# An In-depth Analysis of Reimbursement Processes Using Process Mining Techniques

## BPI Challenge 2020

Aydar Bulatov, Anna Sverkunova, Artem Glagolev, Elvares Geydarov, Sergey Kuznetsov, Elena Khomyakova, and Danil Smetanev

Sberbank of Russia, 19 Vavilova St., 117997 Moscow, Russia
{asbulatov, sverkunova.a.a, ayglagolev, geydarov.e.e, kuznetsov.s.nikolaevi, khomyakova.e.yu, smetanev.d.m}@sberbank.ru

**Abstract.** Expense reimbursement is an integral part of any kind of business travels. Nowadays this process is commonly used not only in companies, but in universities as well, in particular in Eindhoven University of Technology (TU/e). Since its staff takes a lot of trips, both domestic and international, a comprehensive analysis of the conducted reimbursement processes can be performed. In order to ensure that work-related transactions are administered and settled with efficiency, accuracy, and accountability to the satisfaction of both the University and the staff, one can use process mining, the most powerful technique to discover, monitor, and optimize real processes these days. This paper aims to examine the expense reimbursement process at TU/e by answering different questions provided by the process owner. By implementing state-of-the-art process mining tools, objective and deep insights into the process on all levels are revealed. These insights are then used to detect problems such as bottlenecks and deviations and identify potential opportunities for process improvement and optimization.

**Keywords:** BPI Challenge · Process Mining · Expense Reimbursement Process · Performance Analysis.

## 1 Introduction

The Business Process Intelligence (BPI) Challenge is a competition in the field of process mining conducted once a year since 2011. It seeks to showcase the opportunities and potential benefits offered by process mining techniques that allow organizations better audit, analyze, and improve business processes. By providing participants with a real-life event log, the BPI Challenge encourages researchers and practitioners to derive fact-based insights by answering both compliance-related and performance-related questions with the help of advanced process mining tools and solutions. This year, the data are collected from the reimbursement process at Eindhoven University of Technology (TU/e). While traveling to conferences, project meetings, or meetings up with colleagues from

other universities, the TU/e staff makes out of pocket expenses that are considered work-related and eligible for reimbursement. The TU/e compensates all personal expenses employees incurred in their performance of their job duties on behalf of the University. While the reimbursement process at TU/e seems to be straightforward and fairly well understood, a lot of objective and interesting insights can still be found. These insights can be thought of as the starting point of process optimization.

The scope of this report is on analyzing the expense reimbursement process captured in the event log mainly focusing on the list of questions provided in the challenge description. These questions are supposed to be answered with the help of process mining. As an innovative business process management instrument, it automatically performs process mapping, i.e. discovers the real process flow and creates visualization of the generated model. By analyzing the data-driven process graph and assessing process performance indicators, it is possible to understand of how things are working and where shifts are possible. In particular, different process weaknesses such as bottlenecks, loops, and deviations are identified. Process mining allows to zero in on these process inefficiencies that should be rethought or automated, and thus to uncover optimization potentials.

The paper is organized as follows. First, we explore five given datasets in Section 2 for the purpose of data and process understanding as well as data quality checking. Then, Section 3 presents the actual process model discovery for each of the event logs provided. Section 4 describes the results of performance analysis, before Section 5 defines opportunities for process enhancement based on derived insights. Finally, Section 6 sums up the conducted process analysis and draws the conclusions.

## 2 Data Description

To begin with, we took a close look at the data in order to identify patterns, exceptional cases, and potential problems. This stage of data exploration and understanding is of crucial importance for further analysis of the process. To examine different perspectives of the process including data, control flow, and performance, we used Sberbank's Python library for process mining *sberpm* here and elsewhere.

### 2.1 The Data

According to the challenge statement, the provided data from the expense reimbursement process at TU/e from 2017 to 2018 is split into travel permits and several request types, namely domestic declarations, international declarations, prepaid travel costs and requests for payment. Each of these categories is recorded in a separate event log. So, we have five datasets to analyze. Table 1 gives an overview of each data file.

Although we have to deal with multiple datasets divided into travel permits, declarations and requests for payment, they all contain typical process attributes, that is:

**Table 1.** Description of the datasets.

| | Domestic Declarations | International Declarations | Travel Permits | Prepaid Travel Cost | Requests for Payment |
|---|---|---|---|---|---|
| Events | 10,500 | 6,449 | 7,065 | 2,099 | 6,886 |
| Cases | 56,437 | 72,151 | 86,581 | 18,246 | 36,796 |
| Unique activities | 17 | 34 | 51 | 29 | 19 |
| Unique traces | 99 | 753 | 1,478 | 202 | 89 |
| Mean duration (days) | 11.5 | 86.5 | 87.4 | 36.8 | 12.0 |
| Median duration (days) | 7.3 | 66.0 | 71.7 | 24.0 | 8.2 |

- Case ID: Unique identifier of the event trace (*case:id* or *case:Rfp_id*),
- Activity: Name of the task executed (*concept:name*),
- Timestamp: Start time of the activity (*time:timestamp*),
- User: Anonymized performer of the activity (*org:role* and *org:resource*).

The three latter correspond to the event level. As for case level attributes (in addition to the Case ID), they differ between event logs, but tend to include the following ones:

- Requested Amount: Amount of the costs for a travel to reimburse,
- Budget Number: Budget that expenses are reimbursed from,
- Project Number: Project that request or permit relates to,
- Task Number: Task that request or permit relates to,
- Activity Number: Activity that request or permit relates to,
- Organizational Entity: Organizational unit that request or permit relates to.

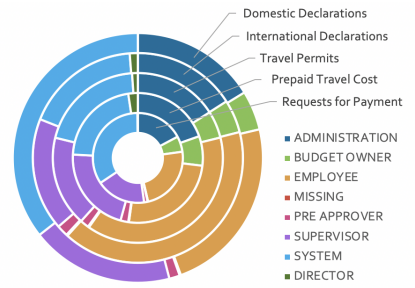


**Fig. 1.** Distribution of task performers in roles.

**Distribution in roles.** It is said in the BPI Challenge description that university staff members cannot be identified in the data. Instead, for all steps, the role of the person executed the step is recorded. If the role is UNDEFINED, then the resource recorded in the data is SYSTEM, otherwise STAFF MEMBER. So, we

decided to replace the role UNDEFINED to SYSTEM and compare the distribution of users in different datasets. Fig. 1 depicts the proportions of roles. As can be seen from the diagram, the top three most frequent task performers are SYSTEM, EMPLOYEE and SUPERVISOR. While the resource SYSTEM dominates in Domestic Declarations (36%) and Requests for Payment (34%) datasets, the role EMPLOYEE prevails in International Declarations (41%), Travel Permits (39%) and Prepaid Travel Cost (25%) logs. The least frequent category is MISSING (0.2%) followed by either DIRECTOR or PRE-APPROVER.

**Relation between datasets.** It is noticeable from the data attributes that some of the datasets are interrelated. In fact, International Declarations, Travel Permits and Prepaid Travel Cost are linked through the Permit ID column (*case:Permit ID* – *case:id* – *case:Permit id*). It is not surprising because all of these event logs refer to the reimbursement of expenses incurred by staff during international trips. A list of common attributes (started form the word Permit) also includes Project, Budget, Task, Activity Numbers and Organizational Entity. The number of different values on some of these columns is estimated in the hundreds. Besides UNKNOWN, the most frequent numbers are:

- Project: 426, 3442, 8761;
- Budget: 922, 698, 635;
- Task: 427, 13772, 648;
- Activity: 2169, 5818, 2401;
- Organizational Entity: 65458, 65455, 65456.

These attributes can also be on the case level, i.e. without the word Permit. For instance, Domestic and International Declarations contain Budget Number with 1 and 719 unique values respectively. In the latter case, the most frequent numbers are 146697, 146681, and 146666. Prepaid Travel Cost and Requests for Payment datasets, in their turn, have the four other columns recorded with the following values (except UNKNOWN) repeated the most times:

- Project: 503, 147556, 147546;
- Task: 14416, 25860, 48624;
- Activity: 505, 13776, 21016;
- Organizational Entity: 65458, 65461, 65463.

Despite the link between the above mentioned datasets, it seems reasonable for us to explore each dataset individually since there can be multiple requests for payment and declarations per permit.

## 2.2 Data Quality

The validity of the results from the process analysis depends largely on the quality of the data provided. So, here we outline some issues found during the data exploration phase.

- *Timeframe Inconsistency*. It is pointed out in the BPI Challenge description that the files contain data from 2017 and 2018. However, some datasets have events with timestamps beyond this timeframe. Hence, we decided to delete cases dating back to 2016.
- *Duplicate Rows*. Duplication is one of the main data quality issues since it causes overload and provides misleading results (e.g. inaccurate performance indicators). We distinguish duplicates in rows and columns. The former mean events identical in case ID, activity name and timestamp. In other words, the same information is recorded twice for one trace. Therefore, such duplicates are needed to be filtered out before proceeding into process discovery.
- *Duplicate Columns*. Another problem is duplicate columns. These are data attributes that have different names, but similar information. For instance, Declaration and Declaration Number differ by 1, but actually report the same. In such cases, either of these columns can be used. However, there are also columns pretending to contain the same information, but turn out to be not on closer inspection, e.g. Permit ID and Permit id. These attributes are much more difficult to process as we do not know which one to choose for further analysis.
- *Missing Attributes*. For some data attributes, there are such values as MISSING or UNKNOWN. This is an issue because these values cannot be interpreted in a proper way. Hence, it seems impossible to identify opportunities for process optimization for such cases when not all information is available.
- *Missing Events*. One more problem relates to the situations when a particular event (or events) is supposed to be recorded in the log, but it is not. For example, Request Payment has to be followed by Payment Handled, but in some cases the latter activity is missing.

Of particular note is the fact that multiple datasets are provided for the analysis. In fact, it is a big enough challenge because, as mentioned above, some event logs are interrelated but cannot be combined at one due to the lack of a number of case IDs in all files at once or the inability to provide one-to-one correspondence between travel permit and declaration. What is more, some events are common to the traces from different datasets, i.e. there is a partial overlap. It results in ambiguous process flows and consequently fallacies or hasty or inaccurate conclusions. In contrast, dataset Request for Payment is not linked to any other event log since, as posed in the challenge description, it refers to expenses which should not be related to trips, and hence to travel declarations or permits.

Last but not least, an explanation on the data attributes was not provided for this year's challenge. So, we could only guess what each column (if it is not obvious like case ID or timestamp) in each file means.

Taking everything into account, the most intricate task is to carry out the compliance check for the reimbursement process cases.

## 3 Process Discovery

The given reimbursement process is divided into several flows, each one has its individual dataset. The desirable behaviour of the process provided by authors of the challenge corresponds to the most frequent event traces in each log. For this reason, for process flow discovery on a high level we assume that chains repeated most of all in the process represent its common flow.

In each case the most frequent trace is a complete reimbursement process without cycles and rejections and the second most typical one differs only in presence of BUDGET_OWNER what is also appropriate. Less frequent chains tend to include deviations such as rejection activities, loops, etc. so they can be safely filtered out in the control-flow discovery.

To track changes over time, we also divided each dataset by year of the first event in ID. The split is conditioned by the fact that 2017 was a pilot year and the data covers only two departments. To discover the process flow, we completed the following steps for each dataset:

1. Split dataset into two parts corresponding to 2017 and 2018;
2. Count occurrence of each trace;
3. Calculate basic statistics of the main trace over the dataset;
4. Filter out infrequent traces;
5. Apply Heuristic miner;
6. Perform conformance checking with the actual process.

For conformance checking we decided to use a custom metric based on the Damerau-Levenshtein distance. The idea is to measure the mean similarity between main traces that represent desirable behaviour and the remaining traces in the dataset.

In information theory and computer science, the Damerau-Levenshtein distance is considered as a string metric that measures the edit distance between two sequences. It is computed as the minimum number of operations (namely deletion, insertion, substitution, and transposition) needed to convert one string into the other. The choice of this metric is conditioned by the sequential nature of the event log where each case is described by a trace, a sequence of executed activities. It means that we can calculate the distance between two task sequences, one taken from the real event data and one produced by the simulation model.

The Damerau-Levenshtein distance between strings $a$ and $b$ with $i$-symbol and $j$-symbol prefixes (initial substrings) respectively $d_{a;b}(i;j)$ is defined recursively as follows:

$$\mathrm{d}_{a;b}(i;j) = \min \begin{cases} 0 & \text{if } i = j = 0 \\ d_{a;b}(i-1;j)+1 & \text{if } i > 0 \\ d_{a;b}(i;j-1)+1 & \text{if } j > 0 \\ d_{a;b}(i-1;j-1)+1_{(a_i \neq b_j)} & \text{if } i;j > 0 \\ d_{a;b}(i-2;j-2)+1 & \text{if } i;j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j \end{cases} \quad (1)$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 1 if $a_i \neq b_j$ and 0 otherwise.

### 3.1 Domestic Declarations

From comparing graphs of the process in 2017 and 2018 (see Fig. 2) it can be derived that the control-flow slightly changed in several ways:

- Declaration APPROVED by ADMINISTRATION precedes final approval by supervisor in 2018;
- Mean time of the main trace increased by 15 hours;
- Approval and rejection are made by ADMINISTRATION and BUDGET_OWNER in 2018 instead of PRE_APPROVER;
- In 2018 the payment is requested 35% faster after Declaration FINAL_APPROVED by SUPERVISOR.

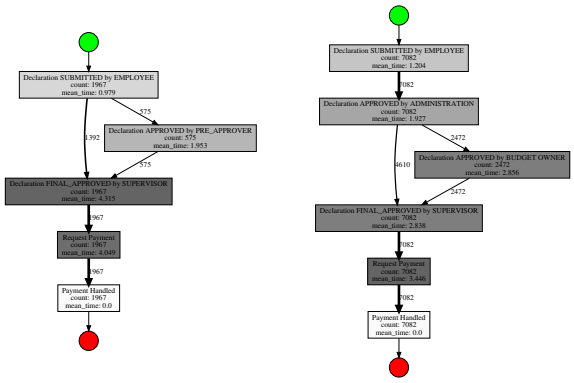


**Fig. 2.** Control-flow of Domestic Declarations dataset in 2017 (left) and 2018 (right).

Similarity score measured by the Damerau-Levenshtein distance is 0.93 for 2017 and 0.92 for 2018 meaning that diversity of event chains slightly increased, but the overall conformance with desirable behaviour is high. Mean time from Request Payment to Payment Handled decreased by 13 hours.

### 3.2 International Declarations

International declarations generally take a more complex path because of necessity of travel permits. Mean time of the declaration is considerably higher as well as the deviation from the target process. Between 2017 and 2018 there are also several discernible changes (see Fig. 3):

- Approval and rejection are made by ADMINISTRATION and BUDGET_OWNER in 2018 instead of PRE_APPROVER;
- Trip starts 23% later after approval of permit by SUPERVISOR in 2018;
- Payment is requested 46% faster after approval of declaration by SUPERVISOR in 2018.

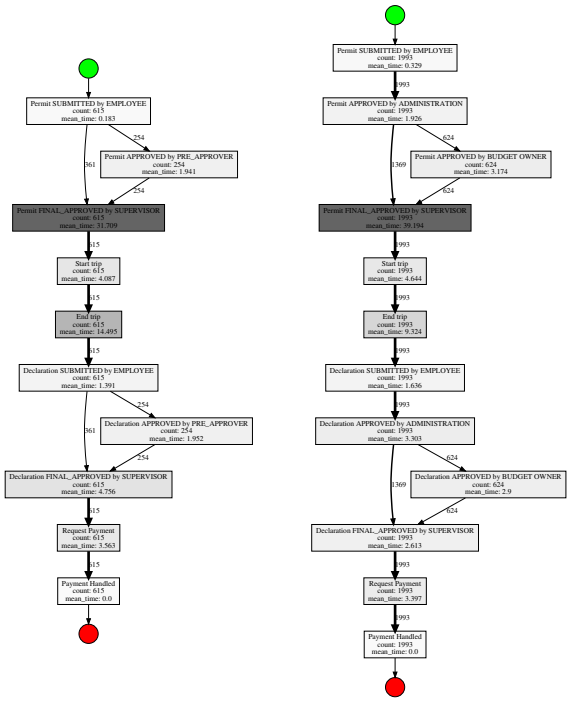


**Fig. 3.** Control-flow of International Declarations dataset in 2017 (left) and 2018 (right).

Similarity score is 0.74 for 2017 and 0.81 for 2018 which is a noticeable improvement. Process became more straightforward and less error-prone. Mean process length increased from 60 days in 2017 to 66 days in 2018 mainly due to the extension of time from Permit FINAL_APPROVED by SUPERVISOR to Start trip.

### 3.3 Prepaid Travel Cost

Declaration documents with preliminary payments unlike others do not contain records of Start trip and End trip but include information about separate requests for payment.

Like in other datasets, approval and rejection activities are made by ADMINISTRATION and BUDGET_OWNER in 2018 instead of PRE_APPROVER. Other things being equal, the overall difference between control-flows in 2017 and 2018 is minimal (see Fig. 4).

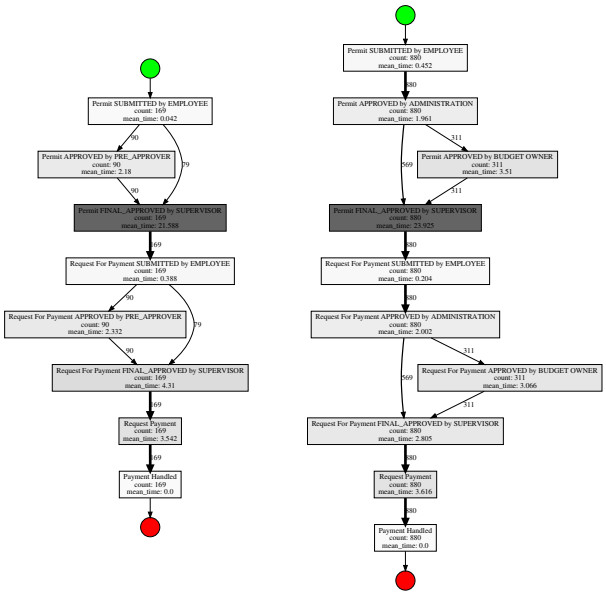


**Fig. 4.** Control-flow of Prepaid Travel Cost dataset in 2017 (left) and 2018 (right).

Similarity score is 0.8 for both 2017 and 2018, thus the process flow and quality did not changed or changed slightly. It is important to state that there are common traces where Request For Payment SUBMITTED by EMPLOYEE is not preceded by a travel permit in both years.

### 3.4 Travel Permits

Travel permits in general follow a more complex flow than declarations since they include permit and declaration approval and in some cases request for payment. Furthermore, there are many deviations from the reference path like incorrect order of trip and permit approval or declarations without existing permits.

The following differences could be derived from comparing control-flow of different years (see Fig. 5):

- Approval and rejection are made by ADMINISTRATION and BUDGET_OWNER in 2018 instead of PRE_APPROVER;
- Reminder is sent to employee once in 2.7 days in 2017 instead of 60 days in 2018;
- Declaration is submitted 41% faster after the end of a trip in 2018.

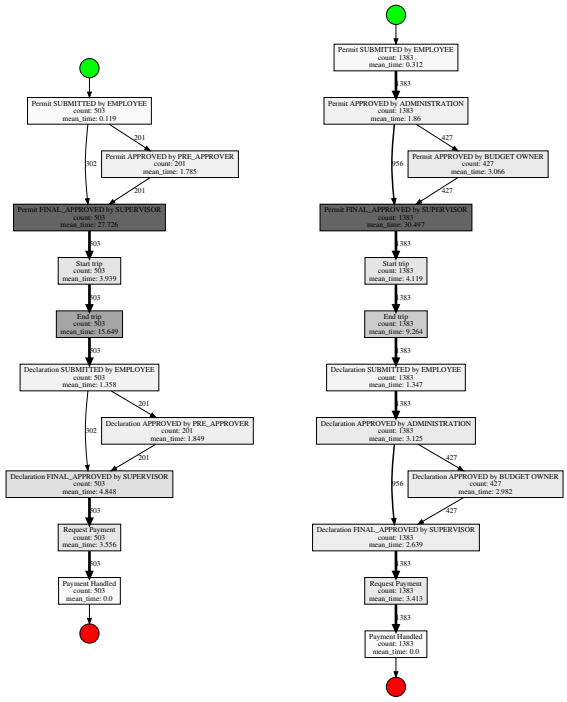


**Fig. 5.** Control-flow of Travel Permits dataset in 2017 (left) and 2018 (right).

Due to the high inherent variance and deviation from the main trace, similarity score of Travel Permits is low, namely 0.71 for 2017 and 0.69 for 2018. Mean time of the main trace decreased from 59 in 2017 to 55 days in 2018 despite the fact that its length increased by two events. This might be a result of effective performance of ADMINISTRATION that is responsible for declaration or permit approval preceding final approval by SUPERVISOR or BUDGET_OWNER.

### 3.5 Requests for Payment

Requests for payment follow a simple path and have similar differences between 2017 and 2018 (see Fig. 6):

- Approval and rejection are made by ADMINISTRATION and BUDGET_OWNER in 2018 instead of PRE_APPROVER;
- Approval by BUDGET_OWNER takes 69% more time than approval of PRE_APPROVER;
- Approval by ADMINISTRATION adds 2.8 days on average to the time of main trace.

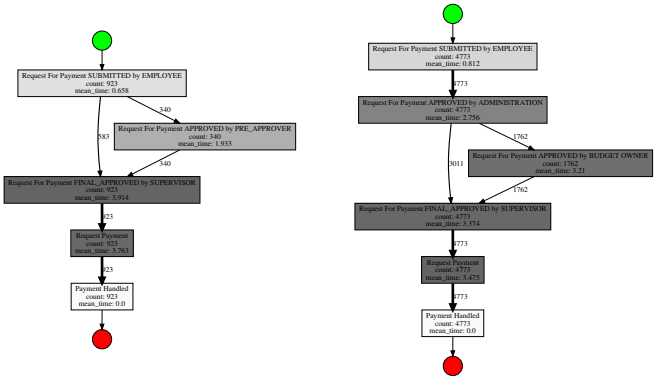


**Fig. 6.** Control-flow of Request For Payment dataset in 2017 (left) and 2018 (right).

Similarity to the common flow is quite high for travel permits – 0.91 for 2017 and 0.9 for 2018. The process is well-established and has few visible bottlenecks. However, mean time of the main trace grew from 8.5 days to 10 days mainly due to the increase of approval time.

## 4 Performance Analysis

First, we will go through the questions on the list offered by the process owner. Then, we will proceed to other new and interesting insights. It is worth mentioning that since 2017 was a pilot year for the reimbursement process and some changes in its flow happened, all questions were answered based on the data for 2018.

### 4.1 Question Answers

**What is the throughput of a travel declaration from submission (or closing) to paying?** First of all, we need to define the term 'throughput'. Actually, it can be interpreted in two ways:

1. the rate at which the number of units goes through the process per unit time, measured in units/per time;
2. the amount of time needed for a case to go through the process or from on activity to another, in other words duration.

Here and elsewhere, we used the second definition, that is the case execution time (either between start and end or between two selected events). So, Fig. 7 illustrates the mean and median throughput times for travel declarations from submission Declaration SUBMITTED by EMPLOYEE to paying Payment Handled. For domestic trips it is in average 11.6 days and for international trips 15.2 days. Median values are 8.0 and 10.8 days respectively. Since the median is less than the mean, the distribution of throughput time is skewed to the right in both cases meaning that there are outliers. Anyway, both domestic and international declarations are processed in a relatively short period of time.

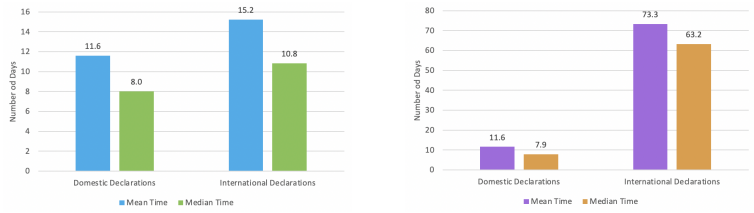


**Fig. 7.** Throughput of domestic and international declarations (left) and trips (right).

One way to reduce the total throughput time is to jump start the approval steps done by BUDGET OWNER and SUPERVISOR as now they take the most time.

**Is there are di erence in throughput between national and international trips?** National and international trips can be compared through the throughput time from submission to paying (calculated in the previous question) or the total case duration. As can be seen from Fig. 7, there are difference in the throughput equal to 3.6 for mean and 2.8 for median time. This can be explained by the fact that international declarations need more careful consideration. As for total duration, including permit submission for international travels, the average values are 11.6 days and 73.3 days for domestic and international trips respectively. Thus, the international trips take more than six times as much days as national.

**Are there di erences between clusters of declarations, for example between cost centers/departments/projects etc.?** To answer this question, we took five most common (containing the most declarations) projects from

*case:Permit ProjectNumber* column and departments from *case:Permit OrganizationalEntity* column of International Declarations dataset and calculated basic statistics to get some understanding about how they correspond to each other. As can be seen from Table 2, there is no apparent differences between clusters of declarations. This outcome was confirmed by an automated hierarchical clustering in Python.

**Table 2.** Clusters of international declarations.

| | Number of Declarations | Mean Case Duration (days) | Trip Length (days) | Mean Budget of Declarations | Mean Budget per Day |
|---|---|---|---|---|---|
| UNKNOWN | 1858 | 74.9 | 7.4 | 721.0 | 97.4 |
| project 426 | 282 | 72.0 | 3.0 | 934.6 | 311.5 |
| project 1623 | 27 | 54.7 | 8.3 | 244.6 | 29.5 |
| project 1115 | 26 | 87.9 | 4.9 | 1511.6 | 308.5 |
| project 1185 | 26 | 65.0 | 5.9 | 821.2 | 139.2 |
| organizational unit 65456 | 882 | 75.0 | 7.2 | 797.0 | 110.7 |
| organizational unit 65454 | 775 | 71.5 | 4.5 | 829.9 | 184.4 |
| organizational unit 65458 | 703 | 68.3 | 9.2 | 772.7 | 84.0 |
| organizational unit 65455 | 509 | 72.3 | 5.8 | 621.9 | 107.2 |
| organizational unit 65459 | 469 | 73.8 | 9.2 | 888.4 | 96.6 |

**What is the throughput in each of the process steps, i.e. the submission, judgement by various responsible roles and payment?** To answer this question, we considered all process steps apart from the ones recorded a few times in the log (like Permit FOR_APPROVAL by ADMINISTRATION and Declaration REJECTED by DIRECTOR) or supposed to be at the end of the event trace (e.g. Declaration SAVED by EMPLOYEE, Payment Handled). Also, we did not take travel duration into account. For each of the process steps not mentioned above, we measured the mean and median throughput time and plotted the results for all datasets on the heatmaps.

Fig. 8 shows the average throughput time in days of the reimbursement process stages. On the left side, there are all activities. The top three steps needed the longest amount of time to be executed are:

- Permit FINAL_APPROVED by DIRECTOR
- Permit FINAL_APPROVED by SUPERVISOR
- Send Reminder

To make other activities more distinguishable, we eliminated these three and draw one more heatmap (on the right). The longest events now are in Travel Permit and Requests for Payment datasets. As for the remaining process steps, they are blues meaning that they differ in the throughput slightly.

Fig. 9 depicts actually the same but for the median values. Again, all activities are on the left while only the part is shown on the right. Activities that were excluded are the longest and the shortest ones. So the heatmap on


**Fig. 8.** Mean throughput of the process steps.

the right-hand side makes the comparison of the process steps more apparent. While Request Payment has the highest throughput time in all event logs, Request For Payment REJECTED by BUDGET OWNER only in one dataset Prepaid Travel Cost.

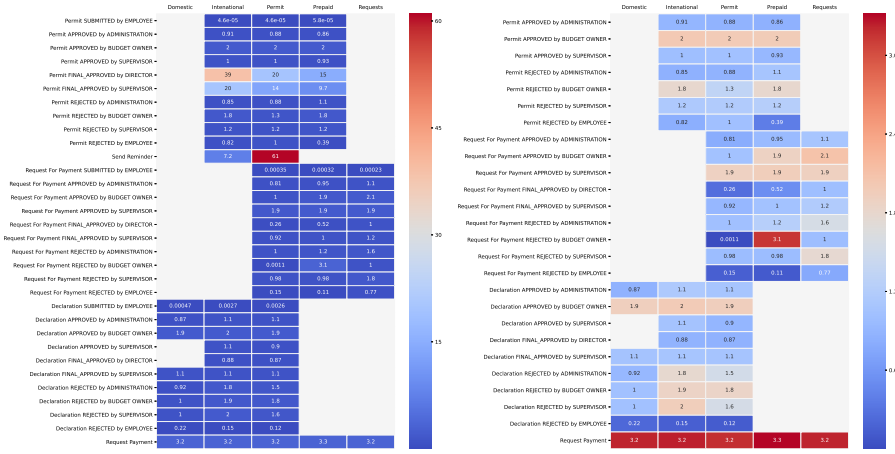


**Fig. 9.** Median throughput of the process steps.

**Where are the bottlenecks in the process of a travel declaration?** To find bottlenecks, we explored process imperfections on different levels, that is case level and event level. On case perspective we check if the given chain of

activities differs from the main flow, and on event level each stage or transition is compared with others by time, cycles and number of uses. Domestic and international declarations are examined separately.

For both domestic and international declarations, there exist traces that consist only of one activity, namely Declaration SAVED by EMPLOYEE. This might be a data quality issue, that is a result of incorrect log recording workflow or it could be a real case when an employee creates an online draft for a declaration and never uses it again. In both cases, this results in dirtier data and incorrect statistics calculation.

Another important bottleneck on case level is declaration re-submission. When declaration gets rejected by ADMINISTRATION, it takes 2-3 days on average to resubmit it again – this happens 19 times for domestic declarations and 23 times for international. Next, we found out that if a declaration gets rejected and then is resubmitted again, it can be rejected once more by the same responsible role. Such cycles happen 12 times in domestic and 11 in international declarations and take 2-5 days on average. It could also happen several times with one declaration. Multiple rejects may indicate inefficiencies in communication between employees and administration or incomprehensibility of submission rules.

Activity level bottlenecks include long and inadequate transitions such as one from Declaration FINAL_APPROVED by SUPERVISOR to Payment Handled (international declarations dataset). It happens 7 times and takes 82 days on average. This could be data error as well as process rule violation to proceed to payment without making a request first.

**Where are the bottlenecks in the process of a travel permit?** Travel permits share several bottlenecks with travel declarations, although they have a different flow. For instance, there exist multiple returns from approval of declarations to declaration submit after they already have been submitted beforehand. That happens 4 times in Travel Permit dataset and has a mean time of 5.4 days. Several event traces also include stages that go in seemingly wrong order. In following cases order of events has been broken with no apparent reason:

- 557 traces start with Start trip before submitting a permit;
- 10 traces contain only saving of permit, start and end of trip.

Returns to the same stage after already completing it occur in:

- Request Payment (214 times, 9 min. avg.);
- Payment Handled (32 times, 5.9 days avg.);
- Declaration SUBMITTED by EMPLOYEE (5 times, 15 min. avg.);
- Declaration FINAL_APPROVED by SUPERVISOR (35 times, 6.5 hr. avg.).

That certainly indicates multiple payments in one travel permit which might or might not be desirable, but multiple submits slow down the process and point to errors in process flow as well as multiple approves by supervisor.

**How many travel declarations get rejected in the various processing steps and how many are never approved?** 100 Domestic declarations are saved and never approved afterwards, 195 ones are rejected in different steps of process (see Fig. 10). For International declarations 192 are never approved, 45 of them end up saved by employee, 47 finish in the end of a trip and 100 are finally rejected by employee. Activities after which rejections happen are also presented in the graph.

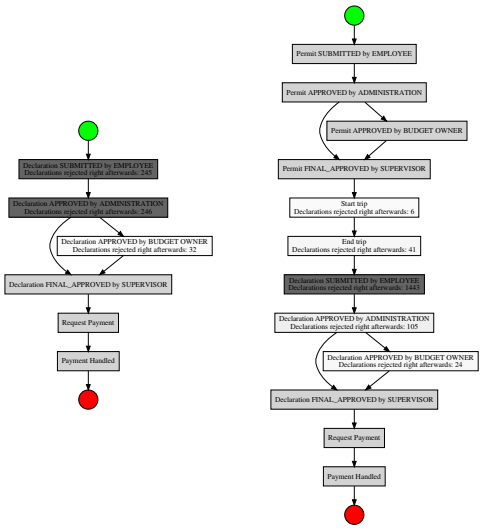


**Fig. 10.** Process flow of Domestic (left) and International (right) Declarations with number of rejected declarations on each step.

**How many travel declarations are booked on projects?** From International Declarations dataset, we computed the number of declarations that are booked on projects, i.e. in *case:Permit ProjectNumber* column the value is not UNKNOWN. The result is 3039.

**How many corrections have been made for declarations?** By corrections we mean rejects for the declarations that were later approved. In fact, we calculated how much times the travel declarations were rejected by various responsible roles, then re-submitted and finally approved. So, no corrections have been made for the domestic declarations and 1163 for the international declarations.

**Are there any double payments?** When activity Payment Handled is recorded multiple times for one case ID, we say that there is a double payment. Among

the datasets containing such payments is Travel Permits only where two types of event traces with double payments are possible:

1. One payment is made before the trip and another one is after. In cases like this, travel expenses often exceed the requested budget.
2. Several payments are made after the end of the trip one by one. This can be either a double payment or a system or input error.

The total number of cases with double payments equals 975.

**Are there declarations that were not preceded properly by an approved travel permit? Or are there even declarations for which no permit exists?** While investigating the International Declarations dataset, we found out that there are 261 cases when declaration submission was not preceded by permit approval. Moreover, in these event traces travel permit was not even submitted by employee. However, for each declaration there is a value in *case:Pemit ID* column, that is to say travel permit number. The latter can also be detected in the Travel Permits dataset where permit submission and approval steps precede declaration. This insight is actually about data quality issues. So, if we consider *case:Pemit ID* column, we can conclude that there are no declarations for which no permit exists. But if we base on the recorded event traces, the answer is opposite and the number of such declarations is 261. This is true for the International Declarations dataset.

**How many travel declarations are submitted by the traveler and how many by a mandated person?** All travel declarations are submitted by the employee who took the trip. Thus, there are no declarations that were submitted by anyone other than the traveler.

**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?** The following International declarations were found according to the provided description: 147028, 147039, 147053, 147425, 147452, 147487, 147514, 1838, 19086, 27582, 40430, 42977, 62541, 63081, 9995 (15 declarations in total). In their traces trip has ended and at some point a reminder was sent after what a declaration was submitted, rejected and then resubmitted again.

**Is this different between departments?** Assuming that by department authors mean case:Permit OrganizationalEntity, we calculated that in declarations from previous question:

- department 65454 was recorded 7 times;
- department 65457 was recorded 2 times;
- departments 65480, 65459, 65464, 65458, 65456, 65460 were recorded 1 time.

**How many travel declarations are not approved by budget holders in time (7 days) and are then automatically rerouted to supervisors?** The following declarations match given conditions: 101126, 101140, 101188, 101194, 101465, 101522, 102063, 104962,104985, 105114, 109603, 110965, 111083, 117441, 119678, 121359, 121820, 121826, 133587, 99462 (20 declarations in total).

**Next to travel declarations, there are also requests for payments. These are speci c for non-TU/e employees. Are there any TU/e employees that submitted a request for payment instead of a travel declaration?** Since the data is anonymized and there is no information about TU/e internal IDs, we cannot answer this question.

### 4.2 New and Interesting Insights

**Project cost and duration.** When comparing projects, we discovered the following dependency between budget per day and trip duration. For each project provided in *case:Permit ProjectNumber* column of International Declarations dataset we calculated the number of declarations it relates to, mean trip duration, and mean budget (*case:Amount*).

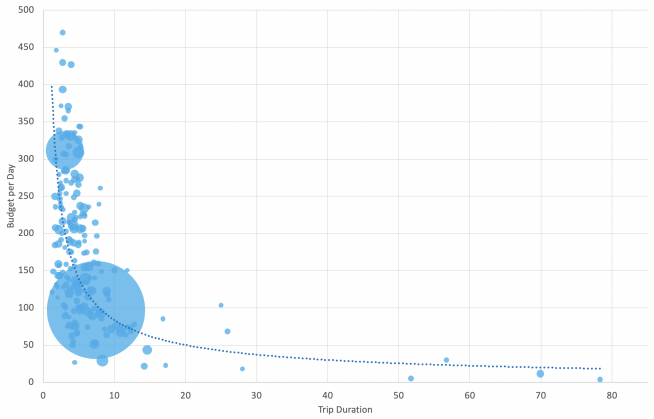


**Fig. 11.** Budget per day for project in dependence of trip duration. Size of bubble indicates number of declarations in project.

Fig. 11 shows that the shorter the trip is, the larger amount is spent per day by the employee. This might be an insight into budget policy of TU/e or a consequence of existence of short travels with significant expenses for flights.

**Automatic insights search.** To summarize insights found in the reimbursement process, we decided to apply an automatic insight search algorithm created

by our team. It takes into account several metrics such as mean time, number of uses, number of unique users and number of cycles. These metrics are calculated for each activity and transition in event log. If an object is beyond either top or bottom 15% quantiles, it is considered as insight by this metric. The results for each object are combined with corresponding weights and an overall optimality of transition (activity) is calculated using the following formula:

$$\forall e \in E : Opt(e) = \sum_{i \in I(e)} w_i; \tag{2}$$

where $E$ is a set of transitions (activities) in event log, $e$ is the activity itself, $w_i$ is a weight of corresponding metric, and $I(e)$ is a set of insights calculated by metrics on given transition (activity).

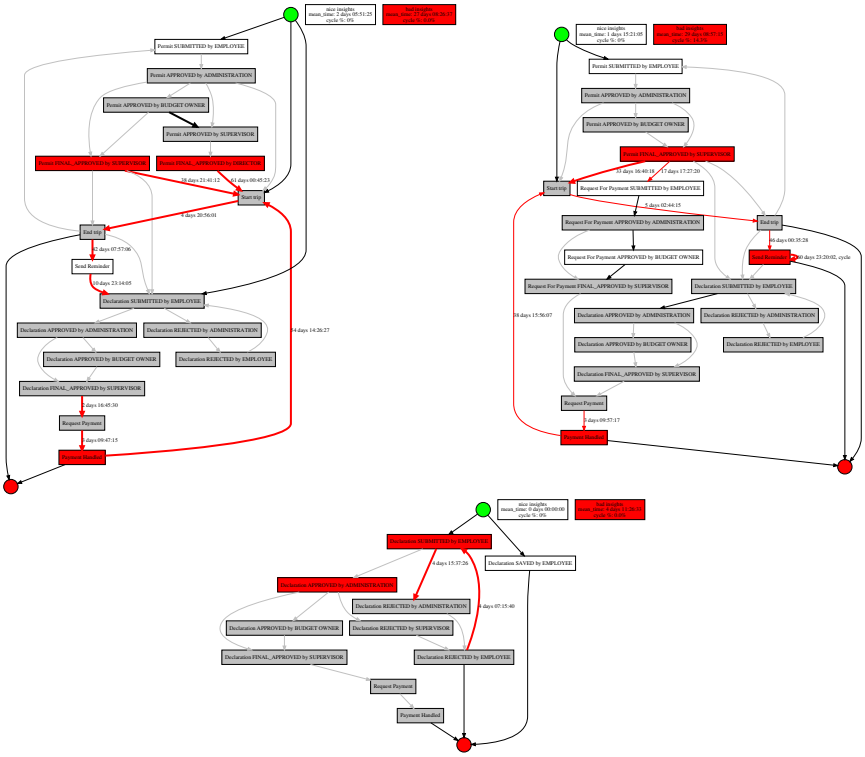

**Fig. 12.** Insights on heuristic miner graph on International (top left), Permit (top right), and Domestic (bottom) datasets.

The results of Autoinsights algorithm are applied to heuristic miner of each process (see Fig. 12). Colors on graph represent how effective (ineffective) a node or edge is. Black means the object is closer to optimal and red indicates

possibility of further optimization. Here we also dropped infrequent event traces to get a transparent image of the process flow.

For each type of travel highlighted insights are divided into two general categories. First category consists of insights justified by the process itself like the long transition between Start trip and End trip or self-cycle in Send reminder. This is up to the process owner to decide their importance. The second category is insights that point out clear inefficiencies in the process. Some of them are connected with the questions provided by authors of the challenge, e.g. international declarations that start before submission of travel permit. Another example can be seen on Travel Permit graph (top right in Fig. 12) when after the end of a trip and sending of reminder, declaration can be submitted, rejected and then resubmitted again.

In fact, insights depicted in Fig. 12 relate to the ones derived above during answering questions. In other words, by implementing our Autoinsights algorithm we provided all-in-one visualization.

## 5 Suggestions for Improvement

After pointing out insights on different levels of the process, we decided to share our opinions on how the process may be improved from the process mining standpoint.

### 5.1 Avoid Multiple Resubmissions of Declarations

As we already stated in the previous section, there exist cycles in stages:

- { Declaration SUBMITTED by EMPLOYEE and
- { Declaration REJECTED by ADMINISTRATION.

Such loops happen up to four times in a declaration path meaning that a declaration is submitted and rejected afterwards four times. One way to address this issue is to establish better communication between employee and administration on the questions why a declaration is rejected and what exact corrections should be made to get an approval. Other solution is to redirect rejected declarations to another mandated person that specializes in re-submissions. This could possibly save time of administration and shorten mean time of the client path.

### 5.2 Prevent Trips from Happening Without Existing Travel Permit

International declarations tend to have deviation in order of events such as submitting a declaration or taking a trip before getting an approved travel permit. First case was examined closely in Performance Analysis section and the second one is the subject of further research. We suggest process managers to take measures to prevent unauthorized trips, for example, one could add Check for Travel Permit stage at the beginning of the trip to make sure that the travel permit was approved.

### 5.3 Reduce Time Interval Between Sending a Reminder and End of a Trip

A rather important aspect of the process is sending of reminder to an employee to submit a travel declaration. Timely sending of a reminder can greatly speed up process performance by reducing waiting time for other employees and enhancing throughput. Although an employee has a 2-month period from the end of a trip for declaration submission, if a reminder is sent too late, an employee may not get expenses reimbursed, hence, such case will be considered incomplete. Therefore, we stick to the opinion that reduction of the idle time before sending a reminder, for instance, to 5 days on average should positively affect both the process flow and staff experience.

### 5.4 Submission of Declaration Strictly After Trip

At the moment, the process structure implies possibility for an employee to submit a declaration in advance for a refund, i.e. receive payment before a travel starts. In fact, we can split all chains of activities into two types depending on whether a declaration is submitted before or after the trip. In the first case process tends to have more bottlenecks and inefficiencies while representing undesirable behaviour in general. If declaration is submitted before the trip, the case duration is 20 days longer on average. What is more, in such cases we can detect additional declaration submissions as the actual expenses exceeded the prepaid travel costs. Therefore, in our opinion, it would be reasonable to restrict employees from submitting declaration before the trip is taken in order to improve process throughput and conformance.

## 6 Conclusion

Since expense reimbursement is directly linked to cost (and hence budget) allocation in the organization, process owner is especially interested in effective monitoring of its execution. Therefore, this report is aimed at the comprehensive analysis of the reimbursement process at TU/e by applying various process mining techniques. The main goal is to ensure that work-related transactions are administered and settled with efficiency, accuracy, and accountability to the satisfaction of both the University and the staff

First of all, we explored the data and process provided. After inspecting some descriptive statistics, relation between event logs and data quality issues, we preprocessed the data in some way and then discovered process models by mapping reimbursement processes as they actually occur. By comparing control-flows of the process in 2017 and 2018 and checking conformance with the Damerau-Levenshtein distance, we identified critical values such as bottlenecks and deviations.

Next, we moved to the performance analysis. In this step, we decided to neglect the cases from 2017 as it was a pilot year. By answering questions provided

in the challenge description and implementing our special algorithm, we derived a number of insights into the process. These findings contribute in several ways to our understanding of the expense reimbursement process and provide a basis for its improvement and optimization. Based on identified optimization potentials, we proposed several ways to enhance the process in the end of this paper.

To sum up, this study may be of assistance to the process owner, that is TU/e, and other organizations practicing employee expense reimbursements for business-related costs. Further research might explore the root causes of the main process problems. Also, experiments with the use of simulation modeling and what-if analysis could shed more light on the question whether or not suggested process improvement alternatives are able to advance process performance.

## References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes, vol. 2. Springer (2011)
2. van der Aalst, W.M.P.: Process mining: Data science in action, Second Edition. Springer, Heidelberg (2016)
3. van Dongen, B.: BPI Challenge 2020, https://doi.org/10.4121/uuid:52fb97d4-4588-43c9-9d04-3604d4613b51.