BPI Challenge 2020

Mohammadamin Gholami

Business Information Systems
Prof. Paolo Ceravolo
a.a. 2021 - 2022

Contents

1 Case study

Many organizations, including Eindhoven University of Technology (TU/e), have policies in place to reimburse employees for their travel expenses incurred while on work-related trips. This may include travel to customers, conferences, or project meetings. TU/e has established clear procedures for arranging and reimbursing travel expenses, making the process straightforward for employees.

The university categorizes travel into two types: domestic and international. Domestic trips do not require prior approval, and employees can simply file a claim for reimbursement of costs incurred. However, international trips require approval from a supervisor, which is obtained through the submission of a travel-permit. This permit must be approved before any travel arrangements can be made.

In terms of reimbursement, employees have the option to file a claim as soon as they have paid for trip expenses, such as flight tickets or conference fees. Alternatively, they can also submit a claim within two months after the trip for expenses such as hotel and food costs, which are usually paid on the spot. This flexibility ensures that employees are promptly reimbursed for their travel expenses.

The actors are the employees and the organization (Eindhoven University of Technology (TU/e)) they work for.

The following questions are of interest:

- What is the throughput of a travel declaration from submission (or closing) to paying?
- Is there are difference in throughput between national and international trips?
- Are there differences between clusters of declarations, for example between cost centers/departments/projects etc.?
- What is the throughput in each of the process steps, i.e. the submission, judgement by various responsible roles and payment?
- Where are the bottlenecks in the process of a travel declaration?
- Where are the bottlenecks in the process of a travel permit (note that there can be multiple requests for payment and declarations per permit)?

- How many travel declarations get rejected in the various processing steps and how many are never approved?

Then there are more detailed questions

- How many travel declarations are booked on projects?
- How many corrections have been made for declarations?
- Are there any double payments?
- Are there declarations that were not preceded properly by an approved travel permit? Or are there even declarations for which no permit exists?
- How many travel declarations are submitted by the traveler and how many by a mandated person?
- How many travel declarations are first rejected because they are submitted more than 2 months after the end of a trip and are then re-submitted?
- Is this different between departments?
- How many travel declarations are not approved by budget holders in time (7 days) and are then automatically rerouted to supervisors?
- Next to travel declarations, there are also requests for payments. These are specific for non-TU/e employees. Are there any TU/e employees that submitted a request for payment instead of a travel declaration?

After downloading the datasets and import them using pm4py the following table represents the data in more details.

The data has been gathered from the TU/e reimbursement process. The records include information from two departments in 2017 and from the entire TU/e in 2018.

File	Cases
DomesticDeclarations.xes	10500
InternationalDeclarations.xes	6449
PermitLog.xes	7065
PrepaidTravelCost.xes	2099
RequestForPayment.xes	6886

In this study, we will limit our examination to only two datasets, which are the "domestic_declarations" and "international_declarations". This decision was made due to the lack of valuable information in the other three datasets, and we believe

that focusing on these two datasets will provide the most valuable insights and results.

2 Organisational goals

The purpose of the research is to examine the data that has been provided with the intention of gaining meaningful information. The organization can utilize the outcome of this analysis to advance towards their desired goals, upgrade the performance of their processes, remove any barriers that may be hindering progress, and take into account any anomalies found in the data. Furthermore, the research aims to provide answers to important questions that need to be addressed.

With the use of process mining techniques such as filtering, process discovery, conformance checking, and variant analysis with visualization, it is hoped to accomplish the previously stated goal. These techniques will allow for a comprehensive analysis of the data, leading to a deeper understanding and the identification of areas that can be improved upon. By utilizing these process mining techniques, the organization will be better equipped to reach their objectives, optimize their processes, eliminate bottlenecks, and address any data anomalies.

3 Knowledge Uplift Trail

1. Log extraction: The first step of the analysis involves transforming the raw input file logs into an event log by importing them into pm4py. This step is crucial in preparing the logs for further analysis.

2. Variant and data analysis: In this step, different variants of the case study are recognized and discussed. This process helps in understanding the different types of processes that exist in the event log. The event log is then converted into variants and their distributions, which allows for a deeper understanding of the data. Also some useful

information and answer to some essential questions has done in this step

3. Filtering: The third step of the analysis involves removing any noisy data or unused logs from the dataset. This is done to increase the precision of the analysis and to eliminate any irrelevant data that could skew the results. The input for this step is the event log, and the output is the filtered event log.

4. Process discovery: Using a variety of algorithms and methods, diagrams or graphs that provide useful insights into the process will be generated with their quality metrics (fitness, precision, simplicity, generalization). The filtered event log serves as the input for this step, and the output is the extracted model. This step is crucial in understanding the underlying structure of the process and identifying potential areas for improvement.

5. Conformance checking: This step involves evaluating the performance of the generated models and comparing them. The input for this step is the model created in the previous step, and the outputs include key factors of the assessment. This step helps in determining whether the generated models are in compliance with the expected behavior and if they can be used to improve the process.

Steps	Input	Analytic	Model	Output
Step 1	File	Extraction	Descriptive	Event log
Step 2	Event Log	Variant Analysis	Descriptive	Variant
Step 3	Data frame	Filtering	Descriptive	Filtered log
Step 4	Filtered log	Process Discovery	Prescriptive	Model
Step 5	Model	Conformance checking	Prescriptive	Metrics assessment

4 Project Results (Domestic Declarations)

4.1 Extract data

To address some of the questions posed by the challenge, variant analysis is utilized on the event log of domestic declarations. This analysis allows us to search for variants and provide answers to some of the questions. In terms of the number of cases, there are a total of 10,500 cases where no double payment occurred. Out of these cases, 1166 were initially rejected due to being submitted more than two months after the end of the trip and were then re-submitted. A total of 10044 cases were handled and out of these, 1301 cases had at least one rejection. There were 365 cases that were never approved, while 10070 cases received final approval from a supervisor.

All of the activities recorded in the domestic declarations dataset is displayed in the table below.

Activities	Count
Declaration SUBMITTED by EMPLOYEE	11531
Declaration FINAL_APPROVED by SUPERVISOR	10131
Request Payment	10040
Payment Handled	10044
Declaration APPROVED by PRE_APPROVER	685
Declaration REJECTED by MISSING	91
Declaration REJECTED by PRE_APPROVER	86
Declaration REJECTED by EMPLOYEE	1365
Declaration SAVED by EMPLOYEE	135
Declaration REJECTED by SUPERVISOR	293
Declaration APPROVED by ADMINISTRATION	8202
Declaration APPROVED by BUDGET OWNER	2820
Declaration FOR_APPROVAL by SUPERVISOR	1
Declaration REJECTED by ADMINISTRATION	952
Declaration FOR_APPROVAL by PRE_APPROVER	1
Declaration REJECTED by BUDGET OWNER	59
Declaration FOR_APPROVAL by ADMINISTRATION	1

Later on, a deeper analysis of the bottlenecks in the process will be conducted. To do this, we will employ various process mining techniques, including filtering, variant analysis, process discovery, and conformance checking. These techniques will help us gain a better understanding of the process and identify areas for improvement.

All the cases in the domestic declarations started with two events: 'Declaration SUBMITTED by EMPLOYEE' with a total of 10365 occurrences, and 'Declaration SAVED by EMPLOYEE' with 135 occurrences.

4.2 Variant analysis

After extracting the variants from the data, we found that there were many cases with a low frequency, having a variant count less than or equal to 3. The goal was to focus on the more frequently occurring variants to find the better process. Despite filtering out the low frequency variants, the top 5 variants still make up a substantial number of our cases, and they are as follows:

index	variant	count
0	Declaration SUBMITTED by EMPLOYEE,Declaration APPROVED by ADMINISTRATION,Declaration FINAL_APPROVED by SUPERVISOR,Request Payment,Payment Handled	4618
1	Declaration SUBMITTED by EMPLOYEE,Declaration APPROVED by ADMINISTRATION,Declaration APPROVED by BUDGET OWNER,Declaration FINAL_APPROVED by SUPERVISOR,Request Payment,Payment Handled	2473
2	Declaration SUBMITTED by EMPLOYEE,Declaration FINAL_APPROVED by SUPERVISOR,Request Payment,Payment Handled	1392
3	Declaration SUBMITTED by EMPLOYEE,Declaration APPROVED by PRE_APPROVER,Declaration FINAL_APPROVED by SUPERVISOR,Request Payment,Payment Handled	575
4	Declaration SUBMITTED by EMPLOYEE,Declaration REJECTED by ADMINISTRATION,Declaration REJECTED by EMPLOYEE,Declaration SUBMITTED by EMPLOYEE,Declaration APPROVED by ADMINISTRATION,Declaration FINAL_APPROVED by SUPERVISOR,Request Payment,Payment Handled	345

The diagram below shows the frequency distribution of the variants before filtering was applied.
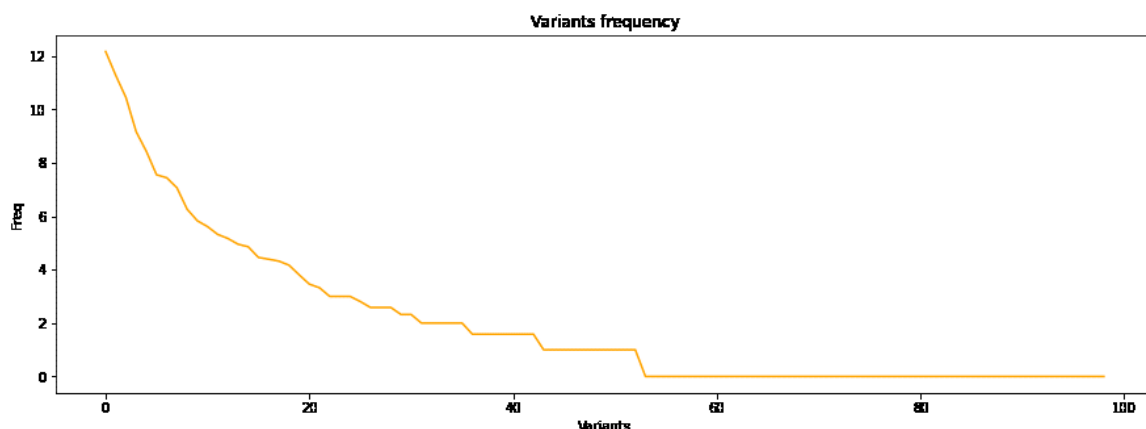
4.3 Filtering

The filtering process was implemented in order to refine the list of cases under consideration. The filter was designed to only include cases where the process started with the submission by the employee. This step effectively decreased the total number of cases from an original larger pool to 10365.

Subsequently, an additional filtering process was applied to focus on the top 6 variants, as it was found that the majority of cases had been handled using these variants. This further reduction in the number of cases brought the total number down to 9591. The purpose of these filters was to simplify the data set and focus on the most relevant cases for analysis.

This is an illustration of the variants that were obtained after the filtering process was executed. The purpose of the filtering was to refine the data set and focus on the most relevant information.
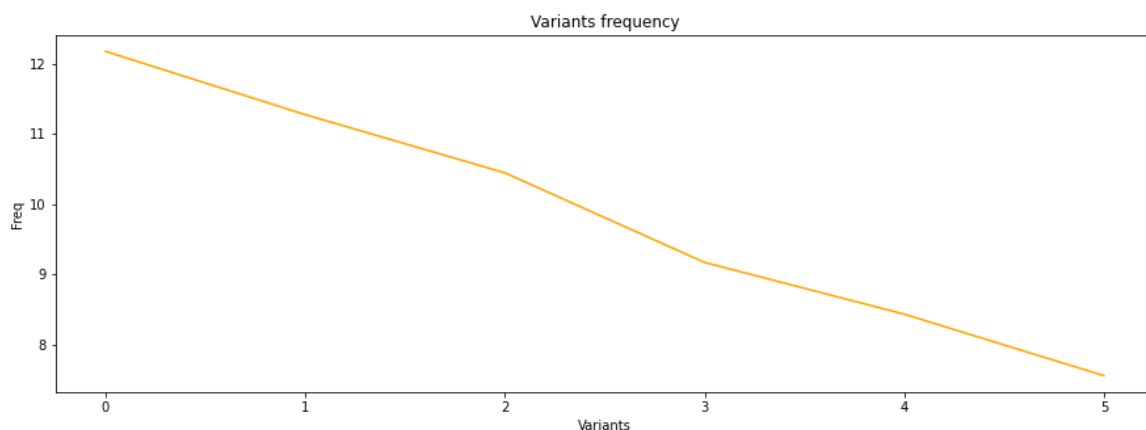

In addition to this representation of the filtered variants, the mean and median of the distribution were calculated. The mean, which is a measure of the average value of the data, was found to be 106.06060606060606. On the other hand, the median, which is the middle value in the data set, was calculated to be 2.

4.4 Process discovery

To thoroughly analyze the processes within the system and create an accurate data model, it is essential to utilize various process discovery methods.

4.4.1 Alpha miner

The Alpha Miner algorithm is a widely used method in the field of process mining that aims to uncover the underlying processes within a system. This algorithm is applied to data sets to analyze and model the behavior of business processes. In the application of Alpha Miner, the frequency of events is taken into consideration, meaning that the number of times each event occurs is taken into account when constructing the process model. This helps to create a more representative picture of the system and provides a more accurate understanding of the processes at play.
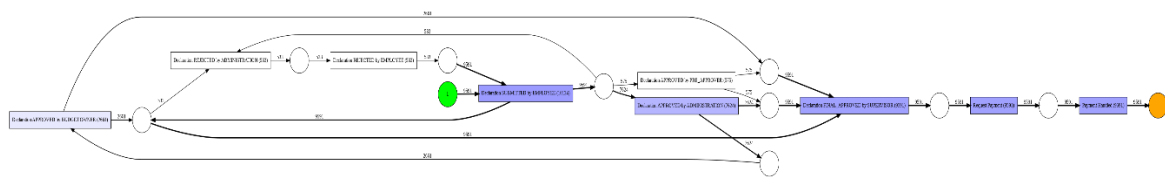

4.4.2 Inductive miner

the inductive miner works by repeatedly finding a split in the event log between within the trace– how to split it. Then detects the operator that describes the splits, and then continues on the sublogs. We also applied the inductive miner on our example data, and this was the Petri net that was discovered.
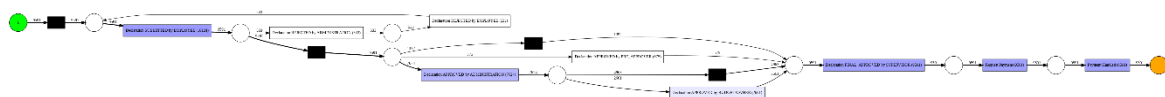
The Process Tree, is a tree-like structure that represents the control flow of a process. It consists of nodes, which represent activities, and edges, which represent the control flow between activities.
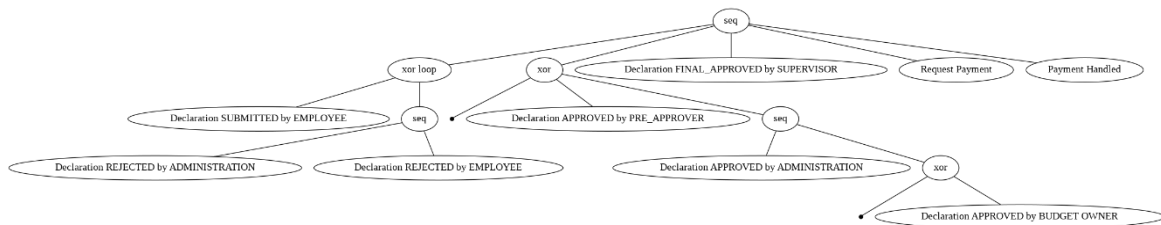
4.4.3 Heuristic miner

The Heuristic Miner in PM4Py provides a simple and intuitive way to extract process models from event logs. It uses a set of heuristics to identify patterns and relationships in the event log and to eliminate noise and irrelevant information. The resulting process model is a process tree, which is a tree-like structure that represents the control flow of a process.

The threshold is 0.99
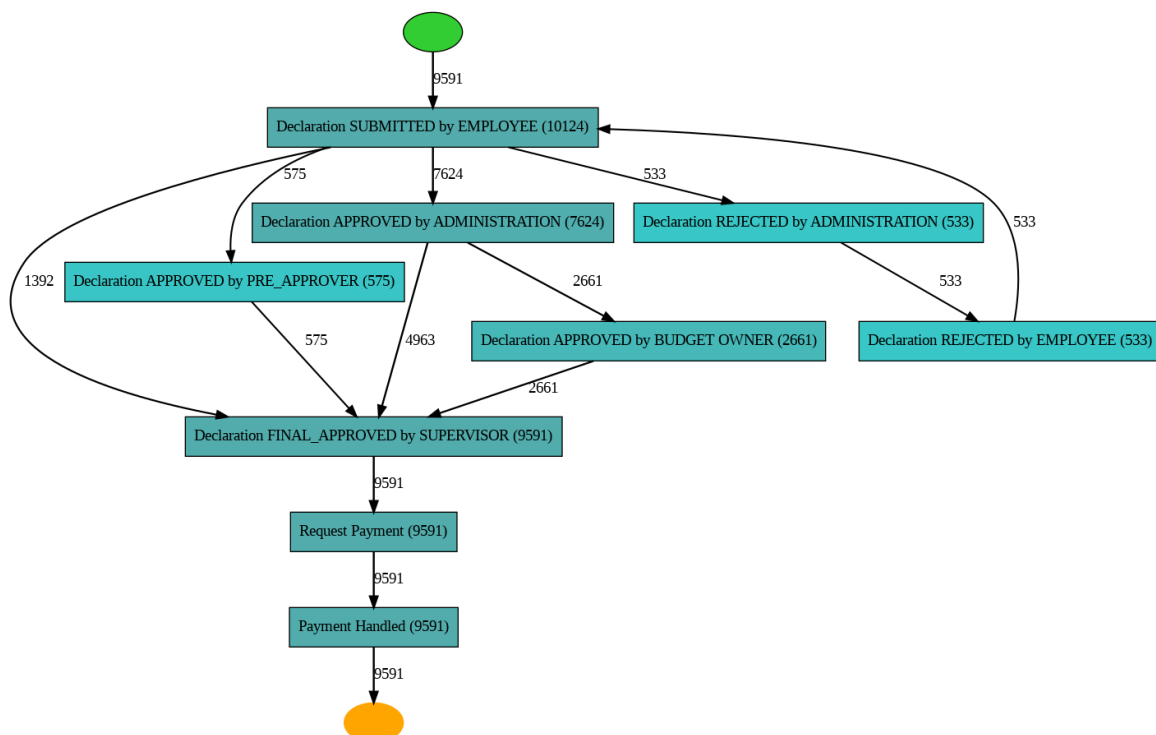
And also using Petri net depicts bellow

4.4.4 DFG

A Direct Follow Graph in process mining is a diagram that displays the interactions between different tasks in a process. It shows the frequency of each task being performed and the sequence in which they occur. The graph is made up of nodes, which represent the tasks, and directed edges, which show the flow of control between the tasks. The weight of each edge is proportional to how often the flow occurs. This graph is useful in analyzing process behavior and identifying any issues, such as bottlenecks or inefficiencies, in the process. It can also be used to compare different process versions and monitor performance over time.
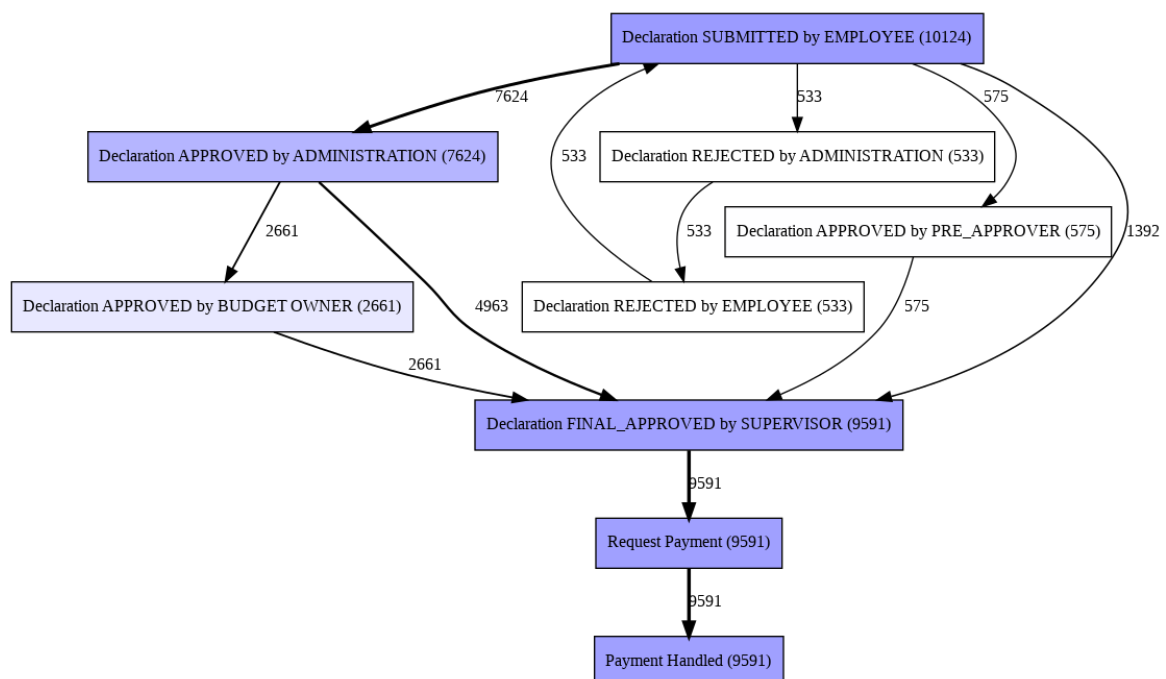
4.5 Quality metrics

The quality metrics (Precision, Fitness, Generalization, and Simplicity) for the algorithms are evaluated using Pm4Py in the corresponding section of the code. The results of these evaluations are displayed in a table below.

Model	Fitness	Precision	Generalization	Simplicity
AM	0.84	1.0	0.97	0.62
IM	1.0	0.98	0.97	0.79
HM	0.90	0.98	0.89	0.76

4.6 Conformance checking

In order to assess the performance of the best models in Pm4Py, conformance checking techniques such as token-based replay and alignment are employed. Token-based replay involves simulating the process execution by following the sequences of events as recorded in the event log. The goal is to see how well the process model represents the actual process execution. Alignment, on the other hand, involves comparing the event log with the process model and identifying any deviations or deviations from the model. These two techniques provide valuable insights into the efficiency and effectiveness of the model and allow for further improvement and optimization.

Inductive:

```
replaying log with TBR, completed variants :: 100%  [████████████████]  99/99 [00:00<00:00, 446.47it/s]
REPLAY
Number of traces 10500
declaration 86809, declaration 86720, declaration 86739, declaration 86750, declaration 86764, declaration 86840, d
Number of anomalous traces  839
Percentage of anomalous traces 7.9904761904761905 %
aligning log, completed variants :: 100%  [████████████]  99/99 [00:00<00:00, 100.39it/s]
ALIGNMENTS
Number of traces 10500
['declaration 86809', 'declaration 86720', 'declaration 86739', 'declaration 86750', 'declaration 86764', 'declarat
Number of anomalous traces  839
Percentage of anomalous traces 7.9904761904761905 %
```

Heuristic:

```
replaying log with TBR, completed variants :: 100% [███████████████████████] 99/99 [00:00<00:00, 557.53it/s]
REPLAY
Number of traces 10500
declaration 86791, declaration 86731, declaration 86735, declaration 86805, declaration 86809, declaration 868
Number of anomalous traces  7234
Percentage of anomalous traces 68.8952380952381 %
aligning log, completed variants :: 100% [███████████████████████] 99/99 [00:01<00:00, 75.02it/s]
ALIGNMENTS
Number of traces 10500
['declaration 86791', 'declaration 86731', 'declaration 86735', 'declaration 86805', 'declaration 86809', 'dec
Number of anomalous traces  7234
Percentage of anomalous traces 68.8952380952381 %
```

5 Project Results (International Declarations)

5.1 Extract data

Upon conducting a review of the international declaration dataset that we imported, we discovered that there was a total of 6,449 cases and 753 variants present. Additionally, we responded to previous inquiries that were related to domestic declarations by utilizing the international declaration dataset. After reviewing the variants, we found that there were 1755 instances where the events were rejected, while 5961 declarations received final approval from a supervisor. Furthermore, 6187 cases were successfully handled and there were 20 instances where the declaration was never approved. In terms of payments, there were no instances of double payments, however, there were 7,915 submissions that were made more than once. This information provides us with valuable insights into the status of the international declarations and the various outcomes associated with them.

5.2 Variant Analysis

In the following chart, the frequency of variants is depicted prior to the implementation of filtering:
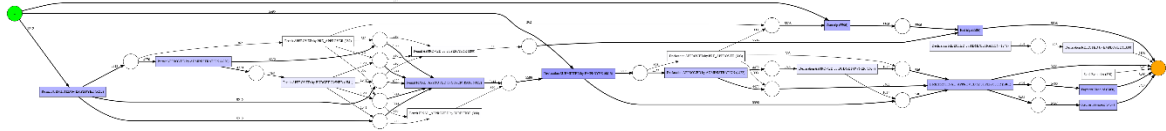
Variants frequency
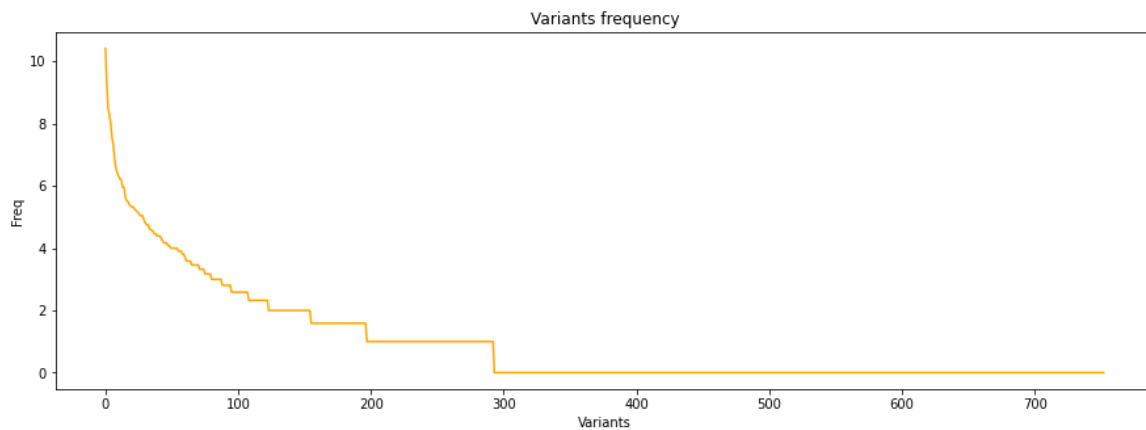
5.3 Filtering

After conducting an analysis of the duration data using the pandas dataframe, it was concluded that while the data was relatively clean from a duration perspective, there was an issue with the high number of activities that were repeated multiple times. This was causing significant bottlenecks and adding unnecessary complexity to the system. To resolve this issue, a number of specific activities were extracted from the event log. These activities included: 'Declaration APPROVED by SUPERVISOR', 'Declaration FINAL_APPROVED by DIRECTOR', 'Declaration REJECTED by BUDGET OWNER', 'Declaration REJECTED by DIRECTOR', 'Declaration REJECTED by MISSING', 'Declaration REJECTED by PRE_APPROVER', 'Declaration REJECTED by SUPERVISOR', 'Declaration SAVED by EMPLOYEE', 'Permit REJECTED by ADMINISTRATION', 'Permit REJECTED by BUDGET OWNER', 'Permit REJECTED by DIRECTOR', 'Permit REJECTED by EMPLOYEE', 'Permit REJECTED by MISSING', 'Permit REJECTED by PRE_APPROVER', and 'Permit REJECTED by SUPERVISOR'.

As a result of these changes, the number of cases dropped from 6449 to 5996, providing a more streamlined and manageable dataset.
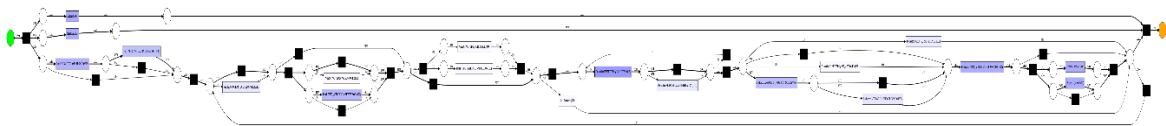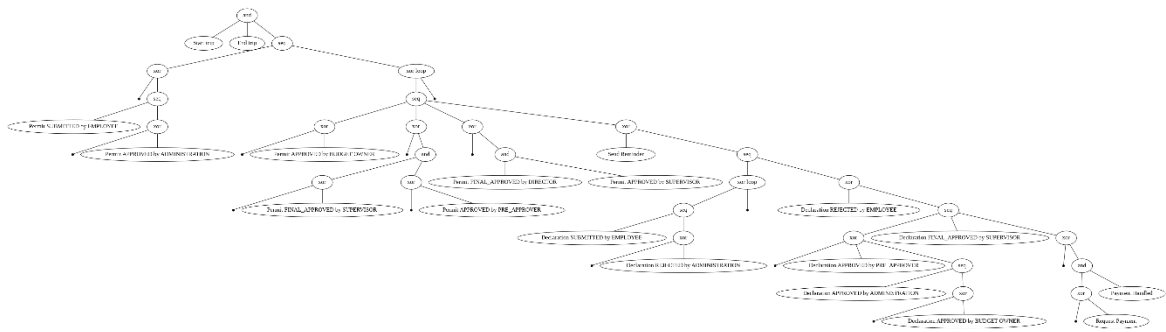
5.4 Process Discovery

Then on the log data the process discovery is applied, using the same algorithms for domestic declaration.
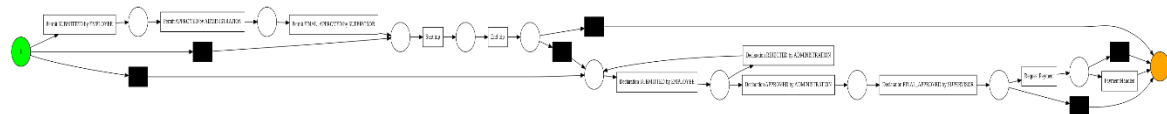
5.4.1 Alpha miner

5.4.2 Inductive miner

5.4.3 Heuristic miner

5.4.4 DFG

Permit SUBMITTED by EMPLOYEE (5212)

| 4262

Permit APPROVED by ADMINISTRATION (4270)

1458 | | 2398

Permit APPROVED by BUDGET OWNER (1541)

Permit FINAL_APPROVED by SUPERVISOR (4822)

| 3817

Start trip (5596)

| 5031

End trip (5596)

| 4040

Declaration SUBMITTED by EMPLOYEE (6815)

| 4399

Declaration APPROVED by ADMINISTRATION (4472)

1618 | | 2808

Declaration APPROVED by BUDGET OWNER (1637)

1626

Declaration FINAL_APPROVED by SUPERVISOR (5461)

| 5439

Request Payment (5457)

| 5436

Payment Handled (5460)

5.5 Quality Metrics

Also, the quality metrics (Precision, Fitness, Generalization, and Simplicity) are evaluated again in the following table.

Model	Fitness	Precision	Generalization	Simplicity
AM	0.69	0.40	0.97	0.45
IM	1.0	0.37	0.92	0.64
HM	0.93	0.97	0.85	0.74

5.6 Conformance Checking

Inductive:

```
replaying log with TBR, completed variants :: 100%  ████████████████  753/753 [00:02<00:00, 335.24it/s]
REPLAY
Number of traces 6449
declaration 74628, declaration 73012, declaration 73010, declaration 143612, declaration 73029, declaration 75907,
Number of anomalous traces  663
Percentage of anomalous traces 10.280663668785857 %
aligning log, completed variants :: 100%  ██████████████  753/753 [00:22<00:00, 41.82it/s]
ALIGNMENTS
Number of traces 6449
['declaration 72590', 'declaration 74628', 'declaration 143644', 'declaration 73012', 'declaration 73010', 'declar:
Number of anomalous traces  853
Percentage of anomalous traces 13.226856877035198 %
```

Heuristic:

```
replaying log with TBR, completed variants :: 100%  ██████████████  753/753 [00:01<00:00, 377.82it/s]
REPLAY
Number of traces 6449
declaration 76457, declaration 76667, declaration 73654, declaration 73596, declaration 73594, declaration
Number of anomalous traces  3467
Percentage of anomalous traces 53.76027291052876 %
aligning log, completed variants :: 100%  █████████████  753/753 [00:08<00:00, 61.95it/s]
ALIGNMENTS
Number of traces 6449
['declaration 76457', 'declaration 76667', 'declaration 73654', 'declaration 73596', 'declaration 73594',
Number of anomalous traces  5073
Percentage of anomalous traces 78.66335866025742 %
```

6 Conclusion

After applying process mining techniques to two datasets, it was discovered that there was no evidence of double payments in either dataset. In terms of handling domestic declarations, 10044 cases were processed out of a total of 10,500. Meanwhile, in international declarations, 6187 out of 6449 were handled. A challenge faced in handling domestic declarations was that a significant number of cases took more than 100 days to resolve. Additionally, in international declarations, there were multiple instances of repeated activities that resulted in rework, constituting a bottleneck in the process. To address these challenges, variant analysis and filtering techniques were utilized to answer questions and improve the process by removing unnecessary data and reducing noise. The filtered process was then discovered by testing various process discovery algorithms and selecting the best one for each dataset based on quality metrics.

Finally, a conformance check was performed by comparing the real log with the discovered model to ensure compliance with the process.

In conclusion, the application of process mining techniques has allowed for the identification and addressing of bottlenecks in the process and ensured compliance, leading to improved efficiency and effectiveness in handling declarations.