BPI 2020 Challenge

Alireza Gholamrezazadeh

Business Information Systems

Prof. Paolo Ceravolo

a.a. 2021 - 2022

Table of contents

Description of the case study

This case study is about staff members traveling for different purposes as it's common in many organizations, staff members travel for work. They travel to customers, conferences, or project meetings, which are sometimes expensive. As an employee of an organization, you do not have to pay for your own travel expenses, but the company takes care of them.
For domestic trips, no prior permission is needed, i.e. an employee can undertake these trips and ask for reimbursement of the costs afterward.
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 travel reimbursed, a claim is filed. This can be done as soon as costs are actually paid (for example for flights or conference registration fees) or within two months after the trip (for example hotel and food costs which are usually paid on the spot).

Challenge provided datasets description

The study dataset contains 5 files that each of which takes part in the case study as below:

- **Requests for Payment**(6,886 cases, 36,796 events): This dataset contains all requests for getting payment of the costs of the travel.
- **Domestic Declarations**(10,500 cases, 56,437 events): This dataset contains all domestic travels, their details, and procedures.
- Prepaid Travel Cost(2,099 cases, 18,246 events):
- **International Declarations**(6,449 cases, 72151 events): This dataset contains all international travels, their details, and procedures.
- **Travel Permits**(7,065 cases, 86,581 events): This dataset contains all information related to the permits according to a travel(domestic or international).

Case study goals

The goal of the case study is completely similar to any other process mining activity. In the report, we will cover the most important 3 steps to recognize the processes and bottlenecks inside our case and try to improve them according to

the algorithms and methods discussed in during the program of **BIS** course. Thus, we will start with process discovery and will check it by conformance checking techniques and finally, try to enhance the performance of existing processes). Moreover, during every step, the next section questions will be answered too.

Challenge asked questions

The following questions are asked by the challenge and during the next steps we are going to answer them as much as possible(because some of them might not be related to our study so we will skip them):

1. ***What is the throughput of a travel declaration from submission (or closing) to paying?***
2. Is there are difference in throughput between national and international trips?
3. Are there differences between clusters of declarations, for example between cost centers/departments/projects, etc.?
4. What is the throughput in each of the process steps, i.e. the submission, judgment by various responsible roles, and payment?
5. ***Where are the bottlenecks in the process of a travel declaration?***
6. Where are the bottlenecks in the process of a travel permit (note that there can be multiple requests for payment and declarations per permit)?
7. ***How many travel declarations get rejected in the various processing steps and how many are never approved?***

Then there are more detailed questions

1. How many travel declarations are booked on projects?
2. How many corrections have been made for declarations?
3. Are there any double payments?
4. Are there declarations that were not preceded properly by an approved travel permit? Or are there even declarations for which no permit exists?
5. How many travel declarations are submitted by the traveler and how many by a mandated person?
6. 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?
7. Is this different between departments?

8. How many travel declarations are not approved by budget holders in time (7 days) and are then automatically rerouted to supervisors?
9. 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

In this project the analytical steps are

1. **Log extraction:** At this step, the raw input file logs will be converted into the event log by importing it into the pm4py.
2. **Variant analysis**: In this step, any different variants of the case study will be recognized and discussed. Thus, the event log at this point will be converted into variants and their distributions.
3. **Filtering**: At this step, the noisy data and unused logs will be removed from the dataset to make our analysis more precise. So, the input is the event log and the output is the filtered event log.
4. **Process discovery**: In this step by using different algorithms and methods, the useful diagrams(graphs) will be demonstrated. The input at this step is a filtered event log and the output is the extracted model.
5. **Conformance checking**: At this step, the performance of the created models will be assessed, and compared. The input at this step is the model created in the previous step and the outputs are the important factors of the assessment.
6. **Process enhancement**: In this final step, the possible improvements at different parts will be discussed. So, the input here is the old model and the output is the new enhanced model.

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

Table 1

Project Results

In the following sections, datasets will be analyzed from different aspects and the result will be discussed. The related code also can be found in this GitHub repository.

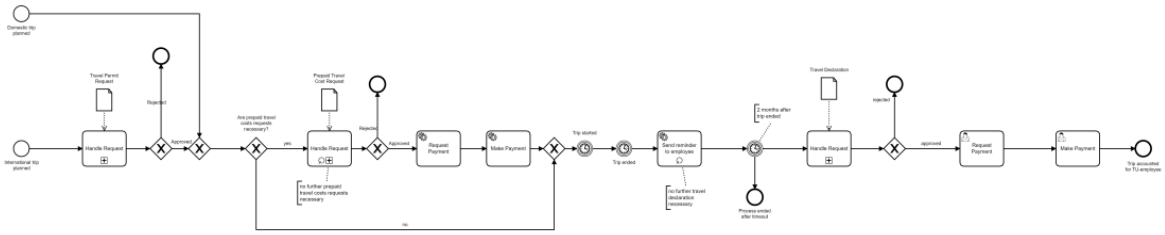

Diagram 1 - BPMN of international declarations

International declaration dataset

The first dataset that we are going to analyze is the international declaration(**idl** variable name in the code), this dataset contains all details about the international travels of the staff members of TU/e university.
This log has more case attributes than its domestic counterpart but is similar to the domestic declarations log, the id, and the **concept:name** is identical. Also, the DeclarationNumber is the id incremented by one in 6,189 cases whereas in 260 cases, it is UNKNOWN. Next to these, there exist four columns with IDs

referring to the related permit, namely: Permit travel permit number, travel permit number, Permit ID & Permit id. The travel permit number and Permit travel permit number are identical in 5,970 cases and different in 479. In the differing cases, the Permit travel permit number is 23 times UNKNOWN. Permit ID and permit id are identical in 6,001 cases and never UNKNOWN. When the two IDs are different, the Permit id is always travel permit 423. Furthermore, in 5,970 cases the travel permit number is the Permit ID incremented by one while in 448 cases, it is decreased by one, leaving 31 cases, where both numbers are completely different. In addition, there are five numerical case attributes Amount, RequestedAmount, OriginalAmount, AdjustedAmount & Permit RequestedBudget. The first three are always identical, whereas the AdjustedAmount is different, only in one single case, where the Amount of 0 is adjusted to 100.49. The requested budget is usually higher than the amount. Next to these, the process has 6 categorical attributes attached to it, which are encoded as IDs. The attributes include 6 different tasks, 207 budgets, 719 permit budgets, 825 projects, 27 organizational entities, and 34 activities (Further details in appendix C).

Start activities

All the events inside the event log started with the following activities, in the table below you can also find their respective distribution:

Declaration SAVED by EMPLOYEE	8
Declaration SUBMITTED by EMPLOYEE	407
Permit SUBMITTED by EMPLOYEE	5294
Start trip	740

Table 1

As the table 1 demonstrate, you can see most of the activities started by Permit SUBMITTED by EMPLOYEE so we can consider this activity as a good option for starting the stream of activities. However, we can see a very low amount of activities started by Declaration SAVED by EMPLOYEE also we can consider this starting point as noisy data and filter it in the next section because it doesn't make any sense in the flow.

Middle Activities

There are 34 activities between the start and end point and among them, 15 activities(as shown in table 2) have very low frequencies which just pollutes the dataset so in the next steps(filter) they will be filtered to make the models clear.

Declaration REJECTED by PRE_APPROVER	84
Declaration SAVED by EMPLOYEE	75
Declaration REJECTED by MISSING	103
Permit REJECTED by MISSING	43
Declaration REJECTED by SUPERVISOR	126
Declaration APPROVED by SUPERVISOR	256
Declaration FINAL_APPROVED by DIRECTOR	252
Permit REJECTED by PRE_APPROVER	25
Permit REJECTED by EMPLOYEE	231
Declaration REJECTED by DIRECTOR	4
Permit REJECTED by SUPERVISOR	92
Permit REJECTED by ADMINISTRATION	83
Declaration REJECTED by BUDGET OWNER	40
Permit REJECTED by BUDGET OWNER	31
Permit REJECTED by DIRECTOR	1

Table 2

End activities

All the events inside the event log also ended with the following activities, in the table below you can also find their respective distribution:

Declaration SAVED by EMPLOYEE	54
Declaration REJECTED by EMPLOYEE	130
Declaration REJECTED by MISSING	11
End trip	593
Payment Handled	5646
Permit REJECTED by MISSING	8

Table 2

Some of the activities with a very low number of frequencies(less than 10) removed to make table cleaner. As you can see most of the activities ended by Payment Handled so we can consider this activity as a good option as an ending point. However, we can see a very low amount of activities ended by Permit REJECTED by MISSING also we can consider this ending point as noisy data and filter it in the next section because it doesn't make any sense in the flow.

Variant analyses

The variant analyses of the dataset show that there are 753 different variants from the start to end activities but you can find their distribution in the following table(to inspect more details in about the variants you can find a detailed table version in the collab code file).

	variant	count
0	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	1369
1	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	624
2	Permit SUBMITTED by EMPLOYEE,Permit FINAL_APPR...	361
3	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	311
4	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	254

Table 3

After plotting the distribution of table 3 into the diagram by using the logarithm function on the frequencies, we can find the below diagram:

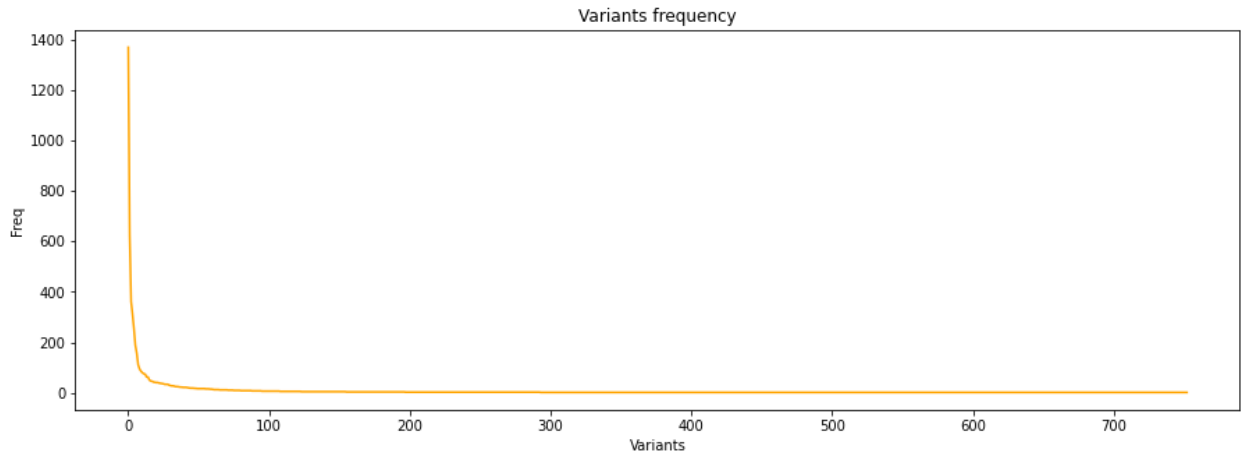

Fig 1 - Variant frequencies

As demonstrated in figure 1 most of the cases are handled by the top 5 variants. So, the top 5 can be very useful in our research and we are going to use them mostly in the next steps. However, before going to the next step it can be useful to find out the duration of each handled case.

case:concept:name	Activity	Activity_list	Resource	Duration
declaration 61184	16	Start trip,Permit SUBMITTED by EMPLOYEE,Permit...	2	742 days 00:00:00
declaration 7483	10	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	2	463 days 01:15:43
declaration 143637	6	Declaration SUBMITTED by EMPLOYEE,Declaration ...	2	458 days 14:12:24
declaration 143644	12	Declaration SUBMITTED by EMPLOYEE,Declaration ...	2	458 days 11:42:47
declaration 143578	6	Declaration SUBMITTED by EMPLOYEE,Declaration ...	2	458 days 08:39:42
...	...	...	...	...
declaration 30793	10	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	2	7 days 07:41:11
declaration 146876	7	Declaration SUBMITTED by EMPLOYEE,Declaration ...	2	7 days 05:45:56
declaration 34709	10	Permit SUBMITTED by EMPLOYEE,Permit APPROVED b...	2	7 days 01:21:22
declaration 81906	8	Permit SUBMITTED by EMPLOYEE,Permit FINAL_APPR...	2	6 days 20:52:15
declaration 28965	10	Start trip,Permit SUBMITTED by EMPLOYEE,Permit...	2	6 days 17:31:11

Fig 2 - Handled case durations

In Figure 2 you can find out the maximum duration for handling a case is **742 days**(it's only 1 and is much higher than the second so it's noise and can be filtered) and the minimum is **6 days** also the average duration of a handled case is **86 days 8 hours**.

Filtering

This part is the most important section in our project which is responsible for making sure the dataset doesn't have huge noises and it's suitable for going on with the next sections and discovering the processes. At this step, the start and end activities are limited to the most important ones, and the very low frequencies activities that have been recognized in the previous sections are removed from the cases to make it as much and possible clean. In the below list you can see the filtered activities for each section:

- Start Activities: `'Declaration SUBMITTED by EMPLOYEE'`, `'Permit SUBMITTED by EMPLOYEE'`, `'Start trip'`

- Activities(skipped ones): `'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', 'Permit REJECTED by SUPERVISOR'`
- End Activities: `'Payment Handled', 'End trip', 'Declaration REJECTED by EMPLOYEE'`

Process Discovery

The variant analyses of the dataset show that there are 753 different variants from the start to end activities but you can find their distribution in the following table(to inspect more details in about the variants you can find a detailed table version in the collab code file).

Alpha miner

The following graph shows the Petri net using the alpha miner algorithm.

Fig 3 - Petri net using the alpha miner algorithm

Inductive miner

The following graph shows the tree using the inductive miner algorithm.

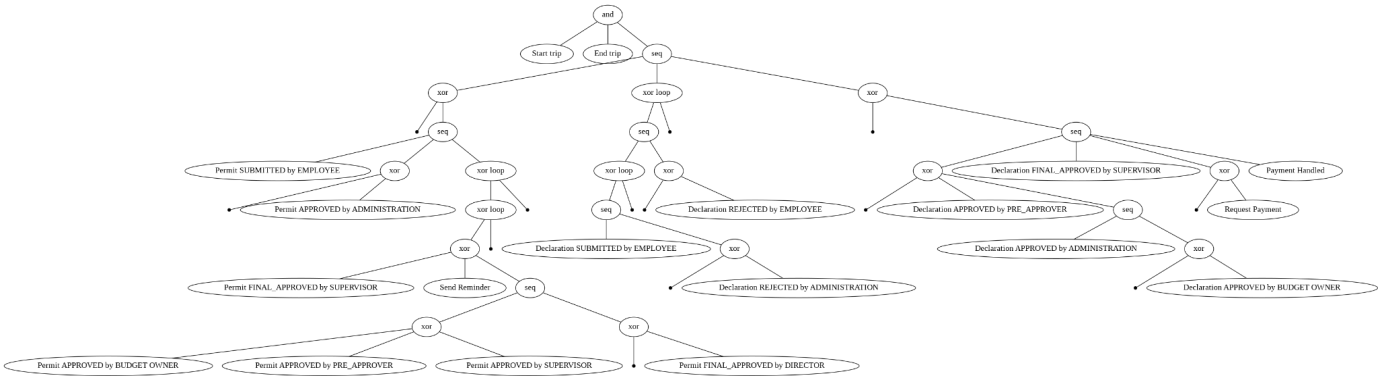

Fig 4 - tree using the inductive miner algorithm

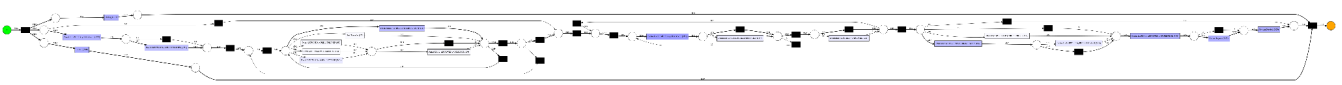

Fig 5 - Petri net using the inductive algorithm.

Heuristics miner

The following graph shows the graph using the heuristics miner algorithm.

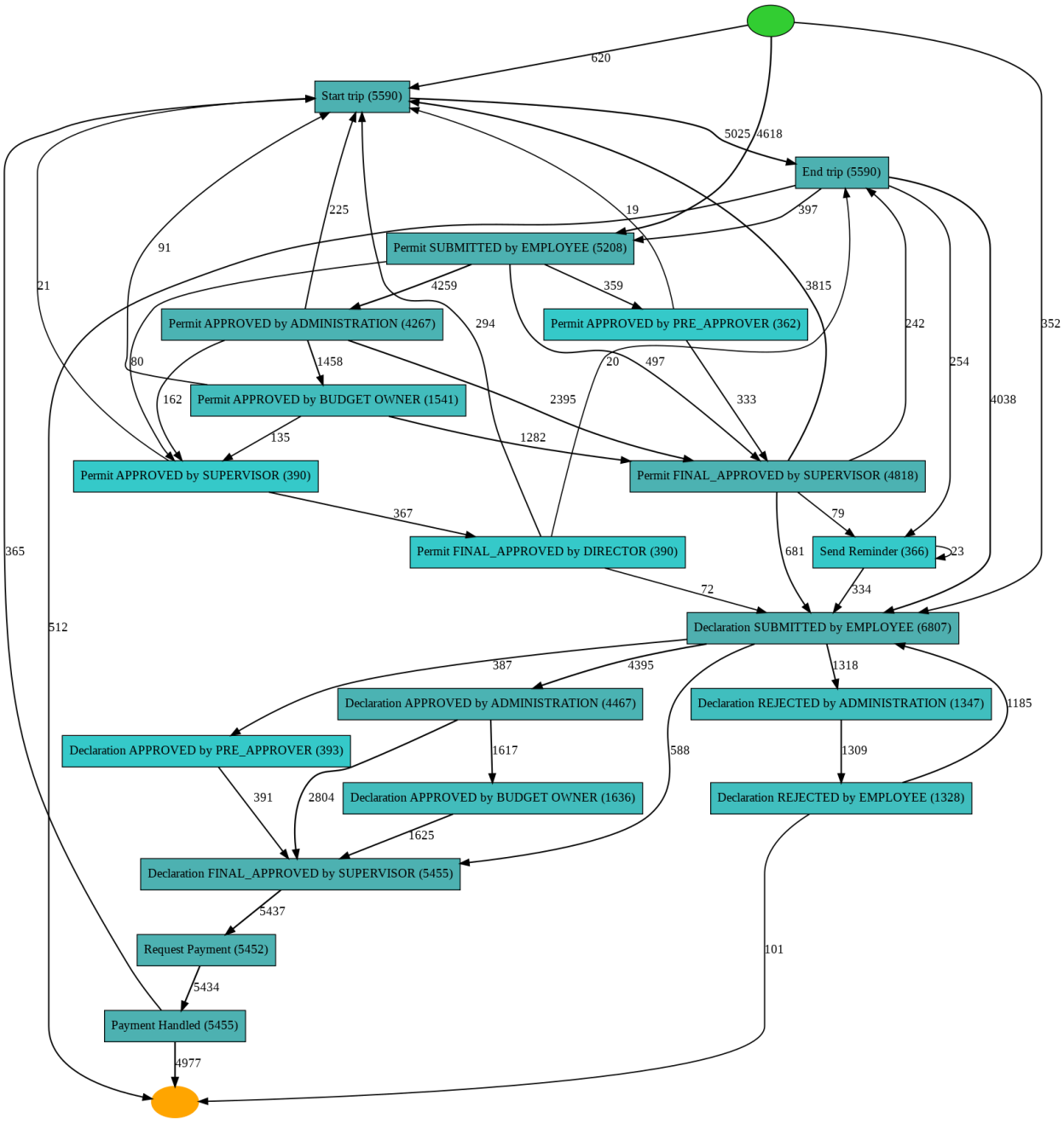

Fig 6 - Graph using the heuristics algorithm.

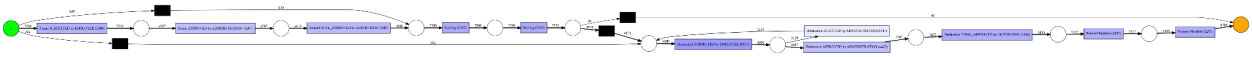

Fig 7 - Petri net using the heuristics algorithm.

Directly-followed graph

The following graph shows the DFG of the filtered event log.

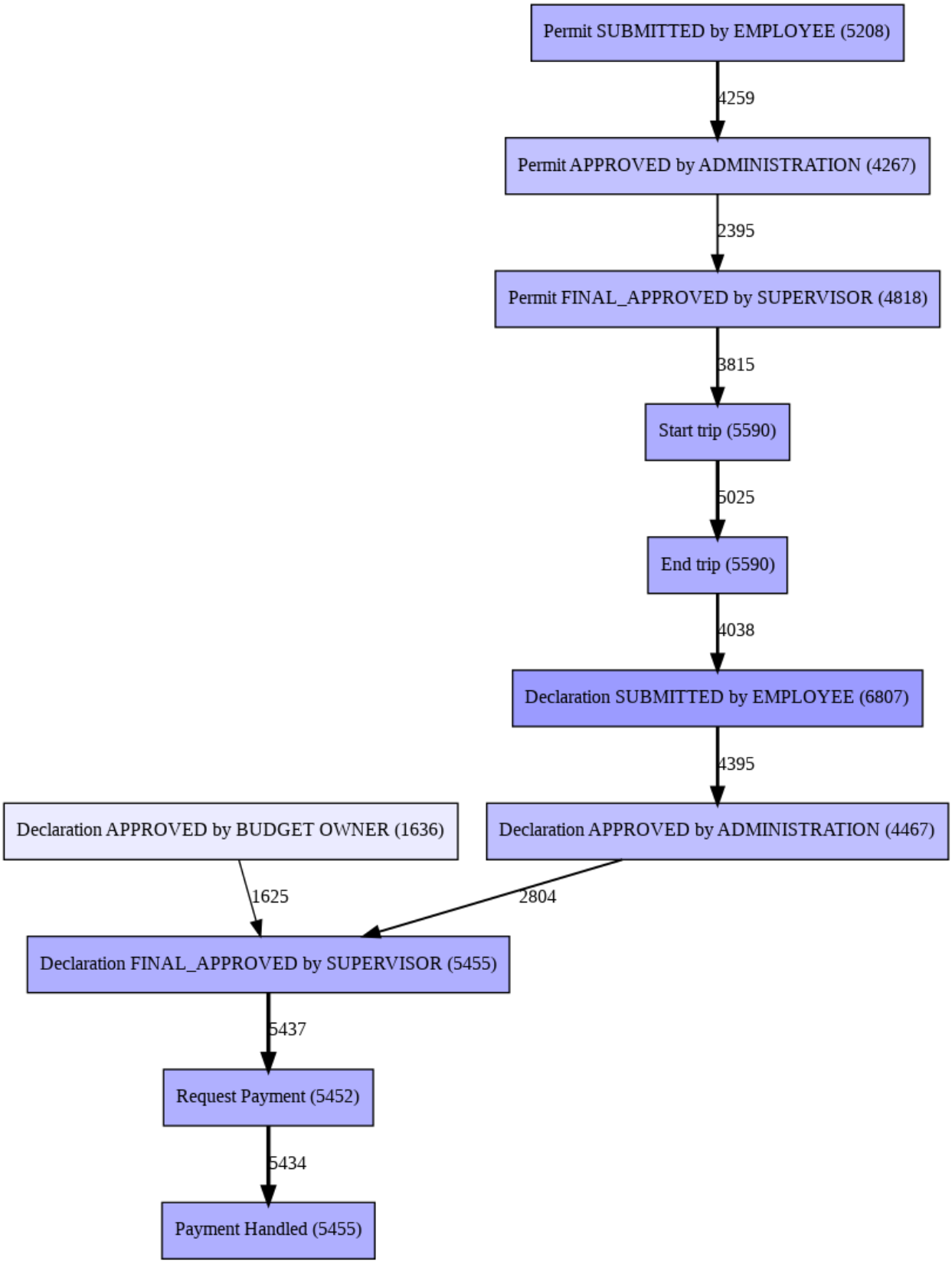
Permit SUBMITTED by EMPLOYEE (5208)
4259
Permit APPROVED by ADMINISTRATION (4267)
2395
Permit FINAL_APPROVED by SUPERVISOR (4818)
3815
Start trip (5590)
5025
End trip (5590)
4038
Declaration SUBMITTED by EMPLOYEE (6807)
4395
Declaration APPROVED by BUDGET OWNER (1636)
Declaration APPROVED by ADMINISTRATION (4467)
1625
2804
Declaration FINAL_APPROVED by SUPERVISOR (5455)
5437
Request Payment (5452)
5434
Payment Handled (5455)

Fig 7 - DFG

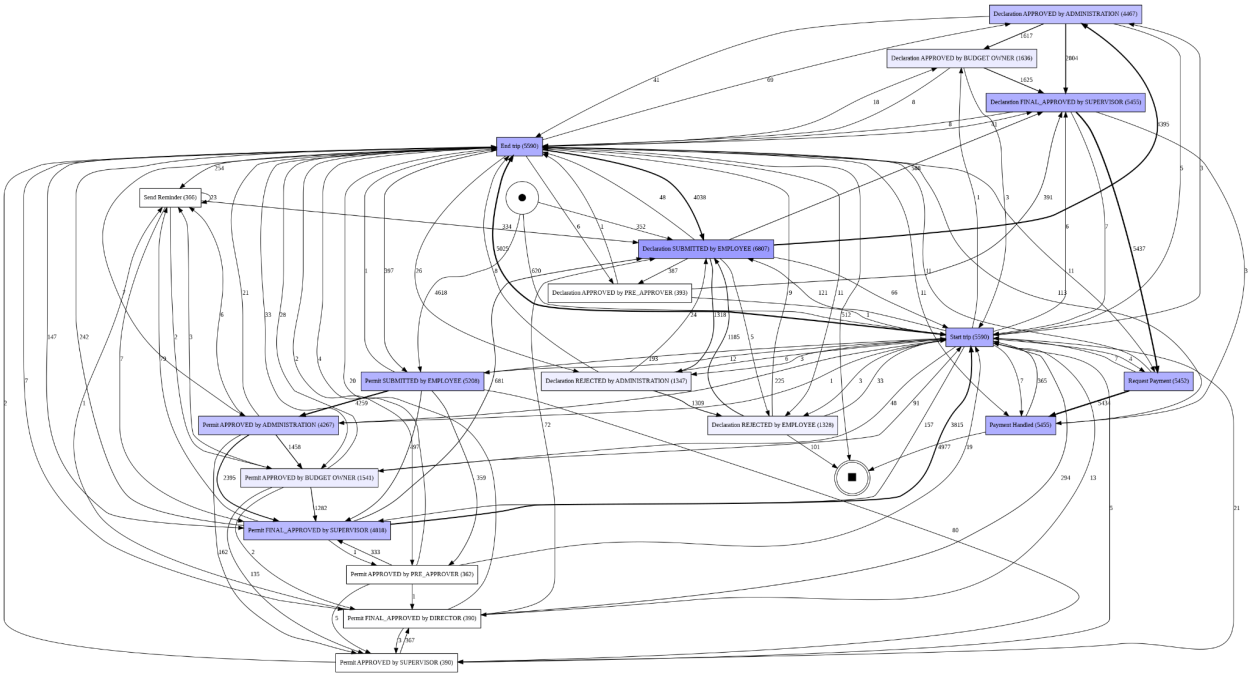

Fig 8 - Complete DFG

Diagrams descriptions in a simple way

This log contains the permit and the declaration sub-process. Ideally, the permit is approved before the trip starts. Nearly all permit related activities can occur after the event Start trip, which means that the international trip was taken without a finalized permission. There are even Declaration SUBMITTED by EMPLOYEE, before Permit SUBMITTED by EMPLOYEE although there are no approval steps, before the permit is submitted. Also, the payment of a declaration can occur before and after End trip, which is inconsistent with the textual description of the process. And there are payment requests that are followed by a Declaration REJECTED by MISSING. Moreover, there are payment requests that occur after the payment has been handled.

Conformance checking

In the previous steps the different models using multiple algorithms have been drawn in this check the performance of each model will be assessed and compared together. Table 4 shows the result of conformance checking for all models that we used.

Algorithm	Precision	Generalization	Simplicity	Fitness
Alpha miner	0.406	0.973	0.482	0.750
Heuristic miner	**0.979**	0.971	0.818	0.932
Inductive miner	0.476	0.934	0.644	1.0

Table 4

As demonstrated in table 4, there is a huge difference in the precision of heuristic miner algorithm, Also in comparison with other 3 metrics, this algorithm has better performance in this dataset.
The bellow figure shows the token replay of the heuristic model:

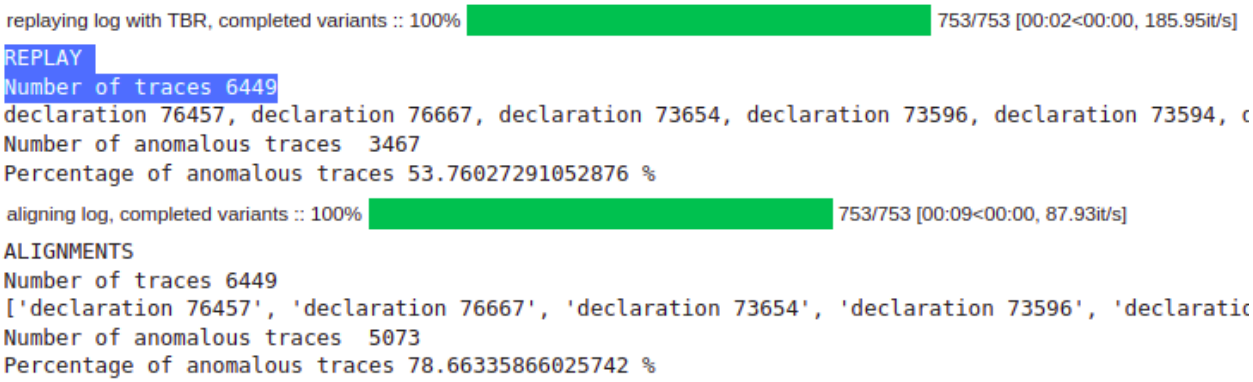

Fig 9 - conformance checking of the heuristic model.

Performance Enhancement

We can analyze and improve the process in many different aspects but for example, if we check the heuristic model(which is the most accurate one), we can understand easily some sections can be improved. However, Fig 6 shows the heuristic miner graph(which is the most accurate model for our dataset), we can understand that about 20% of the REJECTED cases are resubmitted by the employees which is making it a bottleneck and loop in the system so we can improve this section to make the process flow much cleaner than before by extracted many reject statuses and their related reason so if the reject is not reviewable we can prevent the employee from resubmitting the old request and consuming the system performance.

Q & A

Q1: What is the throughput of a travel declaration from submission (or closing) to paying?

The majority of the travel declarations have been paid and a throughput time can be calculated. For the international declarations, 95.94 % of the 6,449 declarations have been paid. The remaining cases are either never approved or just saved as drafts that have not been submitted. To measure the duration of one declaration, the mean of the durations is calculated and it shows a request from submitting to handle status takes about **86 days and 8 hours**.

Q5: Where are the bottlenecks in the process of a travel declaration?

To determine the bottlenecks in the process of a travel declaration, we calculated the duration of activity for each case. Based on this, we calculated statistics over all cases including the mean. We select the activities with the highest bottlenecks based on the median and not based on the mean duration, since we do not want to put too much emphasis on outliers. Furthermore, it is noticeable that the activities that start the trip and end the trip have a very mean duration.

Q7: How many travel declarations get rejected in the various processing steps and how many are never approved?

To answer Q7, we had to examine all activities that suggest a rejection or approval of the declaration. Therefore, we took all activities into account including the terms ”reject” and ”approved”. To this end, we firstly computed a case activity matrix that counted the occurrence of each activity for each case. This allowed us to count the occurrences of rejection and approval activities. We again analyzed international declarations. For the international declarations, we considered approvals and rejections both during the permit request and the travel declaration. We started again inspecting the activities that contain ”REJECT” or ”APPROVED” and counted their occurrences. Then we had a look at the cases. Out of the 6,449 cases, in 5,550 the permit and the declaration had been approved. For 800 cases the permit and the declaration has not been approved. For 279 cases the permit has been approved but the declaration has never been approved and for 406 the declaration has been approved without the approval of the permit.

Conclusions

The business process observed was decomposed and analyzed using various methods of process mining. Variant analysis and process discovery allowed us to answer the questions of the challenge. International declarations had more variety in activities. Many variants were presented in only one case. The process discovery step was held by creating data models. Conformance analysis showed the best models for international declarations. The best model for international declarations was a heuristic miner with **0.93** fitness and **0.97** precision. Though accurate filtering was produced using pm4py methods, leaving only successful traces, the alpha miner showed a low performance of **0.75** fitness. Alpha miner with decreased number of traces showed **0.93** fitness. There are a lot of different rejections in international cases, held by multiple organizational entities, projects, and other classes of declarations. Declarations, that were submitted more than 2 months after the end of a trip, were rejected only by one organizational unit, making up **52** percent of the total performance of this unit. The most important bottleneck is related to approvals made by different organizational entities. Budget owner approvals take additional time for international trips, making about half and a quarter of all approvals correspondingly take more time than direct approval by the administration and then final approval by the supervisor. Probably it might be possible to reduce approval time at this point, or a number of declarations sent for approval in the budget owner unit.

References

- Project Github: https://github.com/sirAlireza/bpi-2020