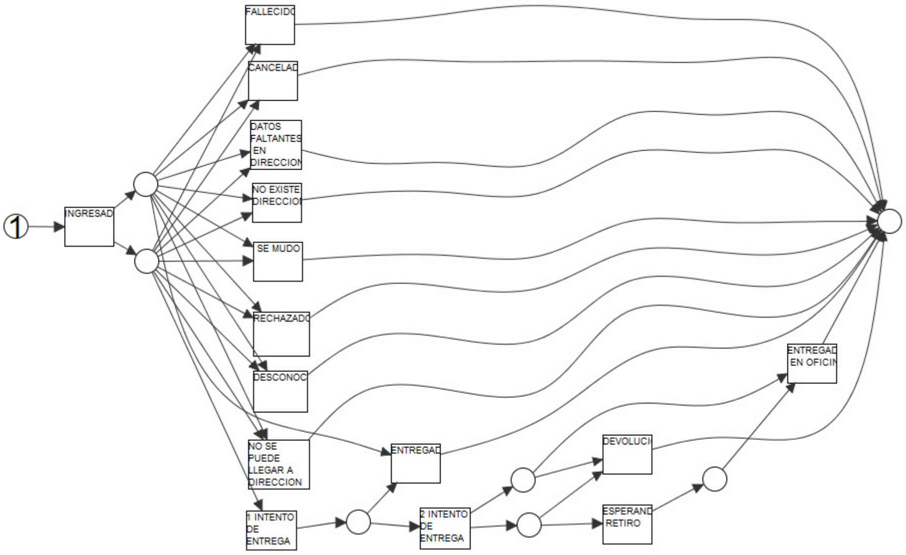


Fig. 2. Process discovered that all traces must fulfill

The discovered process is made up as follows: there is always an entry event that all the traces have to complete, which is followed by the distribution step. If delivery is possible, it is logged and the history of the trace ends. If it is not possible due to a final event (deceased person, the address does not exist, there is not enough data to find the address, etc.), a non-delivery event is logged with the relevant reason and the process ends. If the reason for non-delivery is not a final event (for example, there was no one to receive the parcel), a first delivery attempt is logged and another attempt is made the following day. If on the following day it cannot be delivered once again, a period of time is given to the recipient to go to the office and pick it up (a visit notice is left at each visit). Finally, the delivery event is logged as delivery in the office or the shipment is returned to the sender. Figure 2 illustrates the scenarios of the discovered process.

3.3 Model Verification

To check compliance, all sample traces are taken (a total of 16,811 traces with 43,888 events) and they are compared to the discovered model.

The aim is to determine how much the traces fit the model and analyze those that deviate the most, taking into account which steps are fulfilled, which are not, and if there are steps that are not reflected in the discovered process. The result is shown in Figure 3, with the most common events highlighted in a dark color and the most frequent paths 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 (either delivered or not) fully comply with the discovered process, with a mean of 2.6 events in each trace.

It can also be seen that many traces do not follow the process exactly, either because they do not complete the steps in the correct order or because they skip some step. The more they deviate from the model, the lower the match to the process; therefore, those traces with a fit of less than 50% will be analyzed, since this is considered to be a large operational deviation.

With these results, two different types of analyses can be carried out: 1) traces that have too many movements, and 2) traces whose tasks are not carried out in the correct order or have missing events.

Additionally, those traces that correctly follow the process but do so in an excessively long time interval will also be analyzed.

Given that the number of movements per trace follows a normal distribution with a mean of 2.6 and a deviation of 0.95, the value of the mean plus two standard deviations was used as a representative value of an excessive number of movements.

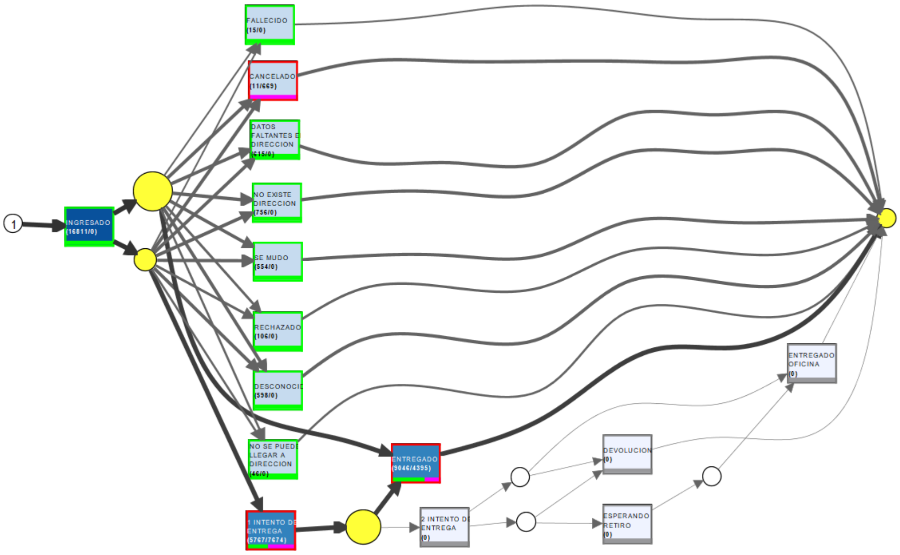


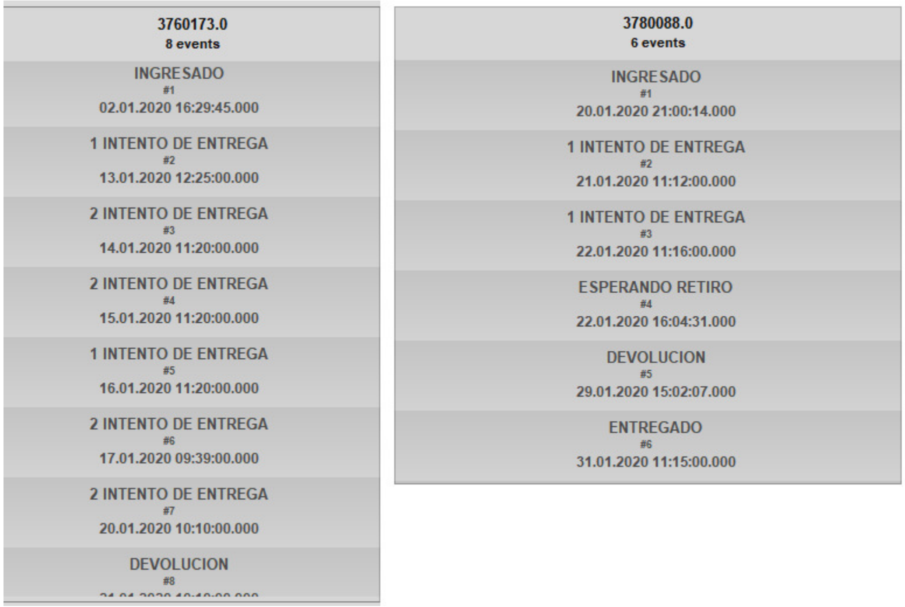
Fig. 3. Trace paths: most common events highlighted in a dark color and most frequent paths represented with a thicker line.

Based on this, in the first case, traces with more than 6 movements were filtered and exported for further analysis. A total of 108 cases with these characteristics were found.

In the second case, the traces that did not exactly match the process discovered were analyzed. For this, an adjustment value of 0.5 was used, considering that traces have at least 50% of their movements outside of the model, representing a large deviation from the standard procedure. As previously mentioned, this can happen because there are too many events in the traces (they do not exist in the model) or they do not correctly follow the sequence.

In this case, it was observed that the same event is repeated on successive days or there are inconsistent logs. Figure 4 shows these situations. On the left side, a case with a first visit after the second is shown. This is impossible, meaning it is probably due to a logging error (for example, the wrong date was used in one of the movements). On the right side, in that same figure, an inconsistency is shown where a delivery movement is logged after the product was returned.

Additionally, the Inductive Visual Miner tool[10] was used to generate a visual animation and represent these cases in a more friendly way. For this, the set of traces with a match rate of less than 50% with the process was used, obtaining the visualization of the traces that deviate the most. Figure 5 illustrates this animation. In this figure, each trace is represented with a circle that goes through the different stages of the process. This visual representation shows that some traces take much longer than others. It also shows that, sometimes, there are incorrect behaviors. For example, in Figure 5, traces go back, i.e., after a task



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

is completed, a previous task is carried out; circles are used to highlight that 25 traces go from the second delivery attempt to the first one, and 20 repeat the same step.

The export of traces that do not follow the model shows the reasons behind the deviations, which allows carrying out a more detailed analysis to identify potential improvements to the process.

3.4 Performance Analysis

Finally, a performance analysis is carried out on the traces in the model to identify those that, although they may have completed the process correctly, did so in a much longer time than the mean resolution time. To do this, the Petri net data sample was replicated using a tool oriented to performance analysis that focuses on resolution times and not so much on seeing whether the traces follow the process or not.

As a result, a visualization is obtained that represents in darker tones the tasks that take more time to complete.

A mean resolution time for each trace of 3.75 days is observed, with a deviation of 2.81 days. Using the same criteria as with the number of movements per trace, the analysis now focuses on those cases that have a completion time greater than the mean plus two deviations (9.37 days), considering that a resolution time greater than this value is an operational diversion.

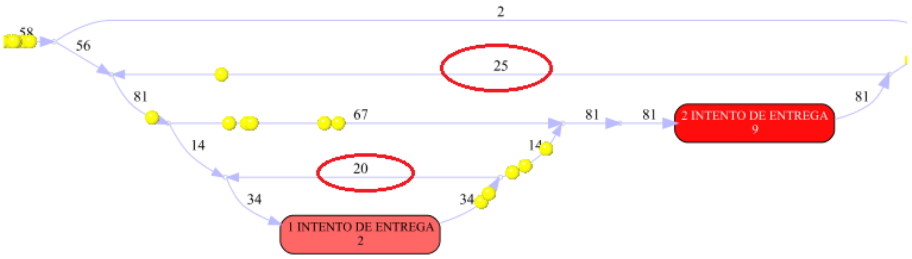


Fig. 5. Traces that do not follow the model; those that go back to a previous state instead of moving forward to the next are circled (see arrow direction)

The traces that meet this criterion are exported, obtaining a total of 687 cases.

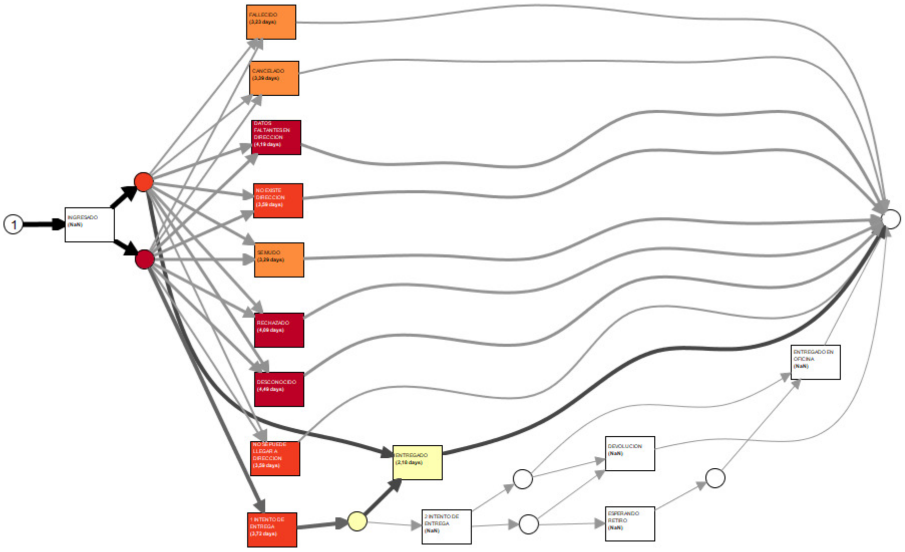
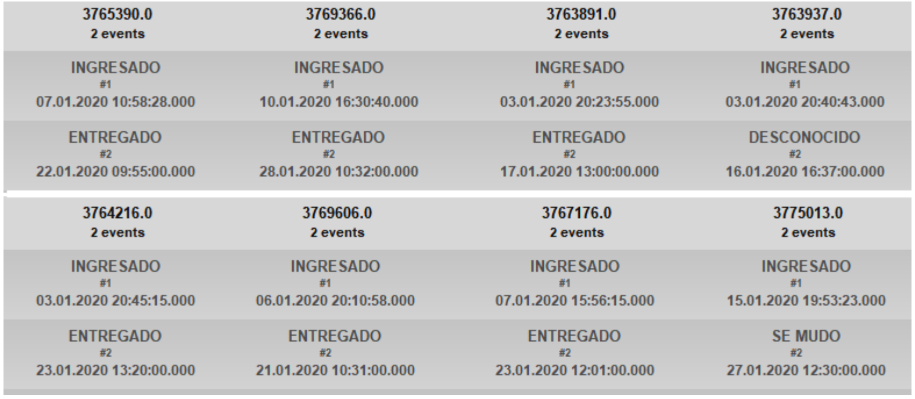


Fig. 6. Performance analysis of sample traces

Among these, there are traces that perfectly follow the process, but with extremely long times. In some cases, traces have more than three movements and an interval of up to 3 days between them, which explains such a long overall time. Even though 3 days between movements is longer than the mean, this could be due to situations that may occur on a day-to-day basis, such as lack of staff due to illness or vacation. On the other hand, most of the cases in this sample have only two movements, where entry was logged and then a period of up to 14 days passed before an exit to distribution event was logged. After this, the process is successfully completed with a delivery or return movement. These cases represent a significant operational deviation, since they directly impact the quality of service by failing to meet product delivery standards.

Figure 7 shows traces that complete the process with only two movements but with a difference of more than 10 days between them.



3765390.0 2 events	3769366.0 2 events	3763891.0 2 events	3763937.0 2 events
INGRESADO #1 07.01.2020 10:58:28.000	INGRESADO #1 10.01.2020 16:30:40.000	INGRESADO #1 03.01.2020 20:23:55.000	INGRESADO #1 03.01.2020 20:40:43.000
ENTREGADO #2 22.01.2020 09:55:00.000	ENTREGADO #2 28.01.2020 10:32:00.000	ENTREGADO #2 17.01.2020 13:00:00.000	DESCONOCIDO #2 16.01.2020 16:37:00.000
3764216.0 2 events	3769606.0 2 events	3767176.0 2 events	3775013.0 2 events
INGRESADO #1 03.01.2020 20:45:15.000	INGRESADO #1 06.01.2020 20:10:58.000	INGRESADO #1 07.01.2020 15:56:15.000	INGRESADO #1 15.01.2020 19:53:23.000
ENTREGADO #2 23.01.2020 13:20:00.000	ENTREGADO #2 21.01.2020 10:31:00.000	ENTREGADO #2 23.01.2020 12:01:00.000	SE MUDO #2 27.01.2020 12:30:00.000

Fig. 7. Traces that follow with the process but take a very long time to complete

The performance analysis helped uncover new operational deviations that had not appeared in the previous analysis. These are traces that perfectly follow the model but whose completion times are well above the mean, and sometimes can be as long as 24 days before delivery. Based on this study, it was decided to take a sample of the 10 longest cases and ask the heads of the corresponding sector/branch about them, in an attempt to gather information about the reasons behind these delays.

The performance analysis helped identify a need to generate an alarm that brings attention to these cases after a certain time has passed, which would help improve quality of service.

4 Conclusion and future lines of research

In this article, different process mining techniques have been used to analyze real data based on the postal distribution of products in Argentina between 2017 and 2020. At the date of generation of this document, no similar investigations have been found that analyze this same case study.

After generating the model from a representative sample of traces, the process that actually takes place was discovered. Subsequently, this model was contrasted with all traces. Through a compliance check, unusual situations that resulted in operational errors were identified. Cases that did not comply with the model were detected, as well as cases that presented task redundancy.

The performance analysis allowed discovering traces that correctly follow the process but have operational deviations due to an excessive time to complete. These cases have a great impact on service quality.

As a future line of work, process mining techniques will continue to be used in an attempt to find a way to insert early warnings into the system aimed at avoiding deviations and operational bottlenecks.

References

[1] Victor Martinez; Laura Lanzarini; Franco Ronchetti. "Process Mining Applied to Postal Distribution". In: *CACIC 2021* (2021). URL: http://sedici.unlp.edu.ar/handle/10915/130342.

[2] Wil van der Aalst. *Process Mining: Data Science in Action*. 1st. Springer, 2016. ISBN: 978-3-662-49850-7.

[3] Xia Hu1; Yanfeng Jin1; Fan Wang. "Research of Postal Data mining system based on big data". In: *3rd International Conference on Mechatronics, Robotics and Automation* (2015). URL: https://www.researchgate.net/publication/300483008_Research_of_Postal_Data_mining_system_based_on_big_data.

[4] Mitchell M. Tseng; Hung-Yin Tsai; Yue Wang. "Context Aware Process Mining in Logistics". In: *The 50th CIRP Conference on Manufacturing Systems* (2017). URL: https://www.sciencedirect.com/science/article/pii/S2212827117303311.

[5] Maikel L. van Eck; Xixi Lu; Sander J.J. Leemans; Wil M.P. van der Aalst. "PM2: a Process Mining Project Methodology". In: *Eindhoven University of Technology, The Netherlands* (2017). URL: http://www.processmining.org/_media/blogs/pub2015/pm2_processminingprojectmethodology.pdf.

[6] Wil van der Aalst. "The Process Mining Manifesto by the IEEE Task Force". In: *IEEE Task Force* (2012). URL: https://www.tf-pm.org/resources/manifesto.

[7] W. van der Aalst; T. Weijters; L. Maruster. "Workflow mining: discovering process models from event logs". In: *IEEE* (2004). URL: https://ieeexplore.ieee.org/document/1316839.

[8] Christian Walter Gunther. "Process mining in flexible environments". In: *Technische Universiteit Eindhoven* (2004). URL: https://research.tue.nl/en/publications/process-mining-in-flexible-environments.

[9] IEEE Std 1849-2016. "IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams". In: *IEEE* (2016). URL: DOI%2010.1109/IEEESTD.2016.7740858.

[10] Sander J.J. Leemans. "inductive visual miner". In: (2017). URL: http://leemans.ch/leemansCH/publications/ivm.pdf.