BPM Challenge 2020

Vadim Sokolov

Business Information Systems

Prof. Paolo Ceravolo

a.a. 2021 - 2022

Case study description

At Eindhoven University of Technology staff members travel to conferences or to other universities for project meetings and to meet up with colleagues in the field. The university pays for travel expenses.

There are two types of trips, domestic and international. For domestic travel no prior permission is required. An employee can undertake these trips and ask for reimbursement of the costs afterwards.

For international trips, permission is needed from the supervisor. This permission is obtained by filing a travel-permit and this travel permit should be approved before making any arrangements.

To get the costs for a trip reimbursed, a claim is filed. This can be done as soon as costs are actually paid, or within two months after the trip.

The data presented in this challenge

The data is collected from the reimbursement process at TU/e. The files contain data from 2017 (only two departments) and 2018 the full TU/e.

The data is split in datasets presented in the following table:

Table 1 - Data representation

| Number | Name | Cases | Events | Size, Mb | Description |
| --- | --- | --- | --- | --- | --- |
| 1 | Domestic declarations | 10500 | 56437 | 20.5 |  |
| 2 | International declarations | 6449 | 72151 | 29.2 |  |
| 3 | Prepaid travel costs | 2099 | 18246 | 7.8 |  |
| 4 | Requests for payment | 6886 | 36796 | 15.2 |  |
| 5 | Travel permits | 7065 | 86581 | 33.2 |  |

Goals

The goal of the research is to analyse the data provided in order to extract useful information. The organisation can use the results of the analysis to reach its goals, improve performance of the processes, reduce bottlenecks and take into account the information about data anomalies.

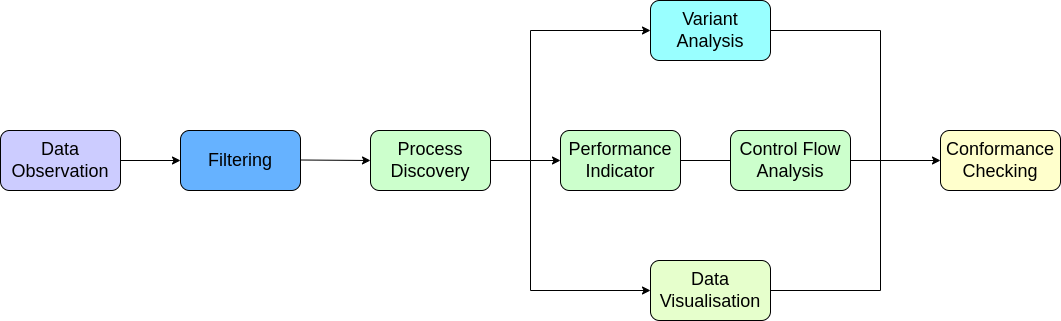


Figure 1 - Goal Diagram

Questions of the challenge

* What is the throughput of a travel declaration from submission (or closing) to paying?
* Is there a difference in throughput between national and international trips?
* Are there differences between clusters of declarations, for example between cost centres/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?

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 traveller 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?

Knowledge uplift trail

The first step of the analysis is filtering. It allows us to get rid of noisy data. One of the most important parts of analysis is the search for bottlenecks.

Specification:

Employees, departments represented by administration, budget owners and supervisors are actors in this process. The values in the model are a travel expenses cost and processing time of the declaration.

Project results

1. Filtering

First, Domestic Declarations were taken into account for analysis. The first step before extracting knowledge is preprocessing that can be made by filtering. Filtering process was implemented using Pm4Py filtering methods. As we can see from table 1, there are 10500 cases in total for domestic declarations. Filtering by start activity allows us to extract events that start with 'Declaration SUBMITTED by EMPLOYEE' event, which is a necessary requirement for starting the process of declaration submission. It corresponds to reducing the number of cases to 10365. The next step is to filter the data on the number of activities. We assume that declarations can’t consist of less than 2 events, that is, “Declaration submitted by Employee”, “Declaration approved / rejected” (by any administrative instance). Also we set the minimum requirement for a total time for each activity, namely, 1 day. It brings us to a reduction in the number of cases to 10349 cases. Assuming that we are interested in only successful declarations, which end with “Payment Handled” event, filtering over end activities was performed, reducing total number of activities to 10042. Successful declarations give us the most useful information about the system. It is noteworthy that all the filtered activities last at least 25 hours. As one can see, filtering reduced the total number of cases to be analysed by 4.3%.

2. Variant analysis

Number of variants for domestic declarations is equal to 63. Basic results of variant analysis are presented in table 2. As one can see, some variants represent cases which were never approved. Overall success rate is high, the quality of the dataset is relatively good.

Table 2 - Variant analysis Domestic declarations

| Property | Value | Description |
| --- | --- | --- |
| Cases number | 10500 | Total number of cases |
| Handled cases | 10042 | Handled case as a final event |
| Never approved | 230 | Cases that were not approved |
| Handled ratio | 0.968 | Ratio of successful cases |
| Rejected | 1301 | Number of cases rejected at least once |
| Resubmitted | 1166 | Number of cases resubmitted at least once |
| Average duration | 11 days 14 hours | Average duration of the case from submission to payment |

To make performance measurements, it is necessary to define a performance indicator. First, a key strategy for an employee in a domestic declaration is to submit it, meeting all the requirements of administrative institutions in order to get the approval as fast as possible. Therefore, it is necessary to avoid resubmissions of declarations. Successful case is a case in which a declaration is submitted by an employee, finally approved by the supervisor and payment was handled.



Figure 2 - Variant distribution

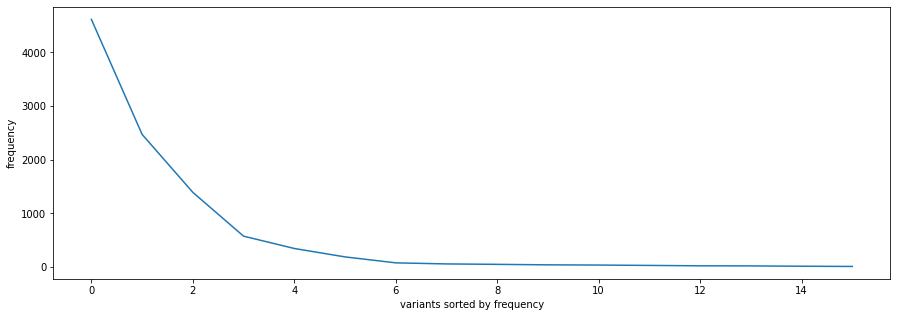


Figure 3 - Variant distribution, most frequent

More than half of the variants have frequency rate less than 10, these cases can be rarely seen. As one can see, this is an example of Pareto distribution.

Table 3 - Variant Distribution

| Parameter | Value |
| --- | --- |
| Mean | 621 |
| Median | 53 |
| Mode | 20 |

3. Process discovery

In order to analyse processes in the system and create the most representative data model, it is necessary to apply different process discovery techniques.

1. Alpha Miner

Alpha miner algorithm is applied taking into consideration the frequency of events.

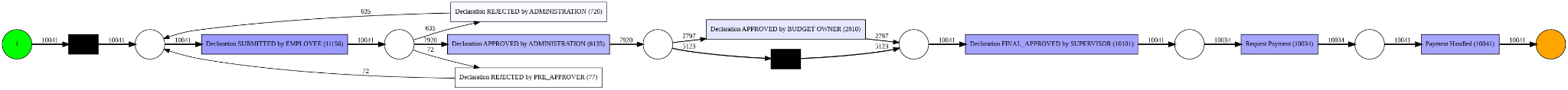


Figure 4 - Petri net for Alpha Miner

2. Inductive Miner

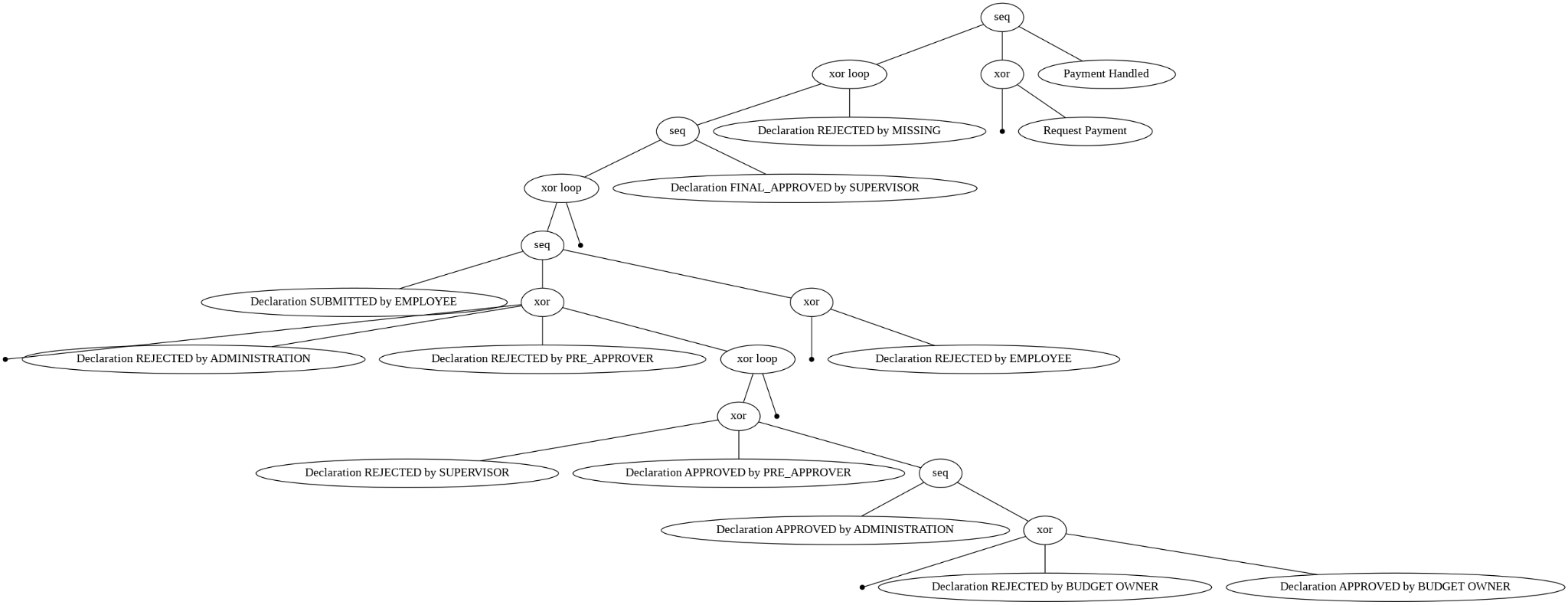


Figure 5 - Process tree Inductive Miner

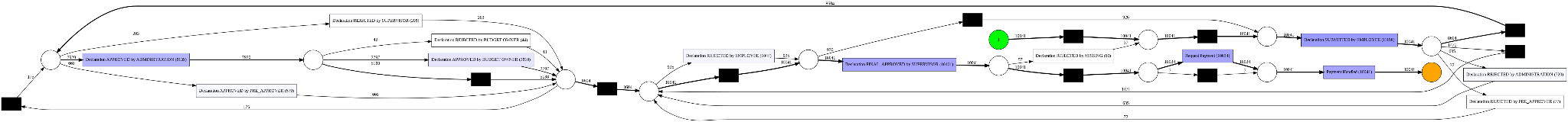


Figure 6 - Petri net Inductive Miner

3. Heuristic Miner

In order to find the most successful case, different values of dependency threshold were observed.

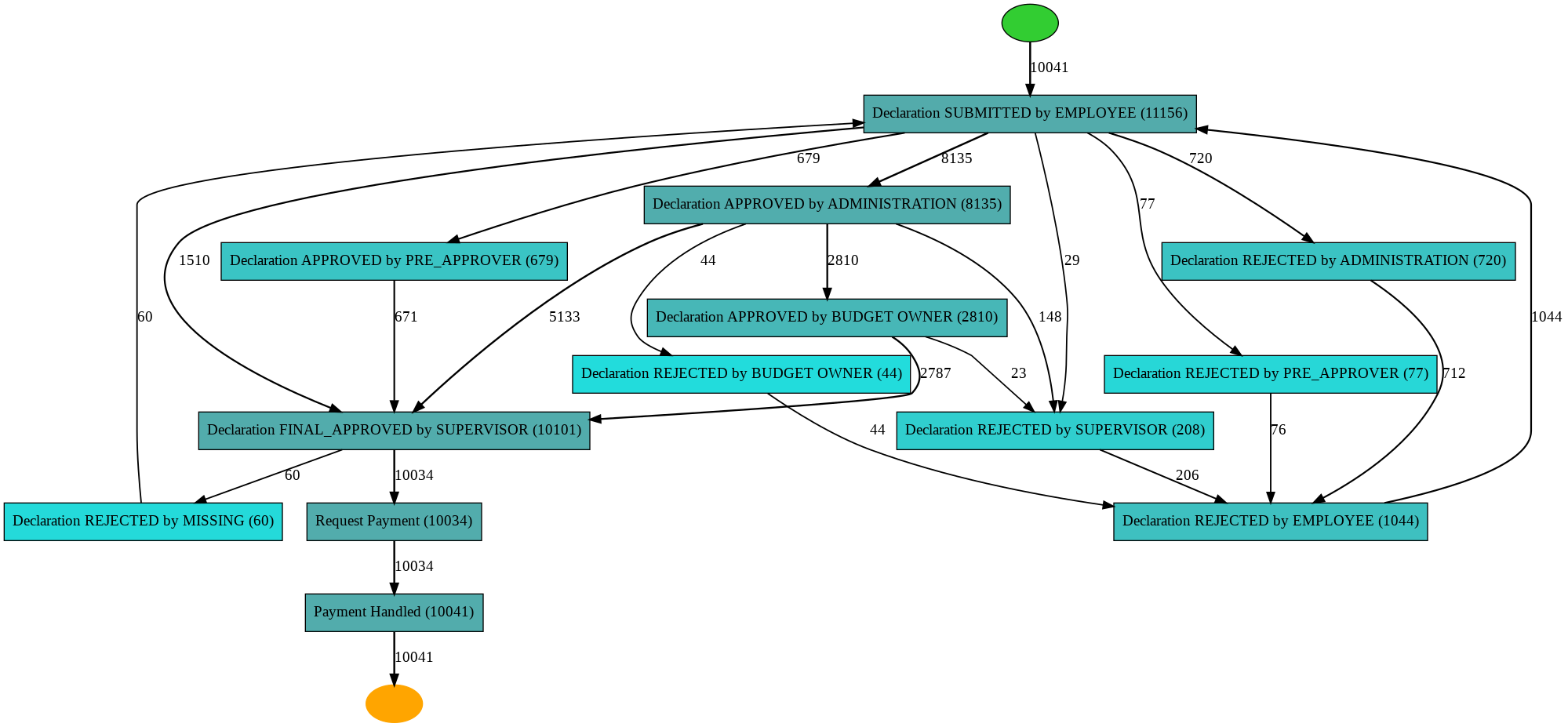


Figure 7 - Process tree Heuristic Miner, dependency threshold: 0.9

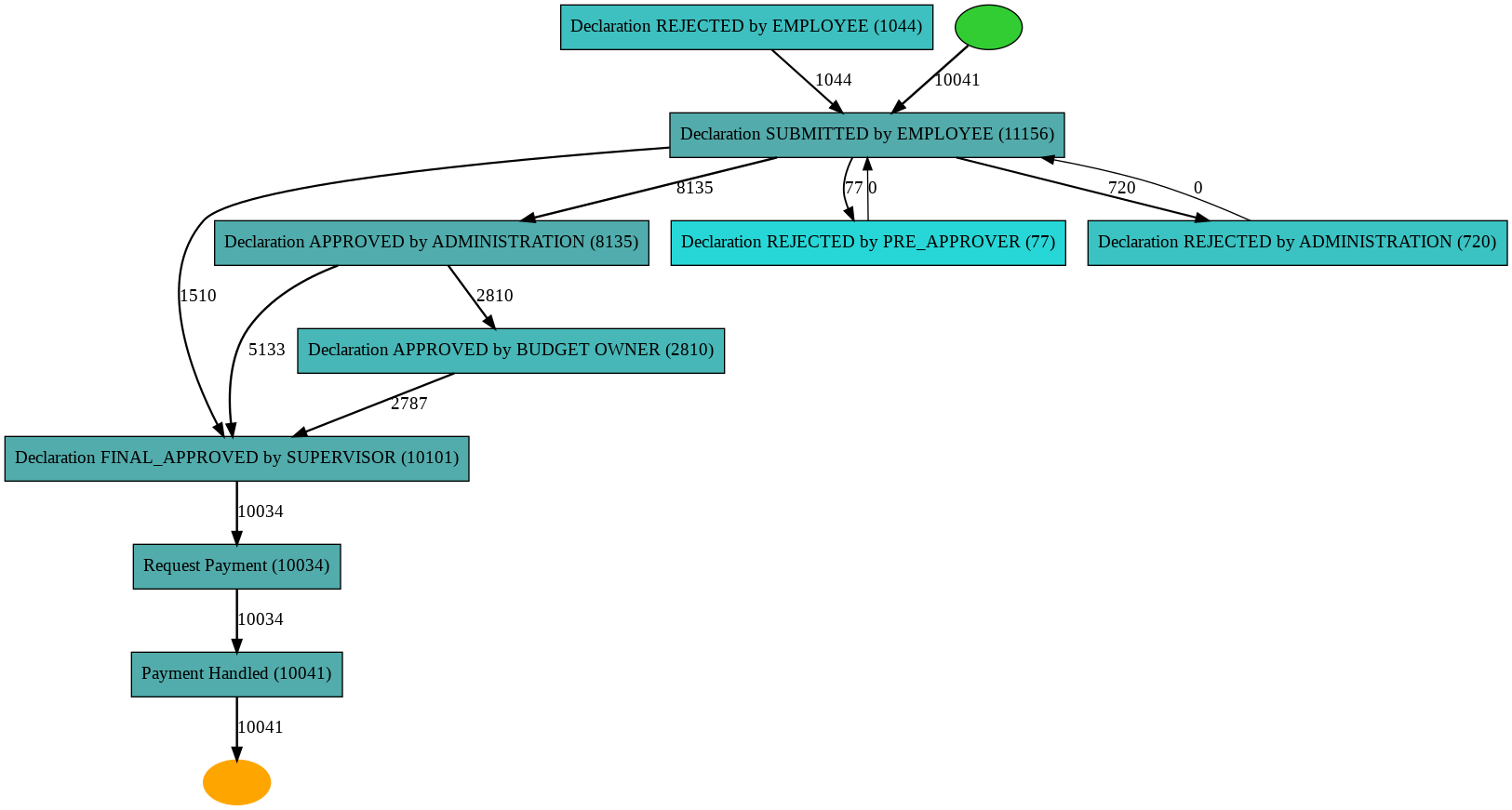


Figure 8 - Process tree Heuristic Miner, dependency threshold: 0.999

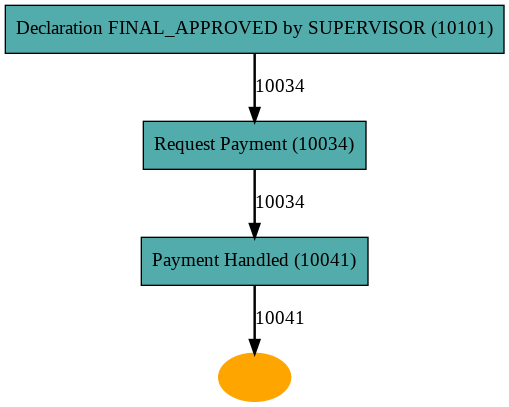


Figure 9 - Process tree Heuristic Miner, dependency threshold: 0.9999

As one can see, the majority of cases fit in the 90% threshold, which means that there are only slight deviations from the successful case.

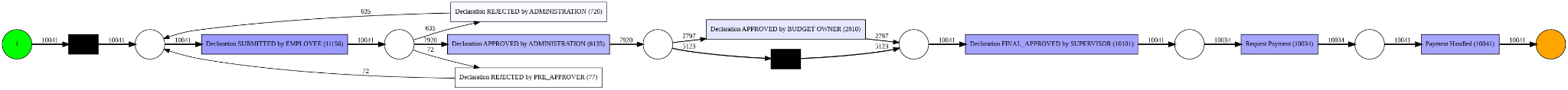


Figure 10 - Heuristic Miner, Petri Net

4. Directly-followed graph that shows good successful case representation contains 90% of all the cases.



Figure 11 - Directly-followed graph

4. Conformance Analysis

In order to understand the accuracy of developed models, it is necessary to provide a conformance analysis - validation of predictive models.

Table 4 -- Conformance Analysis

| Algorithm | Fitness | Precision | Generalisation | Simplicity |
| --- | --- | --- | --- | --- |
| Alpha miner | 0.96 | 0.95 | 0.97 | 0.81 |
| Inductive miner | 1.0 | 0.62 | 0.94 | 0.62 |
| Heuristic miner | 0.96 | 0.95 | 0.97 | 0.82 |
|  |  |  |  |  |

Analysis of International declarations

International declarations have different structure in comparison to domestic ones. In international declarations it is possible to specify start and end dates of trip and deal with all the payments later on. It brings an additional number of variants, additional parallelisation and loops.

1. Filtering

Since it is not necessary to start the activity by submitting the declaration or permit for a trip and huge variability in multiple cases, it would be incorrect to filter all the activities by starting events. The only property of a successful case is supposed to be a handled payment, assuming that all the trips can not be for free. Total number of activities was reduced from 6449 to 6187.

2. Variant Analysis

Table 5 - Variant analysis International declarations

| Property | Value | Description |
| --- | --- | --- |
| Cases number | 6449 | Total number of cases |
| Handled cases | 6187 | Handled case as a final event |
| Never approved | 20 | Cases that were not approved |
| Handled ratio | 0.959 | Ratio of successful cases |
| Rejected | 1755 | Number of cases rejected at least once |
| Resubmitted | 7915 | Number of cases resubmitted at least once |
| Final approval by supervisor | 5961 | Cases in which employee had a supervisor’s approve |
| Permit submitted by Employee | 5761 | Permit submitted by Employee at least once |
| Average duration | 86 days 15 hours | Average duration of the case from submission to payment |

As far as we can see from basic variant analysis, it is more likely that an employee resubmits the declaration. To put it simply, it is more difficult to prepare the declaration that will satisfy the requirements for an international trip in comparison to a domestic one. Total number of variants is 645. Number of variants with frequency more than 10 is 71, dropping 9 times.

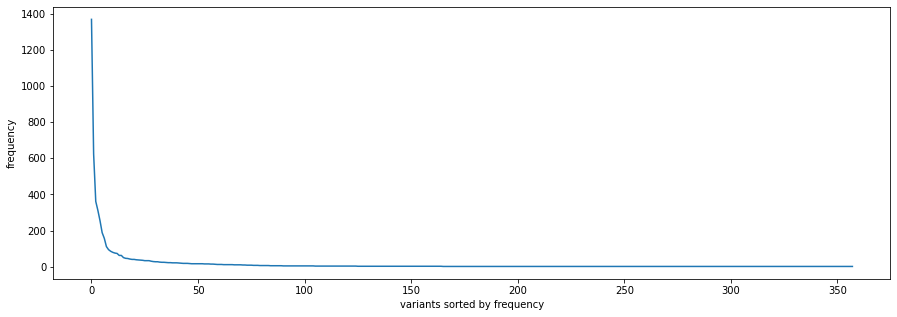


Figure 12 - Distribution of variants

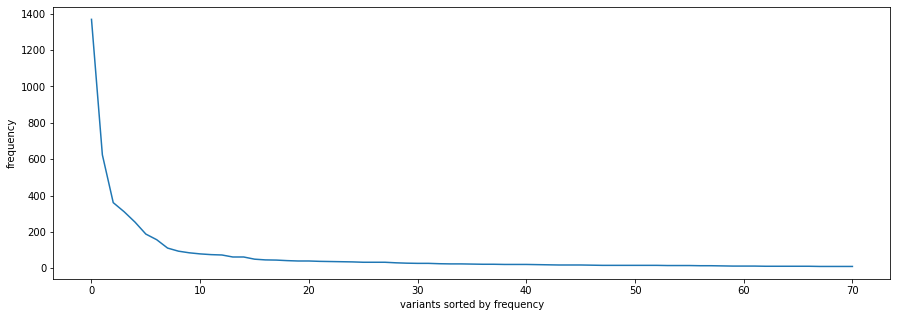


Figure 13 - Distribution of variants, most frequent cases

Table 6 - Variant distribution

| Parameter | Value |
| --- | --- |
| Mean | 72.54 |
| Median | 23 |
| Mode | 16 |

3. Process Discovery

1. Alpha Miner

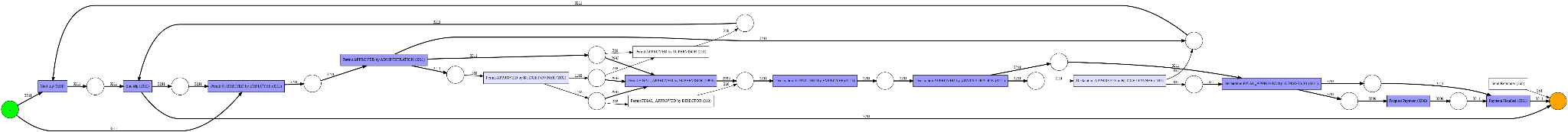


Figure 14 - Alpha Miner Petri net

Even though the number of variants was reduced up to 117 and all the rejections in variants were eliminated, the graphical representation of the Petri net for Alpha miner algorithm is quite complicated. The reason is that there are still many variants because of the presence of shuffles between different events that can be assumed as successful.

2. Inductive Miner

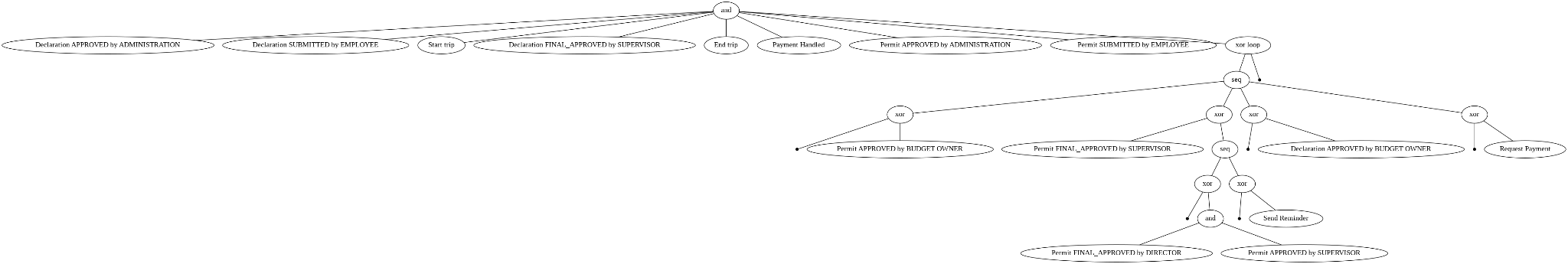


Figure 15 - Inductive Miner Tree

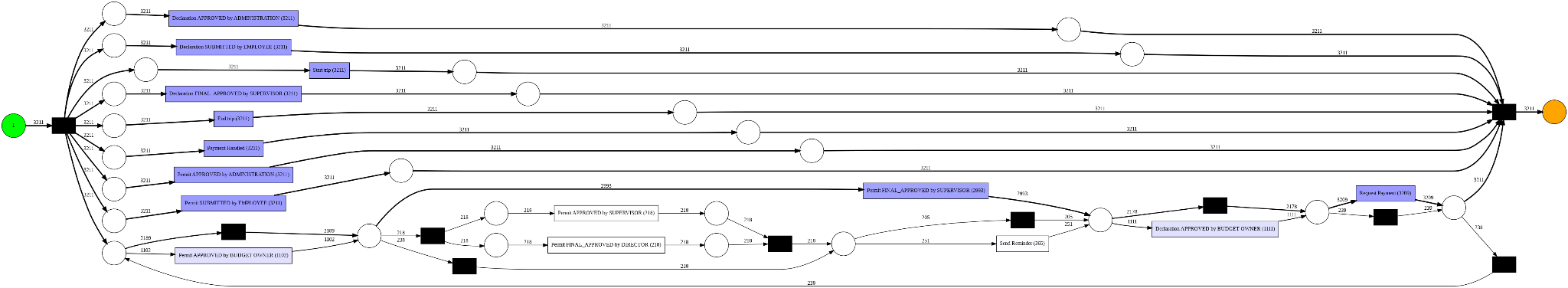


Figure 16 - Inductive Miner Petri net

3. Heuristic Miner



Figure 17 - Heuristic miner Graph



Figure 18 - Heuristic miner Petri net

4. Directly-followed graph

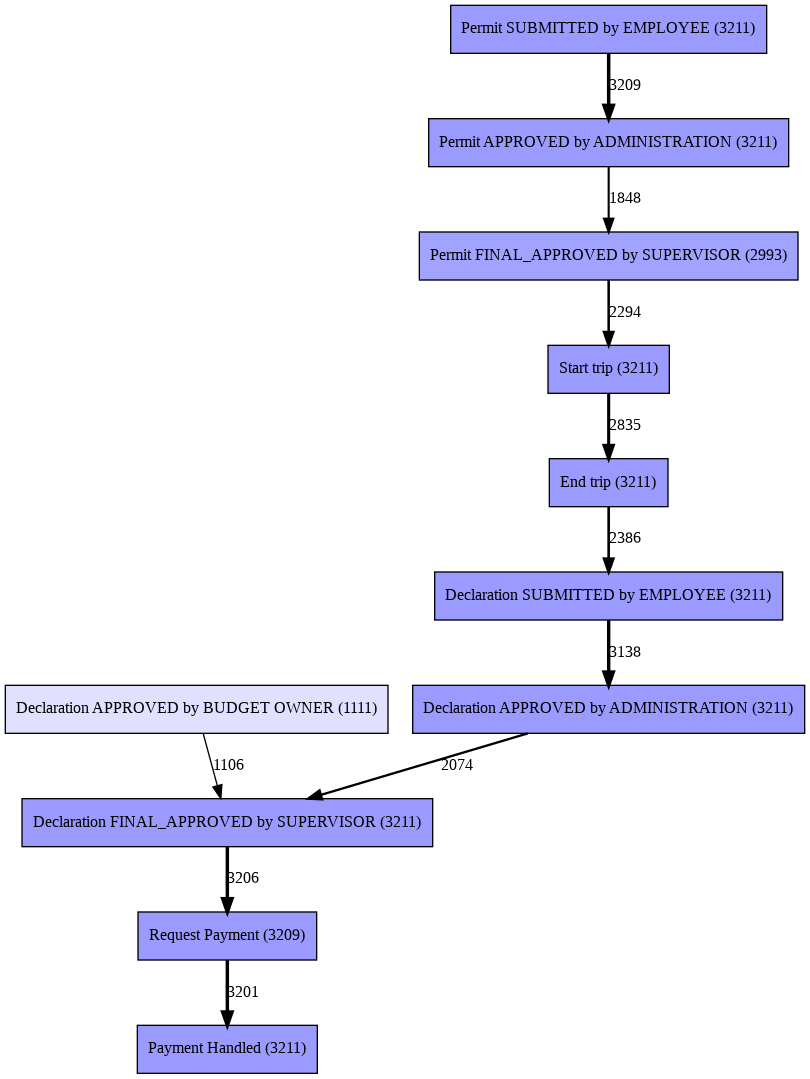


Figure 19 - Directly-followed graph

4. Conformance Analysis

Table 7 -- Conformance Analysis

| Algorithm | Fitness | Precision | Generalisation | Simplicity |
| --- | --- | --- | --- | --- |
| Alpha miner | 0.75 | 0.78 | 0.97 | 0.63 |
| Inductive miner | 1.0 | 0.73 | 0.96 | 0.64 |
| Heuristic miner | 0.94 | 0.97 | 0.88 | 1.0 |
|  |  |  |  |  |

5. Conclusion

Bottlenecks = last approve and payment handled

Trip duration

Trip is a bottleneck?

Performance indicator:

1 Determine the Key strategic objectives

2 Define Success

3 Decide on measurements: msr unit, temporal frame, data source

Time, cost, quality

Processing time, waiting time etc -> calculate it

Cost: material + resource

Quality: process, product, customer satisfaction

Case arrival rate

To measure performance we include all dimensions -> create a variant

Performance measurements -- variant analysis.

Mean median mode

Distribution of variants

Control flow analysis -- identify patterns

Sequence

Synchronization

Parallelisation

Iteration

Combination

Rework, bottleneck, Cancellation, deviant flow

Process Discovery -- construct a representation of a process

DFG, Petri Net, Process tree

UML, BPMN

DIfficult to match Process discovery algo with process model

Alpha miner

Inductive miner

Genetic Miner

Validation of predictive models:

Precision, recall (fitness), Generalization, simplicity

Model quality can be improved by reducing variability of event log: filtering, variant analysis, clustering etc. Probably division of logs in 5 parts is already an example of clustering. Also division of event log by department and then creating a model.

Conformance Checking …

Observed Process VS expected process

Rule based CC: rules:

control flow

time

Resource

Imperative Process Model --

replay techniques

Trace alignment -> alignment matrix

Behavioural alignment

Iterations and parallel executions lead to infinite behaviour

Conformance checking helps to identify deviating behaviour of Event Log or additional behaviour to update model

Classifying traces:

Early - activities executed in the trace before than specified in the model

Late - activities executed in the trace later than specified in the model

Insert - activities executed in the log but not specified in the model

Skip - activities specified in the model but not executed in the trace

Process Mining Epistemic Dilemma

Comparative PM

Compare variants, processes, segments of Event Log.

Significance of difference

Chi-square test

Central tendency

Wasserstein’s distance

Time Series analysis

Time series: base level, trend, seasonality, error

Entropy of Time Series

Encoding

Autocorrelation -- predict the next values based on previous

Moving average

ARIMA and SARIMA models for moving average

ML: knn and random forest