BPI Challenge 2020 Submission

—

Following the money: An exploratory research into the process for international declarations using process mining

Wessel van Bakel 5783615, Rose Mary Hulscher 4272978,
Mitchell Klijs 5863872, and Martijn Sturm 4141369

Utrecht University, Department of Informatics
{w.t.h.bakel, r.m.hulscher, m.klijs, m.j.sturm}@students.uu.nl

Supervisors: Dr. J. Gulden and Dr. ir. X. Lu

Abstract. The annual Business Process Intelligence (BPI) challenge allows students to test their process mining skills: analysing process data and trying to derive valuable insights. This year, 2020, the process data captures the reimbursement process for travels with requests and declarations at TU/e. This paper reports the results of our process analysis, done on the *international declarations* and *permit log*, which revolved around three main questions. We analyzed (i) the characteristics of overspend and underspend permits; (ii) possible new standardization for groups of declarations by amount; (iii) the relation between attributes and highlighted the most interesting and valuable insights. The main recommendations include: look into (i) the differences between overspent and underspent projects, (ii) differences between organisational units. Furthermore, employees should be encouraged to better think about the requested amounts. Additionally, the procedure should be changed the have a different process per budget amount group.

Keywords: process mining, process discovery, process analysis, data analytics, declarations, BPMN, Disco

1 Introduction

Processes are all around us and are present in any business environment. Nowadays, most of these processes are supported by information systems, which often record information about the execution of the processes into so-called event logs. Every event in the log contains information such as the activity performed, a case ID and the person who performed the action.

Process mining is the science that revolves around the analysis of these event logs. The goal of process mining is to collect useful information and gain knowledge about the executed processes and, if possible, improve or support the execution of the existing processes.

This paper is written as a submission to the tenth International Business Process Intelligence Challenge[1] in the student category. The challenge requires us to analyze real-life event data using process mining tools in order to gain insights into the processes captured in the event logs.

The data of this year's challenge originates from the TU/e (Technische Universiteit Eindhoven)[2]. The provided data revolves around the reimbursement of travel expenses. In this report, we positioned the amounts of declarations as central topic. We will analyze the following:

- What process related properties are correlated with the amount of money declared?
- What are the characteristics for overspend and underspend permits?
- Are there differences in the process based on the requested amount?

This report starts with an introduction to the data and the processes. Next, the questions mentioned above are answered in their own chapter. Throughout these chapters, recommendations are made. Concluding, these recommendations are summarized in the final chapter.

[1] https://icpmconference.org/2020/bpi-challenge/

[2] https://www.tue.nl/en/

2 The Data & The Processes

As stated in the introduction, the data is extracted from the reimbursement process at the TU/e. The data is published at 4TU Centre for Research Data[3]. It contains events from two organizational units of the TU/e in 2017 and of the whole TU/e in 2018.

2.1 The Data

The data consists of five event logs, namely:

- **Domestic Declarations:** For domestic declarations, no prior permission is needed and the reimbursement is paid after the cost are made.
- **International Declarations:** For international declarations, permission from the supervisor is needed before arrangements can be made. Permission can be obtained by filling in a travel-permit form.
- **Prepaid Travel Cost:** Prepaid travel costs can be reimbursed before the start of an international trip, after the travel permit is approved.
- **Travel Permits:** The travel permit log contains all events related to the travel permits (for the international trips) and prepaid travel cost.
- **Requests for Payment:** non travel-related reimbursements

As said in the introduction, this report focuses on the different level of costs for international trips. Therefore the *domestic declarations* and *requests for payment* logs in the rest of this report are disregarded.

All data in the event logs is anonymized in such a way that there are no TU/e internal IDs visible in the dataset. All staff members are replaced by their roles. Furthermore, all (payment) amounts in the logs are slightly changed. Summing declarations referring to the same travel permit and then comparing them to the original budget is still possible. On a large enough sample the summarized amounts should also be roughly correct.

The data contains three main concepts *a permit*, *a declaration* and *a payment*. The permits and declarations can be *submitted*, *rejected*, *approved* or *final approved*. The payment can be *requested* and *handled*. This information is included as event attributes.

Also, the actor is often registered with the activity. Unfortunately, sometimes the actor is stated as 'missing'. We assume this was done to anonymize the data. Actors involved in the process are *employee*, *supervisor*, and *director*.

Permits The concepts for the Permit log and the important attributes are presented in Figure 1. (Since the permit log contains all information about international declarations, the concepts overlap). Every *permit* corresponds to exactly one *organizational unit*, *project*, and *budget*. The permit can have multiple *requests for payment*. The project may have multiple *budgets* and the budgets may refer to multiple *projects*.

[3] `https://data.4tu.nl/repository/uuid:52fb97d4-4588-43c9-9d04-3604d4613b51`

International declarations For the international declarations log, each case represents a single *declaration*. Multiple *declarations* can belong to a single *permit* (Figure 1). Some case properties of the declarations are of interest to us. The most important property for this paper is the *amount* property. This resembles the amount of money that is to be reimbursed.

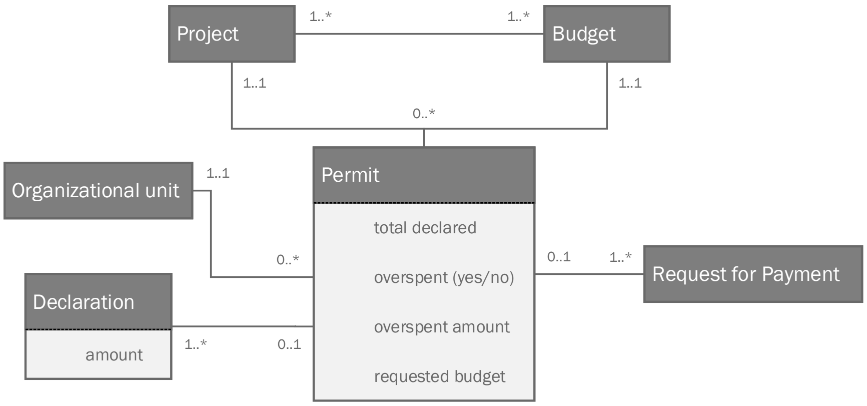


Fig. 1: Datamodel for the main concepts for the international declarations and the permits

2.2 The Processes

In this subsection the process for international declarations and travel permits will be presented. All processes follow a very similar process flow. Important to note is that the processes changed slightly from 2017 to 2018. Therefore, in some of the analysis steps performed in this report only the cases from 2018 were taken into account.

International Declarations For international declarations, the process can be summarized as follows. First, the *employee* that is going to make a trip has to *submit a permit*. This permit needs to be *approved* by the *travel administration* and *supervisor* before the employee can *start his trip*. When the employee's trip has finished, he can *submit declarations* that he wants to be reimbursed. These declarations need to be *checked* and *approved* by *staff members*, before the reimbursement is *handled* in form of a payment. This happy-flow is visualised in Figure 2. If the permit is *rejected*, the employee either *re-submits* the request or *discards* the request.

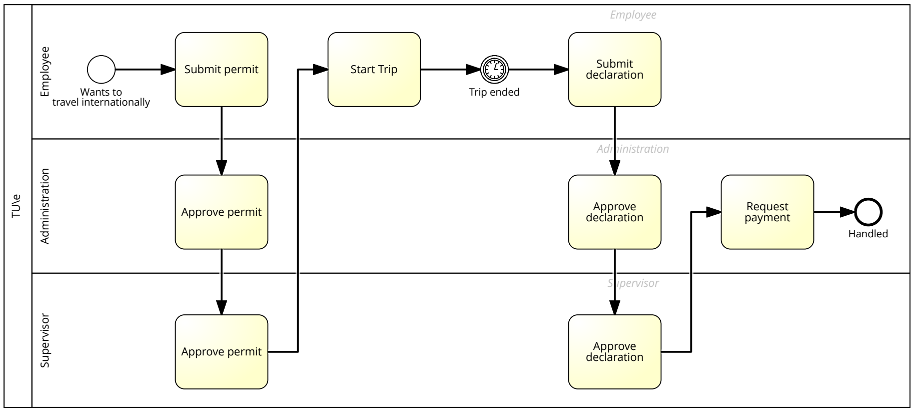


Fig. 2: BPMN model of the happy-flow of the international declarations process

We can also try to discover this process from the International Declarations event log. Figure 3 represents the process model as discovered by Disco[4]. Here, we've filtered on the start-dates of the processes to be after January 1st 2018 (to exclude the cases that went through the 2017-process). We've simplified the model as much as possible to only show the most frequent traces and paths (activities: 0%, paths: 0%).

When we inspect this discovered model, we can identify the happy-flow in the model. However, we can also immediately see deviations from the happy-flow. For example, there exists a flow between *end trip* and *permit submitted*. This should not be allowed, because the permit should be approved even before the trip may start.

Travel Permits The process flow of the travel permits is similar to the process flow of the international declarations and was derived in the same way. It is presented in Figure 4.

[4] https://fluxicon.com/disco/

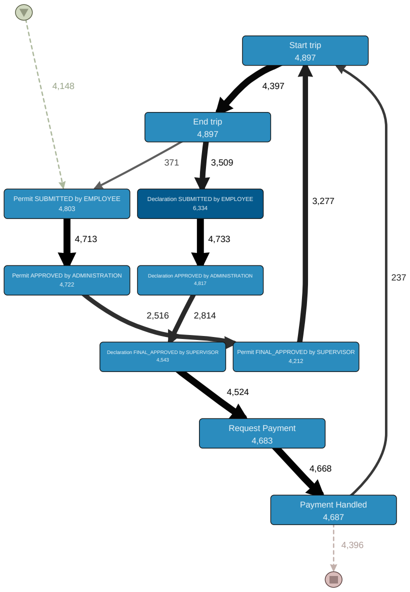

Start trip
4,897
4,148
4,397
End trip
4,897
371
3,509
3,277
Permit SUBMITTED by EMPLOYEE
4,803
Declaration SUBMITTED by EMPLOYEE
6,334
4,713
4,733
237
Permit APPROVED by ADMINISTRATION
4,722
Declaration APPROVED by ADMINISTRATION
4,817
2,516
2,814
Declaration FINAL_APPROVED by SUPERVISOR
4,543
Permit FINAL_APPROVED by SUPERVISOR
4,212
4,524
Request Payment
4,683
4,668
Payment Handled
4,687
4,396

Fig. 3: Basic process model for the international declarations process

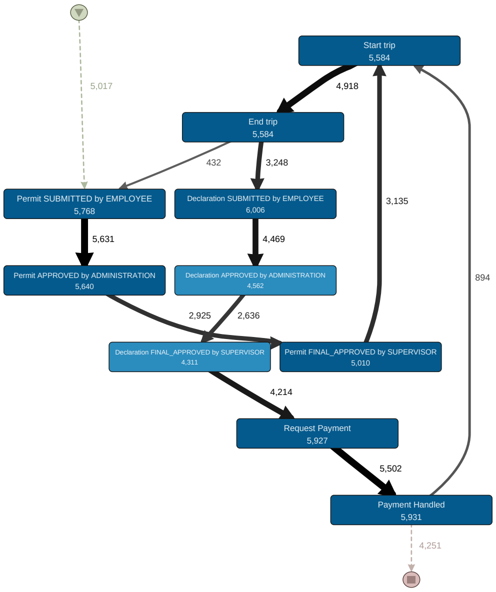

Start trip
5,584
5,017
4,918
End trip
5,584
432
3,248
3,135
Permit SUBMITTED by EMPLOYEE
5,768
Declaration SUBMITTED by EMPLOYEE
6,006
5,631
4,469
894
Permit APPROVED by ADMINISTRATION
5,640
Declaration APPROVED by ADMINISTRATION
4,562
2,925
2,636
Declaration FINAL_APPROVED by SUPERVISOR
4,311
Permit FINAL_APPROVED by SUPERVISOR
5,010
4,214
Request Payment
5,927
5,502
Payment Handled
5,931
4,251

Fig. 4: Basic process model for the travel permit request process

3 Process properties (International Declarations)

In this section we focus on the International Declaration cases from the International Declaration log file. Each declaration can be linked to a permit, as is described in section 2.1. In this section, we look at certain characteristics of the process of each declaration and how this relates to the amount of money being declared.

The following research question was defined: *What process related properties are correlated with the amount of money declared?* This question is answered by plotting the distribution of the amounts, and calculating correlation coefficients between the amount and a single process related property. This section is of exploratory nature, in which we try to deduce some process rules from the data at hand.

3.1 Method

As described in section 2.1, the data contains some activities that are performed on the declaration cases. We can determine for each case: the number of times a particular activity was performed, and whether the activity is performed at all. The same is applied to the actor performing the activity. Metrics are generated on how many times a particular actor was involved per case.

First, declarations with amounts of 0 were discarded. Then, to obtain the new case properties, the *xml.etree.ElementTree* Python package[5] was used. This allowed us to mine the frequencies of activity attributes from the international declarations log.

Boxplots and bar charts were generated using the *Plotly*[6] package. Finally, Spearman's correlation was used to determine the correlation between the case's generated properties and the amount. This because the amounts are related to frequency variables, which are not normally distributed. So Pearson could not be used.

3.2 Analysis

Exploration First, some exploratory statistics about the process were generated. The counts of certain activities are visualized in Table 1.

The counts for the declaration and permit numbers sections relate to the activity types. The counts in the actor section relate to by whom the activity was performed. The total number of declarations that contain a non-zero amount is 6175. This corresponds to the number of declaration submitted activity (as can be seen in the table).

We can identify some activities that are nearly always performed per declaration. This is the case for the activities (Declaration and Permit) that have a number in the Declarations row close the 6175. For both the permits and the

[5] `https://docs.python.org/3/library/xml.etree.elementtree.html`

[6] `https://plotly.com/`

Declaration				
	Approved	Final Approved	Rejected	Submitted
Declarations	5510	6174	1393	6175
Total	7680	6261	3294	7874
Permit				
	Approved	Final Approved	Rejected	
Declarations	5191	5751	245	
Total	7441	5771	490	
Actor				
	Supervisor	Director	Employee	
Declarations	6175	665	6175	
Total	12236	870	15695	

Table 1: Frequency of the activities and actors in the International Declarations log (0 amount declarations omitted). The Declaration rows denote how many declarations contain the activity / actor at least once. The Total rows denote the total number of activities / actor summed over all declarations.

declarations, rejection activities are much more infrequent than approvement activities.

For the actors, at least one activity per declaration is done by a *supervisor* and an *employee*. This is in sharp contrast to the *director* actor. Only about one in ten declarations involve activities performed by *directors*. Therefore, it could be the case that the *director* will only get involved if declarations fulfill an unknown property. This property might be cost-related.

Number of activities Figure 5 shows the distribution of the amount corresponding to the number of activities per declaration. Most boxes in the boxplot show a right-skewed distribution. This indicates that far more variance is present in the higher amounts than in the lower amounts. However, for the righter boxplots, this is less apparent. Also the boxplots for the higher number of activities show that these have higher first-quartile margins.

Overall, based on Figure 5, we can say that for traces that contain less activities, the amounts are more frequently lower, than for higher number of activities traces. This is also apparent in Figure 6. The column denoting the number of activities (n) shows a positive correlation with amount. This means that declarations for which many activities have been performed are concerned with higher amounts of money.

Insight: Declarations with higher amounts have longer process traces and are thus more frequently evaluated.

The most likely explanation for this finding is that declarations with higher amounts are treated with more care, and thus go through a longer process. Of course, the risk (money lost) involved in approving a faulty declaration is much

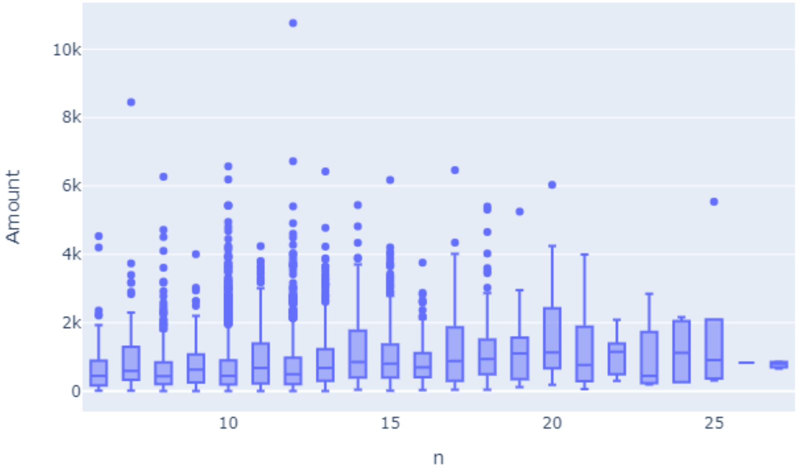


Fig. 5: Distribution of Amount per number of activities

higher with a higher declaration amount. Therefore, during reviewing steps, the reviewers are possibly more inclined to reject a request if having doubt. That would lead to an additional reviewing step before a declaration can be finalized. Hence, leading to a higher number of activities associated with such a declaration.

Actors The correlation between the type of actors that performed activities for a declaration, and the amount of a declaration are shown in Figure 6. The first thing to note is that for every actor type, the correlation is positive. Meaning that if more activities are conducted by a particular type of actor, the declarations amount is expected to be higher. This was expected, since the previous finding was about a higher number of activities being correlated to a declaration's amount. Though what interesting is here, is the difference between the type of actors. The number of activities involving a director is highest correlated with a declaration's amount. Contrary, the supervisor is the opposite.

Insight: The declaration amount and the director are highly positively correlated. I.e.: the director is more involved in declarations that have higher amounts.

We can make some speculations about why this is the case. A director is a high level function within an organisation. The organisational rules might be that for high impact cases, a director needs to be involved. Here, the high

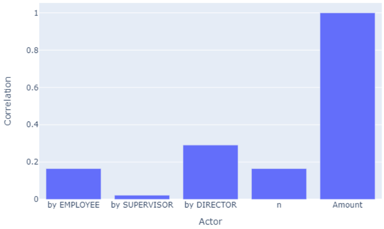


Fig. 6: Correlation between a declaration's amount (y-axis) and the frequency of activities performed by a type of actor (EMPLOYEE, SUPERVISOR, DIRECTOR). Also the correlation between amount and number of activities (n) is shown.

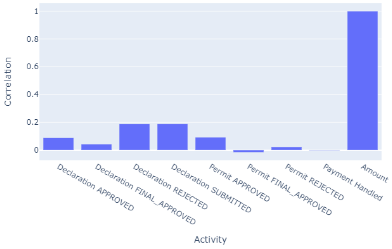


Fig. 7: Correlation between a declaration's amount (y-axis) and the frequency a certain type of activity is performed for that declaration.

impact is derived from the declaration having a high amount of money. Contrary, the supervisor might be involved in all declarations. Therefore, they are not correlated with a higher amount of money being declared.

Activity types In Figure 7 the correlation between the number of a particular type of activity and the amount of the declaration is shown. Most correlations are positive, indicating the more activities of that type appear together with higher declaration amounts. However, the correlations are also small. For both declaration REJECTED and declaration SUBMITTED, a correlation of approximately 0.2 is found. This indicates that clearly not all of the variance of amounts between declarations can not be attributed to the frequency of certain activity types.

4 Overspend permits

This section investigates the permits associated with international declarations. The goal of this exploration is to investigate what permits went over-budget or stayed under-budget, based on the value of the attribute "overspent". Additional measures can be taken based upon these findings. For example, to review certain permits or to evaluate the process for a specific organisation unit. This resulted in recommendations to the TU/e. Thereby the following research question is answered: *What are the characteristics of under-spend and overspend permits?*

For this exploration, the permit log was used. The permit log contains information about the declarations belonging to that permit, if they are overspend (more spend then requested) and by how much (as introduced in Figure 1). A project can have multiple permits and an organisation can have multiple permits and projects (as can be seen in Figure 1). Thus, the projects and organisational units are also investigated, to see what projects costs the most money, or which organisational units are over- or underspending.

4.1 Data preperation

The permitlog was filtered using Disco in the following ways:

- **Time:** The time frame was trimmed from 01-01-2018 00:00:00 to 01-01-2019 00:00:00, because of three reasons: 1) there were a lot of empty values, 2) the process was changed in 2018 and 3) one year is a concrete amount to investigate and has enough cases.
- **Activity:** The activity *payment handled* was selected to be mandatory; to filter out any traces that didn't go through with the payment-process. Hereby it is assumed that this activity actually handles the payment. As there can be multiple activities *payment handled* for one case (which was the case 17% of the time), only one activity was selected; because the focus is on the *full amount* and the partial payments (declarations) is out of scope. This resulted in 65% of the cases and 72% events.

The permitlog was then exported to a CSV-file based on the *overspend* and *overspend_amount* attributes. It was split into three levels; the *underspend* ($overspend\_amount < 0$), *on-budget* ($overspend\_amount = 0$) and *overspend* ($overspend\_amount > 0$) permits.

4.2 Data Exploration

After the creation of the CSV-files, the log was examined using R. The goals of this explorations were the following:

- **Gain insight into the data:** Before the main research question can be investigated, it is important to first gain insight into the data and look at the ranges of different attributes and look at abnormalities in the data.
- **Outlier identification:** Before more processing is performed, it is important to identify the outliers. Existing outliers that are detected can either signify an error, or an extreme case that needs to be investigated further.

4.3 Research questions

The following descriptive research questions were answered, separated into various categories:

1. **Travelpermits:** How many travelpermits are there and what are their characteristics? Are there any outliers?
   - For the outliers: what are their characteristics? What is the process that can be identified?
2. **Organizational units:**
   - Which of the organizational units has the most under-budget, on-budget or over-budget travel permits relative to their total amount of permits?
   - Which of the organizational units has the most under-budget, on-budget or over-budget travel permits based on the total amount of permits in Euros?
3. **Projects:** Which of the projects has the most under-budget, on-budget or overbudget travel permits? And to which organizational units to they belong?
   - What are the top 7 underspend, on-budget and overspend projects?
   - Which organizational unit has the most under-budget or overbudget projects?

4.4 Results

This section presents the results of the research questions presented in the previous section.

Outlier All overspend amounts were plotted in a boxplot, displayed in Figure 9. As can be seen, one point has a much larger overspend amount. This point corresponds to travel permit 54518. The employee that submitted this travel permit requested a budget of 2.000 euros. However, he declared a total amount of nearly 1.500.000 euros. This results in an overspend amount of almost the same number. This declaration belongs to organizational unit 65466.

Figure 8 displays the process that this case went through. If we compare the process model to the happy-flow presented in Figure 4, we can see that the *Request for payment* flow is much earlier on in the process, before the trip started. This is not how the happy-flow specifies this. This flow even lacks the *Declaration* tasks. The average request takes 10 events with a throughput time of 73 days, the outlier contains 24 events and takes 181.3 days to end.

We do not know enough of the details of this process to make any conclusions about how this happened. In all further research questions this case was excluded.

Recommendation: It is recommended to investigate travel permit 54518 further, to determine appropriate steps to be taken.

When this outlier is excluded and we view the boxplot in Figure 10 there are no obvious additional outliers.

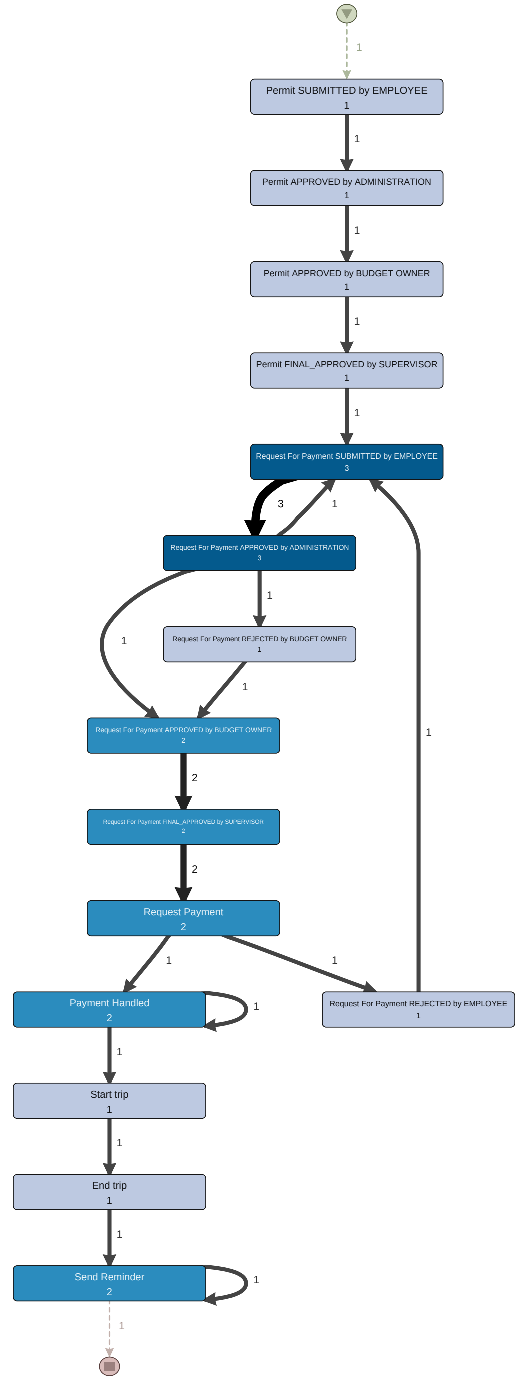

1
Permit SUBMITTED by EMPLOYEE
1
1
Permit APPROVED by ADMINISTRATION
1
1
Permit APPROVED by BUDGET OWNER
1
1
Permit FINAL_APPROVED by SUPERVISOR
1
1
Request For Payment SUBMITTED by EMPLOYEE
3
3
1
Request For Payment APPROVED by ADMINISTRATION
3
1
1
Request For Payment REJECTED by BUDGET OWNER
1
1
1
Request For Payment APPROVED by BUDGET OWNER
2
2
Request For Payment FINAL_APPROVED by SUPERVISOR
2
2
Request Payment
2
1
1
Payment Handled
2
1
Request For Payment REJECTED by EMPLOYEE
1
1
Start trip
1
1
End trip
1
1
Send Reminder
2
1
1

Fig. 8: Process model of the outlier

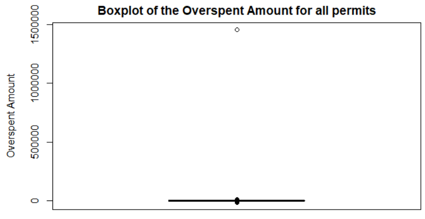


Fig. 9: Boxplot of the overspend amounts for all permits in the PermitLog

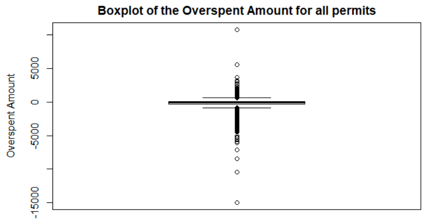


Fig. 10: Boxplot of the overspend amounts for all permits in the PermitLog (outlier excluded)

Organizational units In order to answer the research question, *Which of the organizational units has the most under-budget, on-budget or over-budget travel permits relative to their total amount of permits?*, two figures were created. Namely Figure 11, that shows the relative amount of permits based on the total amount for each organizational unit in a bar chart and Figure 12, which shows the same results, but only using percentages combined with the total amount. Organizational unit 65468 has the most overspend permits in percentage of their total amount of permits (87.5%). The most on-budget organizational unit was unit 65472 with 5.58%. The most underspend organizational units with 100% were units 65488, 65486, 65478 and 65462. Interesting to note is that they only had one permit. The highest count of permits was organizational unit 65455 with 831 permits, thereby 62% was underspend, and 36% overspend.

To answer the question, *Which of the organizational units has the most under-budget, on-budget or over-budget travel permits based on the total amount of permits in Euros?* The table presented in Figure 13 was created. It presents the sum of the underspend and overspend projects per organizational unit. The table is sorted descending by the last column: the total underspend/overspend for that organizational unit. As can be seen from the table, organizational unit 65468 is as a whole (based on the total column) the most overspend, namely around 2250 euros. Organizational unit 65458 on the other hand is the most underspend, namely around 202.000 euros. The most on-budget organizational unit is organizational unit 65486, they only went 4 euros under budget.

If we sum all the totals in the last column, we find that all organizational units combined underspend a total of 816.500 euros.

Projects The tables in Figure 14 presents an overview of the top 7 underspend (14a), overspend (14b), total (14c) and bottom 7 total (14d) projects.

If we look at the top overspend projects (Figure 14b), we can see that even though certain permits for that project were overspend, the total was always still positive (underspend).

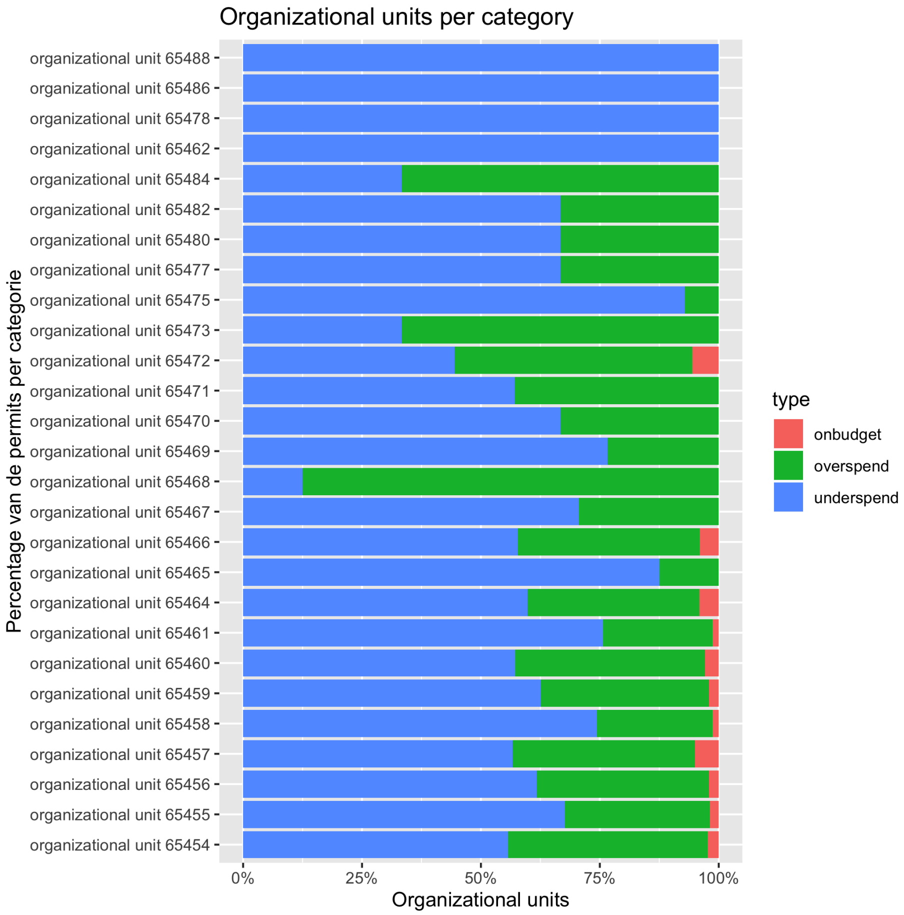
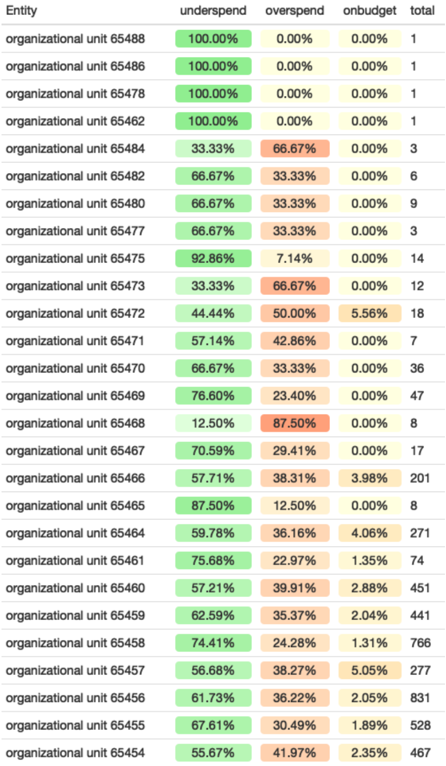


Fig. 11: The relative amount of permits based on the total amount for each organizational unit.

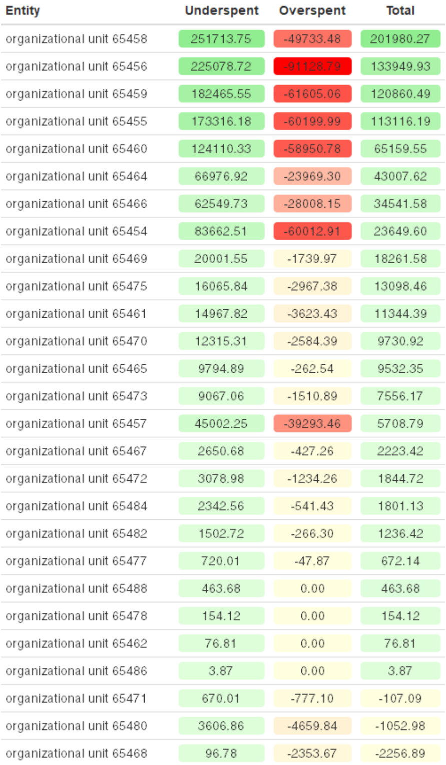
Recommendation: It could be very interesting to look further into the overspend and underspend projects. Was the budget incorrectly estimated and what can be done to make better estimations in the future? Were there unforseen circumstances?

It is noteworthy that organizational unit 65458 appears three times in the top 7 underspend projects. Organizational units 65456 and 65460 appear respectively four and two times in the bottom 7 total underspend projects (so they were overspend).



Entity	underspend	overspend	onbudget	total
organizational unit 65488	100.00%	0.00%	0.00%	1
organizational unit 65486	100.00%	0.00%	0.00%	1
organizational unit 65478	100.00%	0.00%	0.00%	1
organizational unit 65462	100.00%	0.00%	0.00%	1
organizational unit 65484	33.33%	66.67%	0.00%	3
organizational unit 65482	66.67%	33.33%	0.00%	6
organizational unit 65480	66.67%	33.33%	0.00%	9
organizational unit 65477	66.67%	33.33%	0.00%	3
organizational unit 65475	92.86%	7.14%	0.00%	14
organizational unit 65473	33.33%	66.67%	0.00%	12
organizational unit 65472	44.44%	50.00%	5.56%	18
organizational unit 65471	57.14%	42.86%	0.00%	7
organizational unit 65470	66.67%	33.33%	0.00%	36
organizational unit 65469	76.60%	23.40%	0.00%	47
organizational unit 65468	12.50%	87.50%	0.00%	8
organizational unit 65467	70.59%	29.41%	0.00%	17
organizational unit 65466	57.71%	38.31%	3.98%	201
organizational unit 65465	87.50%	12.50%	0.00%	8
organizational unit 65464	59.78%	36.16%	4.06%	271
organizational unit 65461	75.68%	22.97%	1.35%	74
organizational unit 65460	57.21%	39.91%	2.88%	451
organizational unit 65459	62.59%	35.37%	2.04%	441
organizational unit 65458	74.41%	24.28%	1.31%	766
organizational unit 65457	56.68%	38.27%	5.05%	277
organizational unit 65456	61.73%	36.22%	2.05%	831
organizational unit 65455	67.61%	30.49%	1.89%	528
organizational unit 65454	55.67%	41.97%	2.35%	467

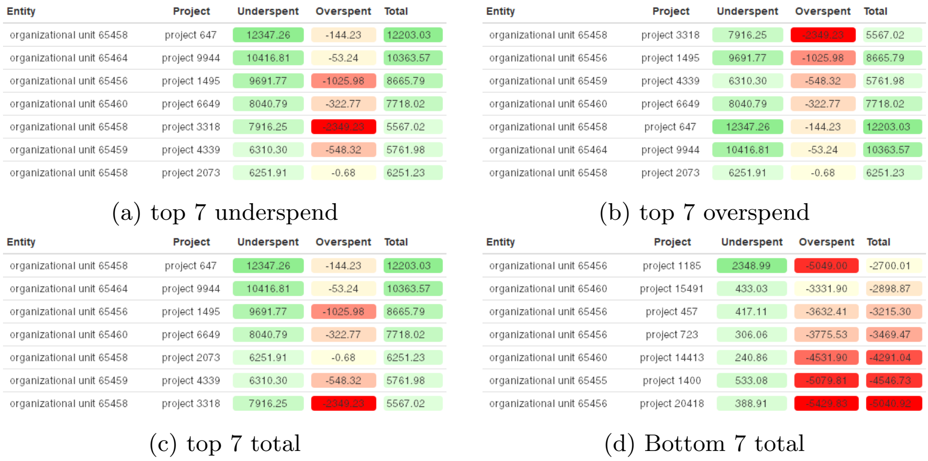
Fig. 12: The relative amount of permits based on the total amount for each organizational unit, combined with the total amount



Entity	Underspent	Overspent	Total
organizational unit 65458	251713.75	-49733.48	201980.27
organizational unit 65456	225078.72	-91128.79	133949.93
organizational unit 65459	182465.55	-61605.06	120860.49
organizational unit 65455	173316.18	-60199.99	113116.19
organizational unit 65460	124110.33	-58950.78	65159.55
organizational unit 65464	66976.92	-23969.30	43007.62
organizational unit 65466	62549.73	-28008.15	34541.58
organizational unit 65454	83662.51	-60012.91	23649.60
organizational unit 65469	20001.55	-1739.97	18261.58
organizational unit 65475	16065.84	-2967.38	13098.46
organizational unit 65461	14967.82	-3623.43	11344.39
organizational unit 65470	12315.31	-2584.39	9730.92
organizational unit 65465	9794.89	-262.54	9532.35
organizational unit 65473	9067.06	-1510.89	7556.17
organizational unit 65457	45002.25	-39293.46	5708.79
organizational unit 65467	2650.68	-427.26	2223.42
organizational unit 65472	3078.98	-1234.26	1844.72
organizational unit 65484	2342.56	-541.43	1801.13
organizational unit 65482	1502.72	-266.30	1236.42
organizational unit 65477	720.01	-47.87	672.14
organizational unit 65488	463.68	0.00	463.68
organizational unit 65478	154.12	0.00	154.12
organizational unit 65462	76.81	0.00	76.81
organizational unit 65486	3.87	0.00	3.87
organizational unit 65471	670.01	-777.10	-107.09
organizational unit 65480	3606.86	-4659.84	-1052.98
organizational unit 65468	96.78	-2353.67	-2256.89

Fig. 13: Underspend/Overspend by organizational units

Recommendation: It would be very interesting to look further into why certain organizational units have more underspend projects and others have more overspend projects.



Entity	Project	Underspent	Overspent	Total
organizational unit 65458	project 647	12347.26	-144.23	12203.03
organizational unit 65464	project 9944	10416.81	-53.24	10363.57
organizational unit 65456	project 1495	9691.77	-1025.98	8665.79
organizational unit 65460	project 6649	8040.79	-322.77	7718.02
organizational unit 65458	project 3318	7916.25	-2349.23	5567.02
organizational unit 65459	project 4339	6310.30	-548.32	5761.98
organizational unit 65458	project 2073	6251.91	-0.68	6251.23

(a) top 7 underspend

Entity	Project	Underspent	Overspent	Total
organizational unit 65458	project 3318	7916.25	-2349.23	5567.02
organizational unit 65456	project 1495	9691.77	-1025.98	8665.79
organizational unit 65459	project 4339	6310.30	-548.32	5761.98
organizational unit 65460	project 6649	8040.79	-322.77	7718.02
organizational unit 65458	project 647	12347.26	-144.23	12203.03
organizational unit 65464	project 9944	10416.81	-53.24	10363.57
organizational unit 65458	project 2073	6251.91	-0.68	6251.23

(b) top 7 overspend

Entity	Project	Underspent	Overspent	Total
organizational unit 65458	project 647	12347.26	-144.23	12203.03
organizational unit 65464	project 9944	10416.81	-53.24	10363.57
organizational unit 65456	project 1495	9691.77	-1025.98	8665.79
organizational unit 65460	project 6649	8040.79	-322.77	7718.02
organizational unit 65458	project 2073	6251.91	-0.68	6251.23
organizational unit 65459	project 4339	6310.30	-548.32	5761.98
organizational unit 65458	project 3318	7916.25	-2349.23	5567.02

(c) top 7 total

Entity	Project	Underspent	Overspent	Total
organizational unit 65456	project 1185	2348.99	-5049.00	-2700.01
organizational unit 65460	project 15491	433.03	-3331.90	-2898.87
organizational unit 65456	project 457	417.11	-3632.41	-3215.30
organizational unit 65456	project 723	306.06	-3775.53	-3469.47
organizational unit 65460	project 14413	240.86	-4531.90	-4291.04
organizational unit 65455	project 1400	533.08	-5079.81	-4546.73
organizational unit 65456	project 20418	388.91	-5429.83	-5040.92

(d) Bottom 7 total

Fig. 14: Underspend/Overspend by projects

5 Groups

Instead of looking at the overspend amount relative to the organization unit, in this chapter we will be looking at the relation between the overspent amount and the requested budget, how they differ in terms of process and possible recommendations. Thereby the following research question is answered:
When splitting the data into groups relative to the requested budget, can we see differences in the process models in terms of throughput time, deviations and probabilities of getting approved?

Currently, as specified by the TU/e, the permit process is uniform and not different based on the requested budget amount. A higher amount of requested budget probably gets checked by more and higher-level management, however it appears there are no established, standardized procedures. This research aims to help discovering possible differences when grouping based on requested budget and wants to advice the TU/e university where to better their processes.

5.1 Method

We continue to utilize the events in the permit log and the filters used in the previous chapter, which are the cases in the year 2018 and the mandatory activity *payment handled*. Since this research question uses both Disco and Rstudio, this research will use two variants of the PermitLog.xes file. One variant will be kept unedited in the Disco Tool to be used for process analysis and the other variant will be exported, prepared and used in Rstudio to perform data analysis.

1. Preparing the Data The first step is preparing the Rstudio variant with converting all values to their respected data type , and deleting unnecessary columns (for this research question) like activity type, cost type, declaration type, rfp_id, task and dec_id. Since all the activities in the same case IDs will have the same attribute values, we deleted all duplicate case ID rows in Rstudio, so that every case ID only exists once. Figure 15 presents an overview of the frequencies relative to the requested budget.

2. Dividing into groups in R After the data was prepared, the following three groups were created by splitting the PermitLog log into three equal groups of 1500 cases relative to the requested budget: low, medium and high requested budget. Table 2 shows an overview of the grouped data. These groups will be used to analyse the differences in data and process. Notice the count does not add up to the total amount of unique cases, since we deleted the common outlier for this research question.

When we further explored the grouped data in R, we explored the relation to the overspent amount like the previous chapter. We divided the overspent amount by the requested budget and got a percentage of overspending amount

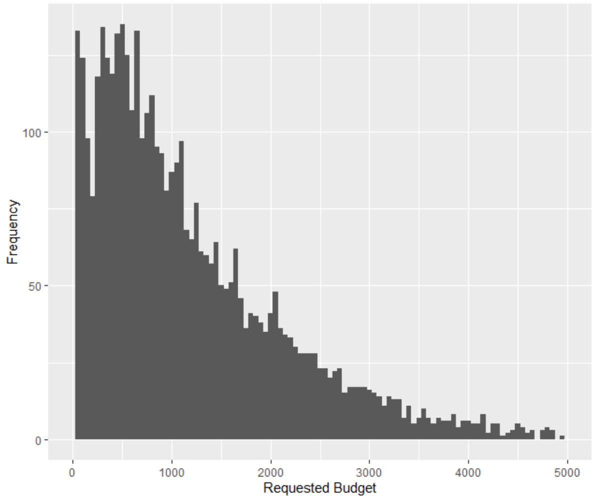


Fig. 15: Frequencies of requested budget

Group	Range	Count	Mean Requested
Low	0 - 588.7	1500	302.3
Medium	588.8 - 1389.7	1499	947.3
High	1389.8 - 13451.8	1500	2511.8

Table 2: Overview splitted PermitLog

relative to the budget group. We made some interesting findings alongside Figure 16:

- The **Lower group** only had 74 out of 1500 declarations on budget, where the overspent budgets averaged a 47% of overspending, and the underspent budgets -42%
- The **Medium group** performed worse on the amount of declarations on budget (23), but did slightly better with the the overspent budgets averaging a 28% of overspending, and the underspent budgets -34%
- Almost no declarations (8 out of 1500) in the **Higher group** was on budget and performed similar to the medium group where the overspent budgets averaged a 23% of overspending, and the underspent budgets -34%

Recommendation: Encourage the employee to think more about the needed/requested budget. Perhaps come with consequences when the overspent amount is above a percentage fault threshold. An average of 40% in under- and overspending in every group is bad for the business in predicting future costs.

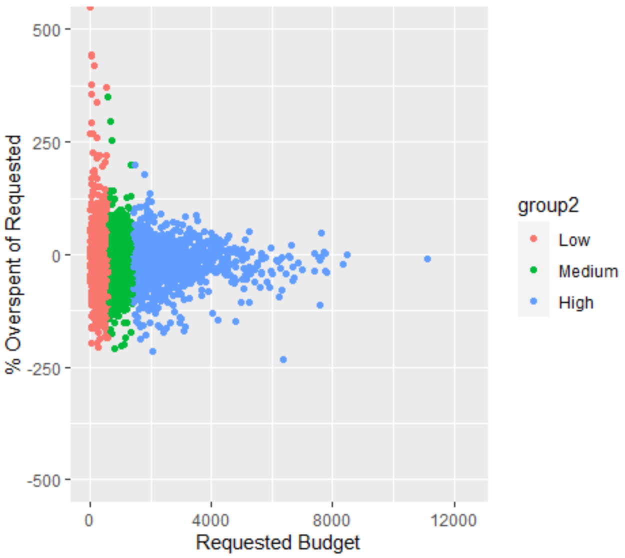


Fig. 16: Requested budget versus Percentage Overspent amount of Requested budget

5.2 Analysis

To find differences in the process of the identified groups we used Disco and filtered the requested budget attributes with respect to the ranges in table 2.

Throughput Time Throughput time was easily discovered by the average events, and the mean and median case duration in the global statistics tab in Disco. As table 3 shows there is a difference between the mean and median within the groups, because of unknown outliers and the skewed distributions of the groups. Therefore, when we analysed the duration time we preferred to use the median case duration and discovered clear differences between the groups. A lower requested budgets has on average a smaller case duration time than others.

Group	Average Events	Median case Duration	Mean case Duration
Low	11.1	46.7 d	61.5 d
Medium	13.6	64.9 d	74.4 d
High	15.6	87.1 d	14 wks

Table 3: Overview Throughput Time

Deviations To analyse deviations we must discover the desired process by analysing the given process flow on the BPI challenge website and appoint which activities are unwanted. Figure 4 shows the happy-flow process model and activities, but does not take in consideration that the budget owner and supervisor could be a different person. We checked two forms of deviations.

1. **Forbidding unwanted activities** We filtered the logs by forbidding all unwanted activities, such as REJECTED and SAVED
2. **The perfect Happy-flow** We added another filter on top of the previous one, were the ten activities in the happy-flow process (from Figure 4) where mandatory.

We divided the amounts of filtered cases by the total amount of cases per group and got the following results in table 4.There are only little cases that resemble the happy-flow process model and when comparing the groups we see that the high requested budget group performs worse with more unwanted activities than the other two groups.
Finally, we made the undesired activities mandatory to check for the most frequent variants of deviations, which were for all groups the same, the declarations were rejected by administration.

Group	Happy-Flow	No unwanted activities
Low	3.2 %	23.4 %
Medium	4.9 %	22.1 %
High	4.6 %	11.8 %

Table 4: Overview Deviations

Recommendation: Create a different standard procedure per requested budget amount group to optimize both duration and to minimize deviations.

Probability of getting approval Lastly, we researched the probability of getting the permit and declaration accepted. Permits and declarations go through administration first and after that to either a budget owner or a supervisor. Permits are approved more than 95% of the time for all groups after submitted by the employee and declarations are approved 81% in the low budget group and 71% in both medium and high.

After administration, several different activities can happen and Figure 17 and Figure 18 show the probabilities of the activities that can occur after approval by administration. Unfortunately, no significant differences were found.

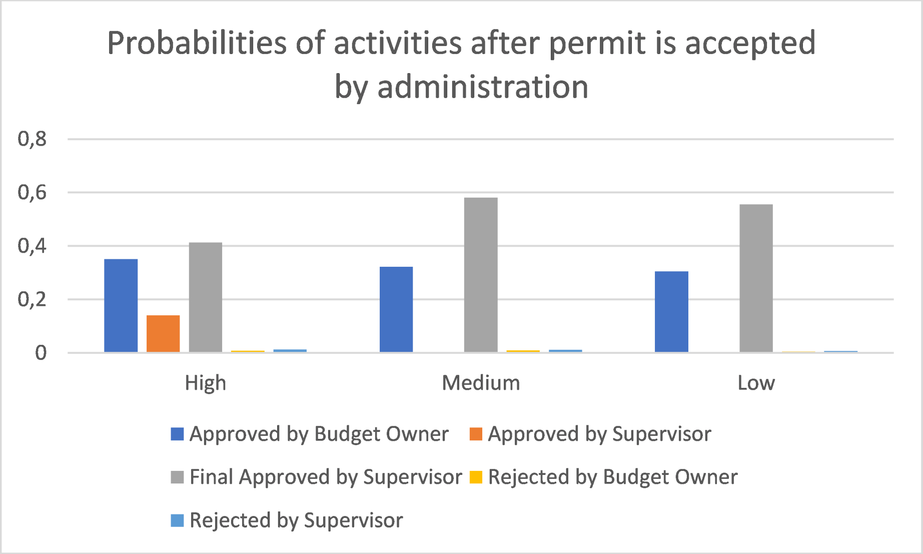


Fig. 17: Probabilities per activity after permit is accepted by administration

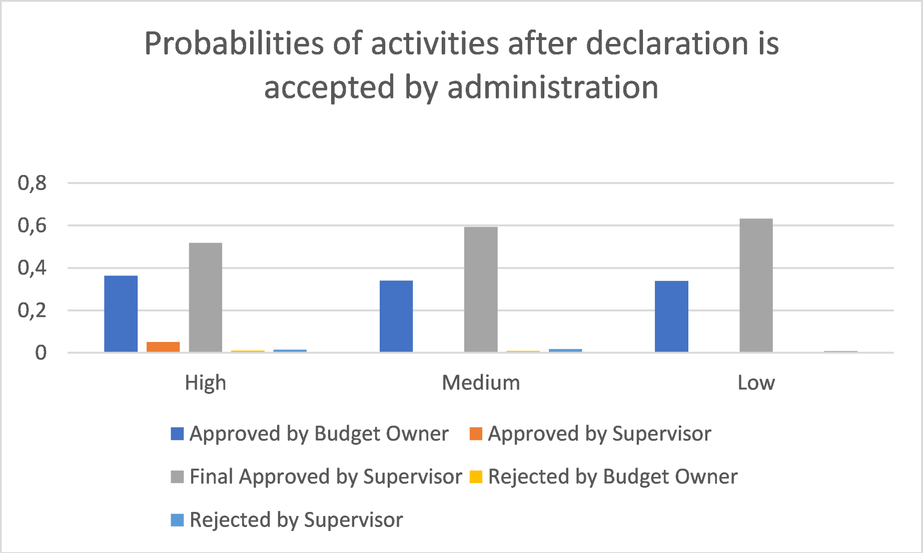


Fig. 18: Probabilities per activity after declaration is accepted by administration

6 Conclusion

The goal of this business report was to investigate the processes related to the international declarations at TU/e. The focus was to present recommendations to the TU/e on the amount of money declared or requested. The following research questions were compiled;

1. What process related properties are correlated with the amount of money declared?
2. What are the characteristics for overspend and underspend permits?
3. Are there differences in the process based on the requested amount?

In order to answer these research questions the following steps were taken; first of all, the relationships between event logs were identified. As a result, two event logs (domestic declarations and requests for payment) were disregarded as they were irrelevant for the determined scope. Secondly, the relationships within the chosen event logs were evaluated, to create a data-model that shows the relationships between entities and attributes.

In order to answer the second and third research questions, the following steps were performed; the events in the event logs were evaluated using R-studio, to identify outliers and determine relevant characteristics. Using Disco, the *happy flow* was identified, which was the dominant process flow. Various assumptions were made, which have to be taken into account with regard to the recommendations. For the exploration and evaluations various techniques were used such as histograms, bar charts, box plots, scatter plots and tables.

The process to investigate the process properties of each international declaration (first research question) was slightly different, as the goal of this analysis was exploratory and the results were insights rather than recommendations. First the characteristics of the activities performed on the declaration case were presented and some filtering took place. Python was used as the programming language to mine the frequencies of activity attributes from the international declaration logs. Boxplots and barcharts were used to visualise the results and for the correlation between the case's generated properties and the amount Spearman's correlation was used.

These processes resulted in the following recommendations;

- It is recommended to investigate travel permit 54518 further, to determine appropriate steps to be taken.
- It could be very interesting to look further into the overspent and underspent projects. Was the budget incorrectly estimated and what can be done to make better estimations in the future? Were there unforseen circumstances?
- It would be very interesting to look further into why certain organizational units have more underspent projects and others have more overspent projects.
- Encourage the employee to think more about the needed/requested budget. Perhaps come with consequences when the overspent amount is above a

percentage fault threshold. An average of 40% in under- and overspending in every group is bad for the business in predicting future costs.
- Create a different standard procedure per requested budget amount group to optimize both duration and to minimize deviations.

Additionally, we gathered the following insights;

- Declarations with higher amounts seem to be evaluated more often than declarations with lower amounts.
- Directors are involved more often in declarations with higher amounts than declarations with lower amounts.