Analysis of Travel Reimbursement Processes Based on Combined Eventlog Using Process Mining: a Contribution to BPI Challenge 2020

Aleksandra Piasecka, Oskar Leligdowicz and Paul Giessler

Private Contribution

piasecka.ola@gmail.com, oleligdowicz@gmail.com, paul-giessler@web.de

Abstract. Travel reimbursement processes are almost ubiquitous. Many organizations put them in place in order to manage costs and ensure compliance with company travel guidelines and allow employees a prompt reimbursement of expenses. Being highly digitized and perceived as very impactful in business contexts, these processes are ripe for being analyzed by means of process mining techniques. In this paper, we provide an in-depth analysis of the travel reimbursement process at TU/e (Eindhoven University of Technology) using data shared within the Business Process Intelligence Challenge 2020, a competition co-located with the International Conference on Process Mining. In order to ensure a thorough analysis of the process, we applied process data modeling techniques. This allowed us to streamline five separate eventlog inputs into one comprehensive data model that we describe in the second chapter of this paper. In the following sections, a target process model has been devised and an exhaustive analysis of the process has been carried out, identifying inefficiencies and events of non-compliance and their varying impact on the actual process. Said analysis allows us to conclude this paper with a set of recommendations for future applications of process mining in the context of travel reimbursement processes.

Keywords: BPI Challenge, Process Mining, Travel Reimbursement Process, Process Discovery.

1 Introduction

The anonymized data set originates this year from the employee travel reimbursement process at TU/e (Eindhoven University of Technology). Employee reimbursement processes can be found in all kinds of organizations, from non-profit institutions to publicly listed global enterprises. All of these organizations need to put in place processes that ensure that their employees receive timely reimbursements. As these processes typically affect a large group of employees they tend to be crucial to the satisfaction of employees. Moreover, these processes have important implications for cash flow management and cost controlling.

Process Mining offers the possibility to analyze how these processes compromise between efficiency and compliance in real life. The expected process has been described by the BPI team in the online challenge statement. Assessing the overall performance and efficiency of the process is hard because general best practices or target values for the most relevant KPIs are rarely available. However, comparing these KPIs for various dimensions within the given data set provided some indications that the process can be improved within certain parts of the organization.

This year's BPI data set has been provided in five different source files, all representing a different process type. Our first analysis of the data sources indicated that merging the data is crucial to get a complete overview of the process. Certain document IDs have been referenced in two or more sources and answering some of the questions provided for in the challenge requires a direct juxtaposition of data from two different sources. Therefore, we dedicated section two entirely to describing the modeling s of the various data sources in order to obtain a single end-to-end data model. In section 3 we provide a comprehensive overview of the travel reimbursement process and a high-level comparison between the to-be and as-is processes. Section 4 covers our detailed real-life oriented answers to the questions provided on the challenge website, together with several forward-thinking performance and compliance analyses. We conclude with a set of recommendations for further analyses.

2 Data Preparation

The data for this year's challenge consists of five .xes files, each containing an eventlog and a case table:

- RequestForPayment.xes (RFP)
- DomesticDeclarations.xes (DD)
- PrepaidTravelCost.xes (PTC)
- InternationalDeclarations.xes (ID)
- PermitLog.xes (TP)

2.1 Initial Analysis

We started our investigation by visualizing each of the eventlog files with a Celonis analysis to find out the relevance of each data file for answering provided questions. Additionally, our goal was to identify overlaps and potential dependencies between the eventlog files. To create a combined overview of activity occurrences per file we created a python script (see appendix 0.1) writing each of the files into a table on a MSSQL Database server. For this task libraries *xml*, *pandas*, *gzip*, *time* and *os* were used.

We developed a SQL script compiling the aggregated event information into a single table (see appendix 1.0 and 1.1 for script and detailed results). This provided us a detailed overview of the 57 activities present in the eventlogs and the following findings:

1. There are 8 activities that only exist in one of the files:

- *Declaration FOR_APPROVAL by ADMINISTRATION*
- *Declaration FOR_APPROVAL by PRE_APPROVER*
- *Declaration FOR_APPROVAL by SUPERVISOR*
- *Permit FOR_APPROVAL by ADMINISTRATION*
- *Permit FOR_APPROVAL by SUPERVISOR*
- *Request For Payment FINAL_APPROVED by BUDGET OWNER*
- *Request For Payment FOR_APPROVAL by ADMINISTRATION*
- *Request For Payment FOR_APPROVAL by SUPERVISOR*

2. Most of the activities exist in two or more files
3. Two activities \ exist in all five files (*Payment Handled*, *Request Payment)*

Additionally, we got a comprehensive overview of the overall data volume in the files. The largest eventlog file (Travel Permits) comprises 86,581 events (rows), while the smallest file has 18,246 events (rows). All five eventlogs combined contain 270,211 events/rows without taking potential overlaps of events between files into account.

To further improve our understanding of the process and validate it with the process description in the BPI challenge statement, we performed a grouping exercise of the activities. The following pattern was noticed in the activity naming:

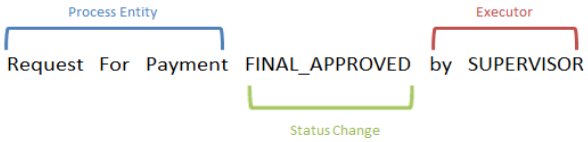


Fig. 1. Activity naming pattern explained by means of example "*Request For Payment FINAL_APPROVED by SUPERVISOR*"

Only 5 activities do not follow this naming pattern - *Payment Handled*, *Request Payment, End trip, Start trip, Send Reminder*. The process description in the challenge statement distinguishes between multiple process entities. Therefore, we grouped the activities into the following categories aligning to these process entities (please see appendix 1.2 for details) - Permit, Trip, Declaration, Request for Payment, Payment.

	TP	DD	ID	RFP	PTC
Permit	X		X		X
Trip	X		X		
Declaration	X	X	X		
Request for Payment	X			X	X
Payment	X	X	X	X	X

Fig. 2. Relation between process entities and eventlog files

It has to be noted that these process entities are partly overlapping with the naming of the .xes files that have been provided. However, the presence of process entities in

the .xes files cannot be deducted through the filename. The detailed relation between process entities and eventlog files is depicted in **Fig. 2**.
The steps mentioned earlier led us to a solid understanding of the activities in the process. In parallel, we started an investigation of the additional table fields to identify potential relations between the provided files. As we assumed the data originated from a relational database we started to analyze the columns structure and samples of the content concerning typical primary key and foreign key patterns.

In each of the eventlog files multiple fields could be identified where the naming indicated a potential usage as primary key to uniquely identify the corresponding process entity. As an example, the file *InternationalDeclarations.xes* contains the columns *ID* and *DeclarationNumber* whose names suggest that they could serve as a unique identifier for the whole set of international travel declarations. However, analyzing the content of the field concept_name and validating the xes-formatting of the source files lead us to the decision to use the field *concept_name* as a primary key for all files.

Apart from the primary key fields, the data set contains multiple columns whose names indicate a potential usage as foreign keys to join data between the provided files. The foreign key candidates found in an initial analysis are listed below:

Table 1. Foreign key candidates in the provided data

Table Name	Column Name	Comment
InternationalDeclarations_caselog	Permit ID	
InternationalDeclarations_caselog	Permit travel permit number	
InternationalDeclarations_caselog	travel permit number	
PrepaidTravelCosts_caselog	Permit id	
PrepaidTravelCosts_caselog	Permit travel permit number	
PrepaidTravelCosts_caselog	Rfp_id	
PrepaidTravelCosts_caselog	RfpNumber	
TravelPermits_caselog	dec_id_x	Suffix 0 to 16
TravelPermits_caselog	DeclarationNumber_x	Suffix 0 to 16
TravelPermits_caselog	Rfp_id_x	Suffix 0 to 14
TravelPermits_caselog	RfpNumber_x	Suffix 0 to 14

All the analyses mentioned above let us to the conclusion that the data needs to be combined into single eventlog and single case table. The main reasons for this decision are following:

- Due to the usage of foreign keys in various tables, case information might be incomplete without combining data from more than one eventlog (e.g. BudgetNumber is only available in the International Declaration and travel permit data set while payment amounts are only visible in the travel permit files).
- Parts of the given questions obviously required information from multiple eventlog files, such as question no 2 which comprises a comparison between domestic and international declarations.

– Most activities occur in more than one of the five eventlog files, the payment related activities even exist in all five. It is therefore fair to assume that the full impact of single activities on the overall process is only visible with a combined data model.

In contrast to these points in favor of a combined data model, we also considered points that would oppose it. Firstly, foreign key candidates are ambiguous which means that a detailed validation is required to ensure the combined data model matches the process flow description in the challenge statement. The validation furthermore needs to include a check for duplicates and a check for completeness. For both of these checks we already had procedures in place, as they are part of the standard routine we applied to Process Mining data sets in the past.

Furthermore, another potential disadvantage of a combined data model might be the complexity and the increased data volume. At this point, we already knew that a combined data model would comprise 57 activities and at most 270,211 events (rows in the eventlog). Both numbers are known to be in the range of complexity where free Process Mining tools on the market can run analyses without any performance issues.

Due to the relatively simple mitigation approaches for both of the reasons against a combined data model, we concluded that the benefits clearly outweigh the risks and therefore continued with this approach.

2.2 Final data model

To unify the data model we made several assumptions based on data itself and additional information provided on BPI 2020 Challenge website. The corresponding code can be found in appendix 3.0.

Order of documents. First assumption was made in respect to the order of documents. Because declarations are most common document across all process instances we selected them as first priority cases (first anchor point for our eventlog). It means that every international and domestic declaration is the case in our analysis. Using this approach there are 10,500 cases for domestic declarations and 6,449 cases for international ones in data model.

The second most common in the process is the travel permit. Due to this fact, we selected them as second priority cases. It means that if there is a travel permit document connected to declaration then it will be part of the case already created for declaration. If there is a travel permit document without such connection then it will be treated as a new case. As a result, only travel permits without declaration generate new cases. In the data model there are 1,457 unique cases added for travel permits and 5,608 were added as part of existing cases.

Third priority cases were generated based on prepaid travel cost documents. Using the same approach, if a prepaid travel cost document is connected to already added cases it becomes part of this case. If it is not connected, a new case is generated. We found out that all documents are connected, thus no new cases were created. All 2,099 cases were added as part of existing cases.

Last document type added was Request for Payments, which instances turned out not to be connected to any other documents type at all. Therefore, every request for a payment document generates a new case - in total 6,886 cases were added.

Connections. Second assumption is that we can connect different documents using provided data. Candidates for primary and foreign keys to connect those documents were described in detail in the previous subchapter. All candidates were checked to find out the best possible connections that result in the highest number of connected documents on both sides.

Domestic declarations are not connected with any other documents so they can be skipped from this step. Subsequently, we tried to connect other documents to international declarations since those are our first priority cases. We found out several candidates to connect travel permits. Best results were achieved using the field "Permit ID" from international declarations and "concept name" from travel permits. Using this connection, we manage to find travel permits for 6,001 international declarations - in total 5,608 unique travel permits were connected. To give one example, we created a case combining events from declaration 47484 and travel permit 47480.

We also managed to connect prepaid travel costs to international declaration via the "Permit ID" field, which is in both data sets. This way we connected all 2,099 prepaid travel costs documents to 1,962 unique international declarations. One of connected documents is declaration 1002 with travel permit 992 and three different prepaid travel costs documents named as request for payment 996, 998 and 1000.

For Requests for Payments we could not find any connection. We tried different fields from Travel Permits data set: "Rfp id x" (with suffix 0 to 16), "Rfp Number" as well as "Rfp Number" from Prepaid Travel Costs. The only common value we found out is value "UNKNOWN". Moreover, we also found out that "Rfp number" on Request for Payment has fewer characters than candidates for foreign keys in other tables. Due to this fact, we also tried to take parts of numbers from fields selected as candidates for foreign keys and compare them to "Rfp Number" from Request for Payments. As a result, we did not get any meaningful connections. To double check results, we also compared values from field "Requested Amount" which is in both travel permits and requests for payments. This test confirmed our suspicions that data sets are not related.

Duplicated events. During data pre-analysis, we also discovered the same events exist in more than one data set. We learnt that every duplicated event has the same value in field "id". Due to those facts, the next assumption we formulated was that we want to avoid event duplication since the same event in a different data set gives us exactly the same information. Moreover leaving such events would be misleading. It would suggest that some actions were executed more than once for a given case.

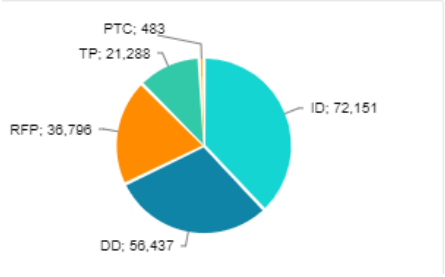
During data model preparation at every step when events were added, we were checking if a given ID already exists in eventlog – if so, the event was skipped. Eventually, we validated if there are any cases without events. It would mean that all events that were part of another case and this particular case could be deleted.

Different time zones and user types. We also noticed two additional smaller irregularities, which could affect the eventlog. Firstly, not all events were in same time zone. Leaving events in those time zones would cause incorrect display of the process, potentially leading to erroneous insights, e.g. by affecting the throughput time. To avoid those problems we recalculated timestamp for every event to one common time zone – CET (Central European Time).

Secondly, most events contained the user role conducting a specific action. Thus, numbers of distinct event description in source files were substantially higher. To avoid unnecessarily high complexity of the eventlog caused by event names we relocated user roles to separate column and kept only the activity name. That improves the transparency of the process and fosters process discovery, as well as in-depth analysis, since it is easier to look into specific type of events as a group.

3 High Level Process Analysis

The process model utilized for that analysis was visualized in the in the process mining software Celonis Snap. It comprises 25,292 cases with 187,155 events in total. The eventlog consists of 23 distinct activities executed by 8 different user roles. Distribution of events and cases is shown below.



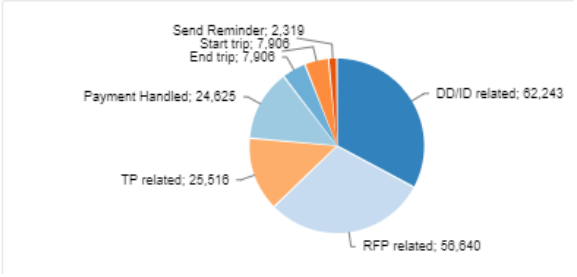


Fig. 3 and 4. Distribution of events by source file (left) and event type (right).

Case Type	Case count
DomesticDeclaration	10,500
Request For Payment	6,886
InternationalDeclaration	6,449
Travel Permits	1,457

Fig. 5. Distribution of cases by source file.

These 4 case types reveal significant differences in the process flow and the process entities involved in the process in general. The most common and most standardized case type is the domestic travel declaration with only 78 different variants. The five most common variants cover 92% of these cases. These variants match the process flow description from the challenge statement. Domestic declarations are submitted and paid. Rejections lead to either an immediate process end or a resubmission as it can be seen in appendix 2.0.

The second most common case type is the Request for Payment. Firstly, it is important not to confuse the request for payment including its submission and approval steps with the final payment request and payment handling which is also occurring for declarations and permits. Secondly, in the BPI challenge description it was explained that these cases are not travel related but cover expenses like representation costs, hardware purchased for work, etc. This can be confirmed in the data as no RfP case is connected to any travel related process entity like permits, declarations or the actual trip (see appendix 2.0).

International declarations are the third most common case type and the by far most complex one. From the ICPM website we knew that international declarations always need a travel permit and in some cases an additional request for payment. These two points can mostly be confirmed with the data. In the five most common variants a travel permit is always submitted and a significant share of the cases flows through request for payment activities. However, the dataset contains 438 international travel declarations that do not have a travel permit approved in advance (see Outlier analysis and in-depth analysis). All in all, international declarations follow 1,144 different variants. The five most common variants cover 39% of the international declarations (see appendix 2.1).

The international declarations furthermore contain so-called pre-paid travel costs. In these cases a payment is made to the employee before the trip starts and a second payment follows after the declaration is approved. They are represented for example in the 5th most common variant of the international travel declarations (see appendix 2.2).

The least common case type is the travel permit. In the five most common variants, these permits are regularly created before an international travel but not converted into an international travel declaration. For most of the cases the process ends with one or more reminders being sent out to the employee that the permit still needs to be transformed into a declaration in order to receive a reimbursement. In total 202 variants can be found for this case type and the top five variants cover 55% of the 1,457 cases (see appendix 2.3).

3.1 Target Process Model

The online challenge statement offered a detailed description of the target process flow in the reimbursement process. Based on the online description and the knowledge we gained through the analysis of the most common variants we were able to derive a process model presented in the figure below.

71% of the cases in the overall data set are conforming with this process model. In the year 2017 the conformance rate is at 65% while it rises to 73% on average when the pilot phase is finished in 2018. Requests for payment are deliberately excluded from this model as they do not seem to be actually travel related and are not explicitly mentioned in the online process flow description.

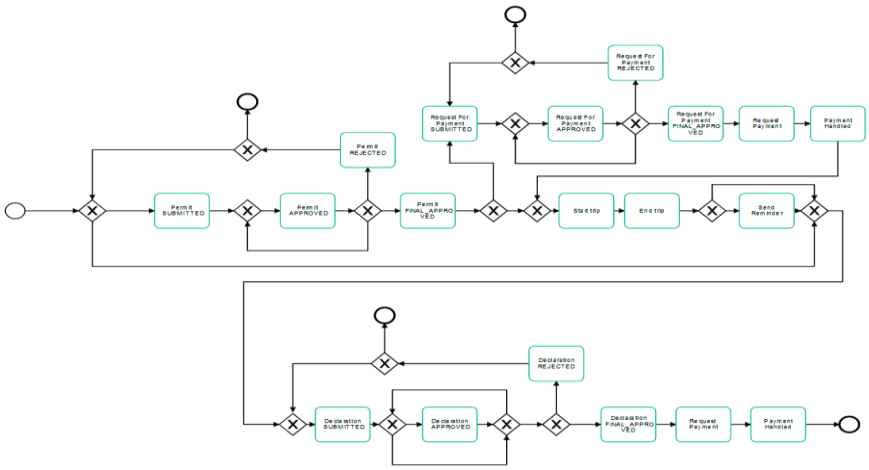


Fig. 6. To-be process flow for travel reimbursement processes.

3.2 Outlier Analysis

In this section, we provide an analysis of properties in the data set that are unexpected based on the data description provided in the online challenge statement. The first part of the section is about outliers that are assumed to be due to data quality issues in the original system while the second part covers outliers that are based on the process flow description and may be due to incorrect usage of the system rather than incomplete data.

Data Quality. The dataset originates from the years 2017 and 2018. It has to be noted that the system has been in a pilot phase in the year 2017 and the actual target process has been implemented since the beginning of 2018. However, the data set contains timestamps from the years 2016 to 2021. Most of the timestamps outside the years 2017 and 2018 originate from the events Start trip and End trip, which is presumably, due to the fact that these timestamps are directly based on user input for the expected travel dates. It appears that users are allowed to enter declarations for trips relatively far in the future or past. There are other events as well frequently occurring with timestamps after 2018 (e.g. Payment Handled, Send Reminder) but all of these events are connected to declarations or permits that have been created in 2018 which is a valid reason to include them within the data set in order to not show more incomplete cases than necessary.

A further data quality issue we noticed is that many of the case properties that have been mentioned in the challenge statement do not contain sufficient data for a conclusive analysis. As an example, we can mention the field PROJECT which is empty for 16,955 cases and contains the value "UNKNOWN" for 496 cases even after combining entries from all of the five source files.

Target Process Conformance. The process flow description provided online and the findings in the data align quite well. Partly, this is for sure because the process flow description is relatively vague at some points. However, the challenge description explicitly stated that international declarations need to be approved by a supervisor and that a travel permit needs to be approved before the employee makes any arrangements. The data set shows that there are 438 travel declarations labeled as "international" that are submitted without a travel permit being submitted or approved before. A dedicated question regarding these cases with a detailed answer can be found in the next section.

4 In-depth process analysis

In the following paragraphs, our detailed analysis and answers to provided questions and an additional financial analysis can be found.

4.1 General questions

What is the throughput of a travel declaration from submission (or closing) to paying? For a comprehensive answer to this question, we decided to consider several aspects. Firstly, we focused on travel declarations containing the steps of declaration submission and payment handling – which means ca. 22k (70%) of all analyzed reimbursement cases. Looking at the average, the throughput time from submission of a declaration to handling of a payment is 12.94 days. However, the median for the same is almost 3 days shorter and equals to 9 days. This difference indicates there are so-called long-runners – outliers with a very long throughout time. This can also be noticed on the histogram below.

There are 163 cases where the throughput time between declaration submission and payment exceeds 100 days. Although only 25% regarded declarations come from 2017, the share of long runners in this group is higher – around 35% (58 out of 163). It could be explained by the pilot phase of the system, which was taking place in 2017.

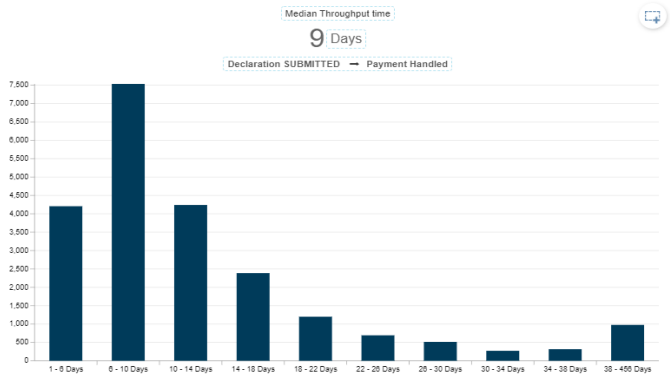


Fig. 7. Median throughput time from submission of a declaration to handling of a payment

Is there a difference in throughput between national and international trips? Yes, there is a significant difference in throughput time between these two types of trips**.** Comparing these two trips based on the declaration type, it was calculated that processes for domestic declarations are on average more than 7 times faster than the international ones (12.47 days vs 91.42 days). An even bigger difference is noticeable when comparing median values: 8 days vs 71 days. It is related to the more complicated process of international declarations – while domestic declarations require less than 6 activities (e.g. no travel permit is required) in the process on average, for international ones it is almost 18.

Case Type	Total throughput time in days	Median total throughput time in days	Case count	Activities count	AVG activities per case
DomesticDeclaration	12.47	8.00	10,500	56,437	5.79
InternationalDeclaration	91.42	71.00	6,449	82,974	17.59

Fig. 8. Throughput time comparison between international and domestic declarations.

Due to these significant differences between the handling of international and domestic declarations the comparison of the total throughput time will not give us a fair picture. Therefore, it is more reasonable to base the comparison on a more precise KPI like the throughput time between the submission and the final approval of the declaration as these steps should occur in the same order and similar context for both categories. However, even for this KPI the domestic declarations are processed significantly faster.

Case Type	Case count	Mean throughput time decl. submission to approval	Median throughput time decl. submission to approval
DomesticDeclaration	10,500	4.91 days	2.00 days
InternationalDeclaration	6,449	7.89 days	4.00 days

Fig. 9. Throughput time declaration submission to approval comparison between international and domestic declarations.

It was observable that international declarations are more often affected by rejections than domestic declarations (rejection rate 12% vs. 24%). Rejections have a high impact on the throughput time between declaration submission and final approval. To analyze if the higher throughput times for international declarations are solely because of the

higher rejection rate, we performed the same analysis as above again only for declarations that have been rejected in the process. The result is that in these cases the difference between domestic and international declarations becomes less significant. On average international declarations are processed even slightly faster than domestic declarations (15.10 days vs. 15.83) while their median is still higher (10 days vs. 7 days).

Case Type	Case count	Mean throughput time decl. submission to approval	Median throughput time decl. submission to approval
DomesticDeclaration	1,301	15.83 days	7.00 days
InternationalDeclaration	1,576	15.10 days	10.00 days

Fig. 10. Throughput time declaration submission to approval comparison between international and domestic declarations for rejected declarations only.

As a conclusion, it is fair to assume that the worsened performance of international declarations is partly due to the higher rejection rate of these cases. The data set only gives high-level indications why international declarations are rejected more often but in general, it is reasonable to conject that international declarations require more complex documentation and therefore provide more opportunities for the employees to make mistakes that lead to rejections by approvers.

Another factor suspected to affect the rejection rate is the amount of the travel declaration. Usually, international trips must be related to higher costs than domestic trips. This tends to be true especially for a country with a relatively small area and a highly developed infrastructure like the Netherlands. Thus, this assumption was taken for the analysis – and afterwards validated. The average reimbursement amount for international declarations is almost 9-times higher (766.88 vs. 86.42).

Case Type	Mean Amount	Median Amount
DomesticDeclaration	86.42	41.51
InternationalDeclaration	766.88	509.07

Fig. 11. Comparison of declaration's amounts by declaration type.

In order to evaluate, if declarations with a higher amount are more likely to be rejected we grouped the declarations into five categories with an equally distributed number of occurrences. The largest category contains 4,342 declarations (category c), amount 50-200) while the smallest category has 2,625 declarations in it (category e), amount of more than 700). In general, our assumption can be confirmed. A higher reimbursement amount correlates with a higher rejection rate. However, category a) including declarations with an amount of less than 25 has a rejection rate that does not match this correlation as it is higher than the rejection rate in categories b) and c).

One of the potential explanations could be that declarations with very small amounts are created with less care and are therefore more likely to be rejected while declarations with a high amount require more detailed documentation, which might be missing in the first submission of the employees and therefore causing a rejection. Analyzing the roles responsible for the rejections does not really provide additional information towards this hypothesis as the main roles employee, administration and supervisor occur with a similar share for all five categories (see figure below).

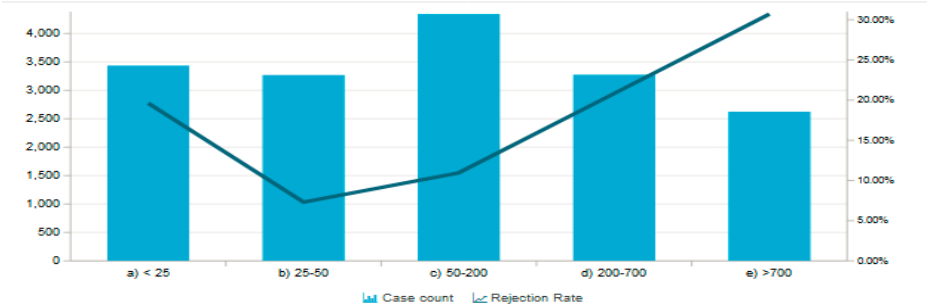


Fig. 12. Rejection rate by declaration amount.

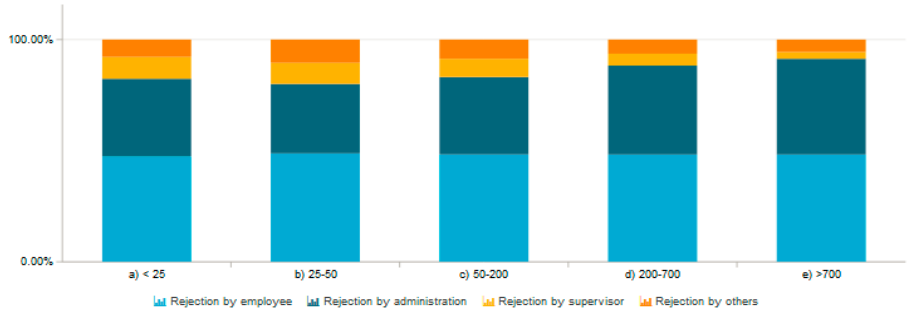


Fig. 13. Rejection roles per reimbursement account.

One of the significant observations proves that for higher amounts the share of rejections by the administration is relatively bigger which could mean that obvious flaws in the declarations are rather found in the first check by the administration than by supervisors in the second check. To clarify these assumptions, feedback from the workers within the process regarding their treatment of different reimbursement amounts would be required to validate the findings within the data.

Overall, we conclude that the throughput time of the declaration approval is more affected by the occurrence of rejections and the reimbursement amount rather than the difference between international and domestic trips. This is also supported in the final graph showing the relation between reimbursement mount categories and the throughput time between submission and approval of a declaration. In this figure, a similar distribution is visible as for the rejection rate. Smaller amounts are approved slower than medium-sized declarations while the largest amounts take the longest time for approval.

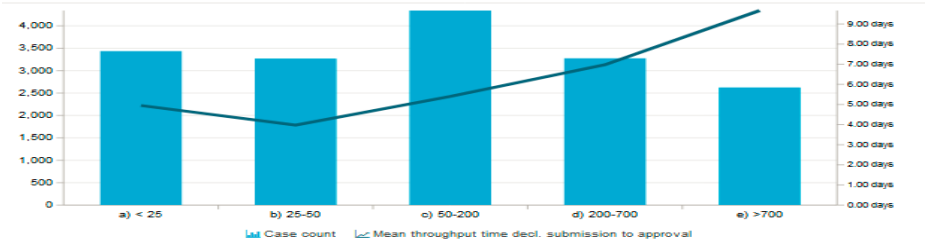


Fig. 14. Throughput time between declaration submission to declaration approval per reimbursement amount

There is another set of activities that domestic and international declarations have in common. This set consists of the two payment related activities Request Payment (not to be confused with the process entity Request for Payment) and Payment Handled. When comparing the throughput time between these two activities for international and domestic declarations it needs to be mentioned, that international declarations often use pre-paid travel costs, i.e. the process flows through the two payment related activities twice, before and after the actual trip. Therefore, it has to be ensured that throughput times are not calculated between the Request Payment activity before the trip and the Payment Handled activity after the trip as this would give an unfair disadvantage to the process structure of the pre-paid travel costs.

Taking this into account it can be seen that there are no significant differences in the throughput time for the handling of neither the first nor the last payment:

Case Type	Mean req. payment activities per case	Mean throughput time first req. payment to first handling	Median throughput time first req. payment to first handling
DomesticDeclaration	0.96	3.35 days	3.00 days
InternationalDeclaration	1.25	3.28 days	3.00 days

Case Type	Mean req. payment activities per case	Mean throughput time last req. payment to last handling	Median throughput time last req. payment to last handling
DomesticDeclaration	0.96	3.35 days	3.00 days
InternationalDeclaration	1.25	3.25 days	3.00 days

Fig. 15. Comparison of the throughput time differences between handling of first and last payments.

This leads to the conclusion that as soon as the payment is requested in the system the process is executed with similar speed disregarding any properties of the declaration. Therefore, this part of the process seems to be rather disconnected from the rest of the process and apparently does not bear a high potential for improvement. This is reasonable as this part is normally handled by the organizations accounting department and therefore potentially even implemented in a different system that is not affected by any properties of the process entities that have been checked in the previous steps.

Are there differences between clusters of declarations, for example between cost centers/departments/projects etc.? Metadata information is not fully available – in many cases, it is impossible to determine the differences due to unknown or unavailable dimensions. For instance, for 69% (21.9k) of travel processes project number is unknown or unavailable. The biggest known project cluster is ‘project 503’ with above

4.3k travel-related processes, with significantly higher throughput time (both average and median, 81 and 66 days respectively). While the throughput times for the whole population are 48 and 22 days (average and median), the results within project 503 are standing out.

Project	Case count	Total throughput time in days	Median total throughput time in days
-	21,405	48.10	22.00
project 503	4,343	80.73	66.00
project 147546	1,052	12.41	8.00
project 147556	947	13.75	10.00
UNKNOWN	525	12.92	2.00
project 147620	313	14.69	9.00
project 148052	301	14.41	10.00
project 147572	270	11.46	8.00
project 147582	253	13.39	10.00

Fig. 16. Throughput time and distribution of cases by project (excerpt)

A noticeable impact on these numbers is related to the origin of the processes – not only there are no domestic declarations within this project (which are being processed quickly), but instead there are extremely long travel permits (262 days of the throughput time in average, 171 permits). The international declarations are also more time-consuming on average (95 vs. 71 days)

Source Event	Target Event	# Occurrences	Throughput time
End trip	Declaration SUBMITTED	4,409	12.64 days
Declaration REJECTED	Declaration SUBMITTED	1,709	2.80 days
Permit FINAL_APPROVED	Declaration SUBMITTED	831	21.28 days
Send Reminder	Declaration SUBMITTED	374	14.55 days

Fig. 17. Basic KPIs per case type.

What is the throughput in each of the process steps, i.e. the submission, judgement by various responsible roles and payment?

Submission of Declarations, Requests for Payment and Permits. The submission of a travel declaration either happens immediately at the start of the process (domestic) or after the travel permit has been approved and the trip has ended (international). For domestic cases, it is difficult to measure the corresponding throughput time as no preceding activity is available in the data. For the international declarations, it can be noted that the submission happens normally immediately after the trip has been ended or a previous declaration has been rejected with the mean throughput times shown below:

Case Type	Total throughput time in days	Median total throughput time in days	Case count	Activities count	AVG activities per case
InternationalDeclaration	115.02	95.00	1,283	25,647	30.54
Prepaid Travel Costs	101.51	78.00	1,870	21,396	14.14
Request For Payment	15.71	9.00	1,021	5,651	6.05
Travel Permits	262.96	171.00	169	2,824	100.71

Fig. 18. Throughput times to declaration submission.

Submissions of request for payment are occurring either immediately at the beginning of the process (non-travel-related requests) or after the travel permit has been approved.

For the non-travel-related requests, there is no preceding event available to calculate a throughput time. For cases with a preceding travel permit, the submission normally follows on average 18.04 days after the permit has been approved.

For the submission of the travel permit the internal guidelines clearly state that this activity needs to be the very first step in the process. However, one can see that there are some cases where employees admit that the trip actually took place before the permit is created in the system. In these cases, the throughput times are distributed as follows:

Source Event	Target Event	# Occurrences	Throughput time
End trip	Permit SUBMITTED	512	30.28 days
Start trip	Permit SUBMITTED	327	21.38 days

Fig. 19. Throughput times to permit submission.

All submissions are executed by employees. Therefore, a benchmarking between various roles is not meaningful for this process step.

Approvals of Declarations, Requests for Payment and Permits. Travel declarations are mostly approved by administrative staff and budget owners. Supervisors and so-called pre-approvers support this step in some cases. It can be noticed that the approvals go much faster if they are executed by administrative staff or the pre-approvers. Especially for budget owners, who are responsible for a relatively large share of approvals, the throughput time is significantly higher (2.38 days).

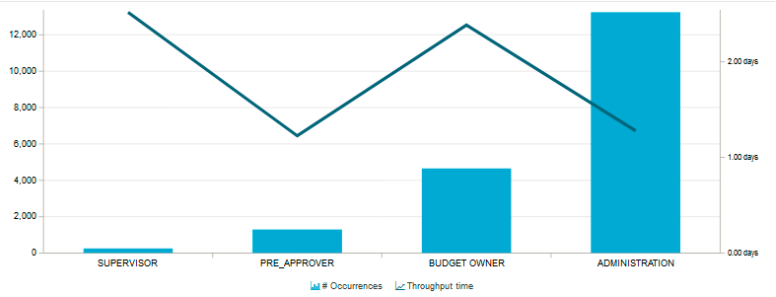


Fig. 20. Approval occurrences and throughput time per role.

The final approval step for declarations is most commonly executed by the supervisor role. As the final approval is conducted by the director role so rarely, the higher throughput time of their approvals does not have severe consequences for the overall process.

For *approvals* and *final approvals* of requests for payment and travel permits similar patterns can be recognized.

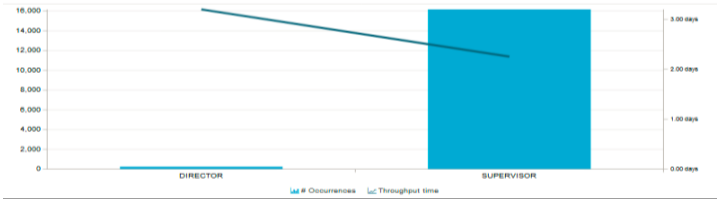


Fig. 21. Final approval occurrences and throughput time per role.

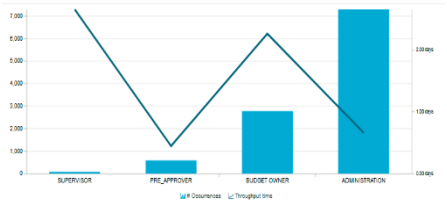


Fig. 22. Approval occurrences and throughput time per role for requests for payment.

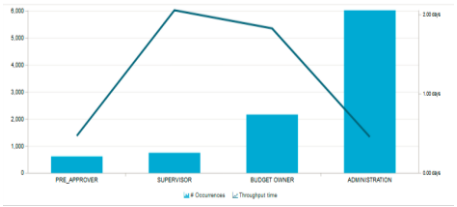


Fig. 23. Approval occurrences and throughput time per role for travel permits.

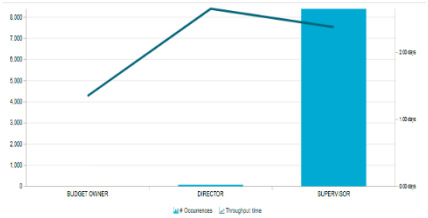


Fig. 24. Final Approval occurrences and throughput time per role for requests for payment.

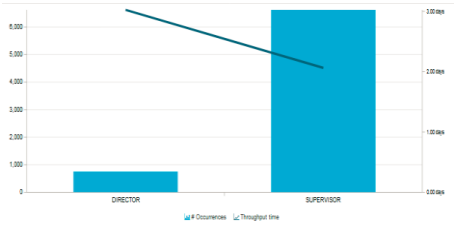


Fig. 25. Final approval occurrences and throughput time per role for travel permits.

Fig. 26. Rejections and approvals within process flow.

Where are the bottlenecks in the process of a travel declaration? Rejection of a declaration (happening in 17% of all declaration processes) extends the throughput

time by 3 days in average. With every repeated rejection, the throughput time is extended by another 3 days. If there is a resubmission of such a rejected declaration, the throughput time increases by 4 days.

Happening in 28% of all declaration processes the re-approval of a declaration implies another 2 days of the throughput time extension.

Where are the bottlenecks in the process of a travel permit (note that there can be multiple requests for payment and declarations per permit)? Analyzing the most common activities between the submission of a travel permit and the final approval shows picture presented on the figure 32. One can easily see that there are some activities happening before the permit is finally approved. The final approval is mostly done by the supervisor or director role (see question 4) which means that the activities like Start trip, End trip or Request For Payment SUBMITTED should not occur before this activity. This is clearly a deviation from the process flow. Furthermore, it can be noticed that the throughput time between the first approval and the final approval takes longer when there are other activities occurring in between. Direct connections take 2 days on average while the presence of Start trip leads to an increased throughput time of 4 days.

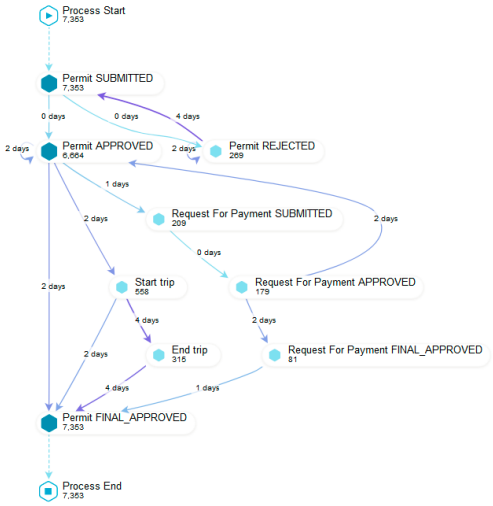


Fig. 27. Process flow between submission of a travel permit and the final approval.

Another obvious bottleneck is the rejection of the permit. After the rejection, it takes on average 4 days until the permit is resubmitted. In some cases, another rejection occurs 2 days later. As a final finding, cases with more than one request for payment or more than one travel declaration per permit indicate an increased throughput time as well. We would not consider this finding as a typical bottleneck of the permit process itself. The employees are aware that submitting additional requests for payment or additional declaration will start the corresponding approval process all over again. The permit steps of the overall process are not directly affected by unnecessary requests

for payment and declarations that are submitted after the permit is finally approved. This can be seen as there are no cases where the submission of a request for payment or declaration leads to a resubmission of a travel permit (see process graph above).

How many travel declarations get rejected in the various processing steps and how many are never approved? 17% (2,877) of all declarations get rejected in the declaration-related steps. For 465 declarations it is the last process step – nothing happens after the rejection. It is important to note that almost 80% of rejected declarations (2,276) get rejected twice and there are declarations with even more rejections, up to 12 times.

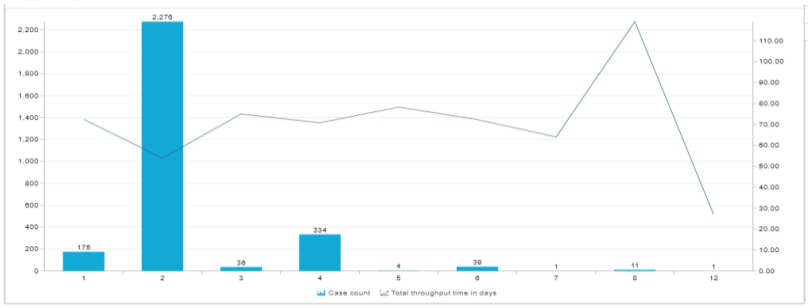


Fig. 28. Number of rejections per case.

There are 2 types of approval: a normal approval and a final approval, whereas the second one is more common (14.2k vs. 16.3k). 2,536 (out of 16,740, so 15%) submitted declarations did not get approved, 466 never flew through final approval. 380 (above 2%) declarations were submitted and never received any kind of approval.

How many travel declarations are booked on projects? Only ca. 8% (1,283) declarations are booked on projects and all of them are booked on one project (project 503). For the remaining 92% of declarations, the project was not specified.

Comparing domestic and international declarations, it can be noticed that no project is available for any domestic declaration (out of 10,500); for international declarations, the share of declarations booked on the project equals to 20%.

An analysis of travel related documents different from declarations reveals interesting insights. 93% of requests for payments (RfP) are booked on projects, most commonly on project 147546 (1,040, 15%), project 503 (1,021, 15%) and project 147556 (931, 14%). Only 7% RfP have no projects assigned.

Looking at the process, specifying a project seems to be mandatory for submitting a RfP as opposed to declarations where it is only optional or even impossible (compare domestic declarations). It appears that the project number is not added to the declaration at the later stage either. Lack of the project number may impede comprehensive assessments of the travel process and its optimization.

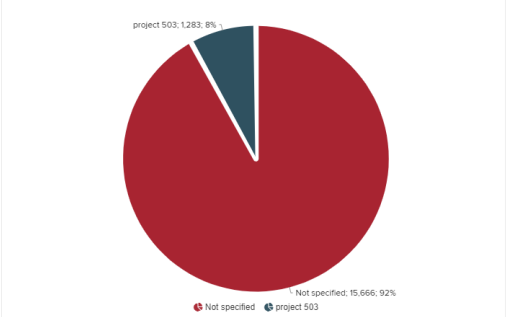


Fig. 29. Declarations and their respective projects.

How many corrections have been made for declarations? Since the data does not contain any exact information on corrections made on declarations, for the purpose of this analysis, it was assumed that every resubmission of a declaration is related to its correction. In that sense, 2,890 corrections have been made for 2,424 declarations. It means that 14.5% declarations have been corrected, and some of them (393) more than once. These corrections took 12 days on average (throughput time between the first and the last submission of the declaration, median 5 days), contributing towards slowing down the declaration process.

Are there any double payments? There are 1222 process instances with double payments (ca. 6% all process instances with payments). Some of these payments are related to pre- and post-travel payments.

Are there declarations that were not preceded properly by an approved travel permit? Or are there even declarations for which no permit exists? In the whole data set we spotted only one case where declaration was submitted after the permit in test period of 2017. What is more concerning is that there are 438 unique cases where a permit was never submitted. If we exclude those cases for which the first activity was before 2018 we have still 261 declaration submitted without permit. For 250 of those cases the payment was handled. In the process for those cases, we can see several irregularities.

Firstly, they are 29 cases where the trip actually started before the declaration was submitted (e.g. declaration 147020). For one of them the declaration was actually submitted during the trip, rejected and resubmitted after trip ended (declaration 147198).

Secondly, we have 56 cases where initially declaration was rejected but resubmitted and in 52 cases, it was accepted although there is no travel permit. For 126 cases (so more than half of them), the declaration was accepted at least twice before it was finally approved. We can clearly see that such cases raise attention but usually they were accepted anyway.

After checking available dimensions, we spotted that all of those cases are related to „project 426“. Moreover, for all of them there is information about travel permit number but those travel permits are not available in provided data set. With closer examination, we can see further irregularities. E.g. declaration 145020 has permit travel 145022 but ID of event start trip and end trip is „rv_travel permit_423_6“ and „rv_travel_ermit_423_7“. So it looks like generated for travel permit 423. If we check travel permits we can find travel permit 423 which actually include events „rv_travel permit_423_6“ and „rv_travel permit_423_7“ with exacly same timestamps but related to declaration 429. If we check for this specific event ID, we clearly see that all events without travel permit actually contain start trip and end trip activity with this specific event ID. It may look like a trivial system error but the value of all declarations with such situation is nearly 250,00 0€, which is substantial.

How many travel declarations are submitted by the traveler and how many by a mandated person? Unfortunately, it is not possible to distinguish whether a declaration was submitted by a traveler or a mandated person. The only information available is the resource type and role name, which for all submissions are displayed as ‘staff member’ and ‘employee’ respectively, what can be seen in the table below.

EVENTNAME	RESOURCENAME	ROLENAME	Activities count
Declaration SUBMITTED	STAFF MEMBER	EMPLOYEE	19.630
Declaration SAVED	STAFF MEMBER	EMPLOYEE	210

Fig. 30. Declaration submission and saving with respective resource and role names.

However, we have noticed there were also declarations saved – 210 times. While it could mean a traveler just saved the intermediate results, it could be also assumed that the declaration was prepared by a mandated person and only then submitted by traveler. Nevertheless, after a thorough investigation of these declarations we concluded that only 1 of these cases got eventually submitted, therefore we cannot assume the action of saving is linked to mandated person.

For a detailed answer to this question, a distinction between traveler and a person submitting would be required, e.g. in form of an additional resource id (pseudonymized) to distinguish different users and an extra dimension about who the declaration concerns.

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? 309 declarations were submitted later than 2 months after the end of the trip (first submission not earlier than 61 days after trip end). Of those, 84 were rejected (and each of them re-submitted as well). Interestingly, among these declarations there are 15 approved first and rejected afterwards, usually with first approval from the administration and further rejection from supervisors. It is worth mentioning that the vast majority of late-submissions (225 out of 309) was approved without any rejections. Therefore, while rejections for late-submissions are to be found within the data in scope, it cannot be confirmed that these

rejections are due to late submissions. The data is not detailed enough to determine any causation and at the same time there are examples of direct approvals.

If, according to the guidelines and process, every declaration submitted later than 2 months after the end of the trip should be rejected, the system settings could be adjusted to either not allowing any submission after this period or redirecting such a declaration on a special approval path, eventually requesting some extra input. That could save time for both submitters and approvers, eventually leading to savings.

Is this different between departments? It was noticed that within the most common organizational unit (considering permits), organizational unit 65458, the share of late submissions is higher (7%) than in other common organizational units (usually oscillating around 3%). Nevertheless, the rejection rate is constant – regardless of organizational unit, there are no statistically significant deviations.

How many travel declarations are not approved by budget holders in time (7 days) and are then automatically rerouted to supervisors? In the data provided, there were 536 declarations approved by a supervisor 8 or more days after the first approval by administration. These declarations did not have budget owner involvement, what suggests an automatic rerouting after the 7-days-deadline (since there might be cases where the budget owner and supervisor are the same person, so there is only one approval foreseen, these circumstances were taken into consideration by filtering on exceeding the approval timeslot). 65% of these declarations were international, whereas 35% of them were national – all domestic belonging to the same budget number: 86566.

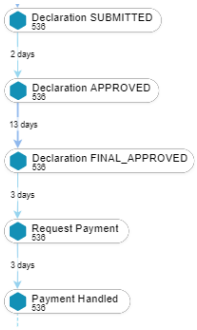


Fig. 31. Process flow with throughout times for declarations rerouted to supervisors.

For these rerouted declarations, the *final approval* by *supervisor* took in average 13 days after the first declaration *approval* by *administration* (median 11 days), extending the submission and payment process to 23 days in average. Within these rerouted declarations, only 18 come from the pilot phase (2017), therefore system implementation cannot take the fall for missing decisions.

Unfortunately, the eventlog data does not indicate rerouting, therefore only the time component and the role were taken into consideration – having the rerouting information would allow a more detailed investigation of this issue. Moreover, it would be

interesting to analyze data about original budget owner, e.g. his/her department, declarations approval workload as well as share of declarations for which a decision was missing. Often, such analysis, linked to time series analysis reveals trends and interesting insights and helps in optimizing approval workflows.

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?. During the dataset discovery, we noticed that all events *Request For Payment SUBMITTED* were conducted by 'staff members' assigned to a role 'employee'. Since this division completely contradicts the process description and the classification might be too general, we decided to approach this question in a different way, namely by the analysis of the whole process flow.

Request for Payment submitted followed by *Declaration SUBMITTED* indicates a suboptimal process flow, which might be related to incorrect request type at the beginning. In other words, if an employee submits the request for payment instead of a declaration, it will be discovered at a later stage and the employee will need to submit a declaration as well. In the analyzed dataset, such process flow was observed 1217 times, most commonly at the beginning of 2018 (see figure 38). That would suggest employees were successfully learning how to use the new system.

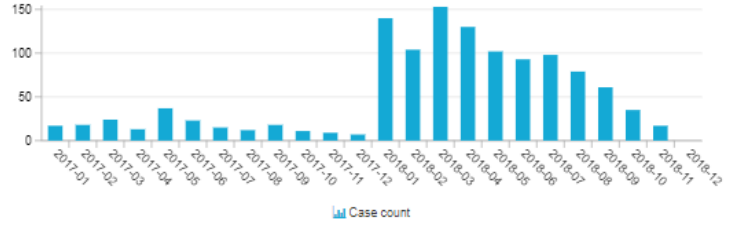


Fig. 32. Number of submitted requests for approvals followed by declarations per month.

It should be mentioned that all these requests were assigned to *"project 503"* accounting for almost 50% requests on this project in total (1217 out of 2472). Even though this project is the most frequently used one, this significantly high share of RfP-Declarations booked on this project implies questions on what is this project precisely.

4.2 Financial analysis

To provide a comprehensive and real-life analysis of this travel reimbursement process, we decided to look into financial details of requests since provided questions do not refer to financial data at all. Additionally, in the provided data set several dimensions are linked to financial aspects of the trip, especially topic of undervaluation of the trips. This is the reason why the first question we tried to answer is:

What is the cause for trips to be overspent from process perspective? The overspent in 2018 equals to 12,820.27€, on average 248.69€ per overspent trip. These are 11% of all cases in 2018. This information was taken from travel permits and added to all cases

where a permit is available. In addition, we excluded events from 2017 since those are from a pilot period of the system and might therefore influence the analysis negatively. Afterwards, we grouped same event types and calculated the ratio of overspent cases for all event types. Furthermore, the overall case count for each event type has been added to the figure 38.For all cases with a trip included as an event on average 26% of them were overspent. However, not every trip ended with payment handled. For paid requests, 35% were overspent. Everything above 35% we can treat as events that correlate positively with overspent cases. We can see two main findings. Firstly, if request for payment is involved, those cases have a higher chance to be overspent especially for those where request for payment was rejected or saved. Secondly, if the declaration is initially rejected there is slightly higher chance that this trip will be overspent (38%). We also checked if overspent cases affect throughput times but there is no visible correlation.

Fig. 33. Ratio of overspent (top) and underspent (bottom) cases per event type compared with case count

What is the cause for trips to be underspent from process perspective? We also spotted that for many more trips the estimation is much higher than actual cost of the trip. Only for 12% of trips, the estimation was perfect, 77% of cases were underspent. It seems that the impact of these requests is smaller than the overspent cases. Nevertheless, it indicates that before the actual payment is made, money is allocated into the trip

that means that the financial liquidity of the organization can be affected. It is relevant since in 2018 the trips were overvaulted by 192,558.71€, on average 109.19€. After closer examination of specific activities, we concluded that rejected travel permits and reminders are correlated with underspend trips.

Ultimately, it was verified that estimating the trip perfectly is a challenge. Although overspent cases tend to happen less often, yet they account for 11%. For some events, a chance of whole trip to be overspent is higher. We suggest looking closely for trips where the declaration was rejected initially or RfP was submitted and processed. Furthermore, over ¾ of all cases for 2018 were underestimated, particularly if a travel permit was initially rejected or a reminder was sent.

5 Conclusions and Outlook

Performed analysis of the data set allows us to conclude that the as-is process is not far apart from the description provided in the challenge statement. However, our efforts to answer the more detailed questions showed that there are obvious deviations in the data set and that the performance of some process steps is not entirely consistent across all cases.

Our analysis indicated that rejections of declarations correlate with many other incorrectly executed activities like overspending and delays in the declarations' approval. Unfortunately, the data set does not include details on the rejection reasons. We were only able to conclude that low and high reimbursement amounts are more likely to be rejected than declarations with a medium sized amount. A critical unanswered question is therefore: What are the reasons for rejections and what do employees need to change in the declarations so that they are approved after resubmission.

Two approaches could provide answers to this key question. A data driven approach would be to check if the underlying system is able to track changes in the travel declarations. In many state-of-the-art systems designed to implement processes of such financial impact, a change log or history table is maintained in order to track the exact changes in the fields, the users who made those changes and the time in which they were made. Availability of said data in the system would allow us to understand what changes are made in order to get an approval after resubmission.

The classic approach would require conducting interviews with experienced employees responsible for rejections who would be able to provide the most common reasons for rejections. However, as interviews are costly, time-consuming and often subject to personal bias, the first approach is preferable.

To carry out a profound analysis with focus on the root-causes, it would be necessary to expand the data. Currently, several relevant dimensions are either not available at all (e.g. department) or not correctly populated (project ID, budget ID). This acts as a powerful constraint for a thorough root-cause analysis. In some cases, it might be related to GDPR and personal data. If a more detailed analysis was desired, we would definitely recommend a project setup that allows the usage of a more complete data set.

6 Appendix

Appendix_0.1	Python script transforming the .xes data to tables in MS SQL database.
Appendix_1.0	SQL script creating a merged overview of activities occurrences per source file.
Appendix_1.1	Result of the SQL script from appendix 1.0.
Appendix_1.2	Assignment of activities to process entities.
Appendix_2.0	Image depicting the most common process variants for case types domestic declaration and request for payment.
Appendix_2.1	Image depicting the most common process variants for case type international declaration.
Appendix_2.2	Image depicting the most common process variants for case type international declaration with payment before start of the trip.
Appendix_2.3	Image depicting the most common process variants for the case type travel permit.
Appendix_3.0	Source code to create the merged data model.