Business Process Intelligence Challenge 2020: Investigating Business Trips Arrangement Process at the Eindhoven University of Technology (TU/e)

Anastasiya Pakileva[1], Ekaterina Skvortsova[1], Nikita Zakoryuchkin[1], Sergey Tsaplin[1] , and Vsevolod Zarubin[1]

[1]Sberbank, Department of Centralized Data Processing, Russia
{appakileva,evikskvortsova,nyzakoryuchkin,tsaplin.s.t,
vvzarubin}@sberbank.ru

Abstract. A clear process of papers submission for business travels is a key point in effective use of resources. As such automated processes are usually logged, there is an opportunity to conduct a complex analysis and evaluate the system health and find areas for potential improvements. The main purpose of this research is to conduct a detailed analysis of the provided logged data from the Eindhoven University of Technology (TU/e) in order to highlight bottlenecks and deviations in the existing process of travel arranging and costs reimbursement and to propose relevant solutions for the current difficulties elimination and costs reduction. In order to perform a deep analysis we used the state-of-the-art tools of data and process mining combining two approaces: out-of-box solutions such as *Disco* and *Einstein Analytics Studio* and python libraries *pm4py* and *sklearn*. An approach based on statistical tests and time-series analytics accompanying the analysis with the explicit visualizations helps researchers to deduce valuable business and technical inferences. The gained results show that the processes that consist of several etalon traces require the clarification among the university staff members and as far as the technical part is concerned, there should be system restrictions installed in order to escape the confusions.

Keywords: Business Process Intelligence, Process Mining, Bottlenecks and Anomalies Detection, t-SNE, DBSCAN, KMeans

1 Introduction

To extract the real value from the presented data we need to be able to store it properly, to accurately check its conformance to the expected behavior, to uncover potential deviation cases. Using the knowledge gained above helps to improve the efficiency of the held processes, extract business value. This start-to-end path became real thanks to relatively novel and cutting-edge field of applied data minig - process mining. [6]

The process owner of the BPI Challenge 2020, Eindhoven University of Technology (TU/e), provides 5 real-time event logs for 2017 and 2018 years, that reflect processes of travel arranging and costs reimbursement.

The reimbursement process at TU/e - is an important part of the organization's operations. This process can be the subject to business risks, such as incorrect submission of declarations, decision-making delays by various responsible roles, unpaid declarations, multiple paid declarations. Investigating and understanding these risks means understanding how effectively the company's monetary, time and human resources are used and allocated in the process.

Process-mining is considered to be an applied data analysis that is why lots of techniques used for data mining can be applied in terms of log-based data analysis. In this paper we demostrate the efficiency of *exploratary data analysis* supplemented with powerful illustrations for a high-level analytics, *clustering analysis* [8] for splitting log into homogenous subsets and further analysis of common patterns and deviations, *anomaly detection* techniques for bottlenecks detection [3].

2 Materials and methods

2.1 Genral assumptions

There were five event log files provided by TU/e. For the further convenience, the following abbreviations are used in the provided research:

- Domestic declarations log – DD
- International declarations log – ID
- Requests for payment log – RfP
- Prepaid travel costs log – PTC
- Travel permits log — TP

The data inside the provided logs is organized in a similar way and often follows a similar flow. However, when analyzing the raw data, the redundancy and ambiguity of the information can be immediately noticed.

Event log traces were organized into groups, and thus separated by logical correspondence. Using this distinction, we managed to combine several event logs that have similar processing features. This allowed us to describe a complex case of international declarations and highlight the deviations.

2.2 High-level analysis of throughput of travel declarations

To figure out the throughput of a travel declaration we considered processing time from the first submission (logged re-submissions were not considered) to the first handled payment event (subsequent payment events were ignored). The sample

included declarations launched in 2018.

Table 1. DD 2018, descriptive statistics

	Throughput, hours
average	278.9
std	339.3
min	25.5
25%	146.7
50%	193.1
75%	312,9
max	6981.7

According to the Table 1, the average is very far from the mean, the large gap between the minimum and the maximum processing time, volatility is confirmed by the high value of the standard deviation, that also exceeds the average.

Table 2. ID 2018, descriptive statistics

	Throughput, hours
average	364.9
std	428.8
min	26.1
25%	169.7
50%	260.1
75%	414.4
max	10298.7

The situation is similar in ID. The average throughput is very different from the mean and there is a huge gap between the minimum and maximum values, which indicates the presence of anomalous observations in the sample. Since the average is not much shifted towards the maximum value, it can be assumed that such cases are few.

More detailed analysis of the processing time will be facilitated by detailed checking of the time spent on transitions between events.

2.3 Detailed analysis of each step in travel declarations throughput

For the DD and ID, the average time of transition from one event to another was

calculated with the help of pm4py library tools.The duration of each process step is reflected in the Fig. 1.

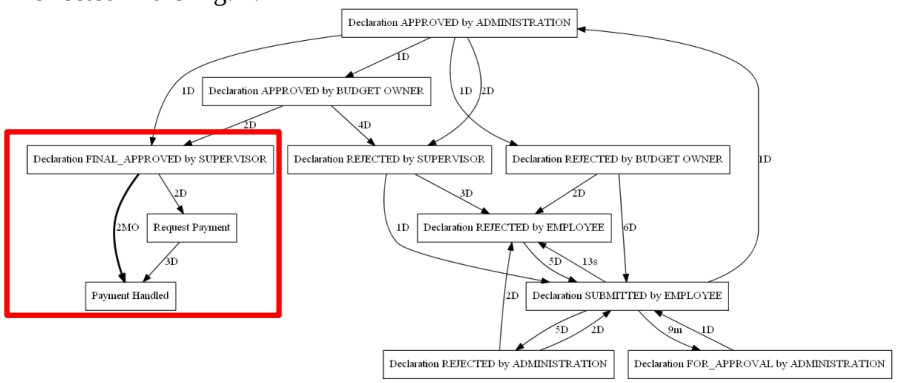

Fig. 1. DD processing graph with timings, 2018

On average, the longest transition is from supervisor approval to payment handled states. Table 3 shows five longest transitions:

Table 3. Top-5 longest transitions for domestic declarations, 2018

№	*Start case*	*End case*	*Throughoutput, days*
1	Declaration submitted by employee	Declaration for_approval by administrantion	20,5
2	Declaration rejected by employee	Declaration submitted by employee	7,1
3	Declaration submitted by employee	Declaration rejected by employee	7,1
4	Declaration rejected by administration	Declaration submitted by employee	6,5
5	Declaration submitted by employee	Declaration rejected by administrantion	6,5

According to the Table 3, it took 20 days for the Administration to approve the declaration. Since the number of such events is low, it can be assumed that the bottleneck is execution by staff member with Administration role. Either the monitoring system did not alerted this need or the staff member did not know that he takes part in the process as the approval executor.

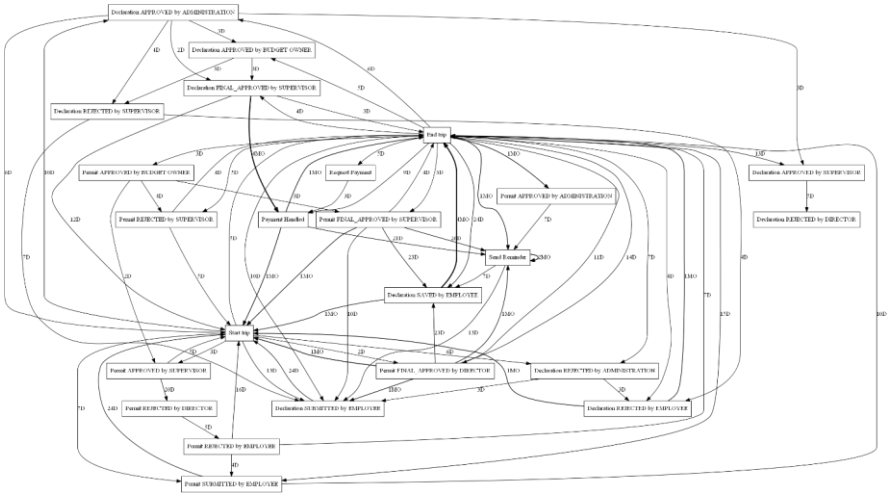

Fig. 2. IDs processing graph with timings, 2018

A similar table on the longest processing times for international declarations is provided.

Table 4. Top-5 longest transitions for ID, 2018

№	*Start case*	*End case*	*Throughoutput, days*
1	Permit final_approved by director	Send Reminder	88,5
2	End trip	Permit approved by supervisor	83,5
3	Permit approved by supervisor	End trip	83,5
4	End trip	Permit final_approved by director	81,1
5	Permit final_approved by director	End trip	81,1

The numbers are 4-10 times higher than for the DD throughput. To understand the reason of the differences the most frequent declaration traces were considered and differences in processing that caused deviations in timings were visualized. In domestic declarations, 85% of cases go by the following process flow:

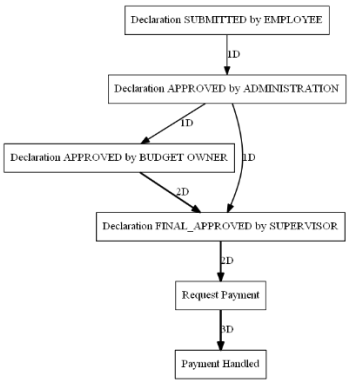

Fig. 3. DD process flow, 2018

As for IDs, only 40% of the cases follow the process flow described in competition rules. Such statistics indicates that most cases are processed much longer because of rejections, loops and failures, in other words, because of suboptimal procedure runs.

Thus, it occurred possible to visualize the difference in the processing of the declarations. Ideally it takes for a domestic declaration 7-9 days to flow all steps before the payment. The processing of international declarations takes longer because travel permits are submitted in advance, about a month before the trip and the processing of the declaration after the trip is two days longer than for the domestic declarations.

2.4 Exploratory analysis of differences in travel declarations throughput

As the first step in the analysis of the differences between the throughput in presented processes was – the visual analysis and graph plotting. The metric to reflect was the number of simultaneously processed declarations as it reflects the workload of the staff members that take part in the travel declarations processing.

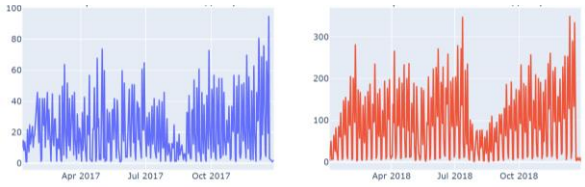

Fig. 4. Time-series graph for the amount of simultaneously processed DD in 2017 (left), 2018 (right)

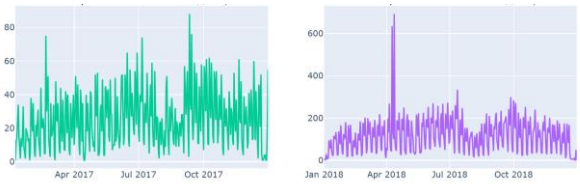

Fig. 5. Time-series graph for the amount of simultaneously processed ID in 2017 (left), 2018 (right)

There can be derived a common feature for both years and both declaration types: there is a constant decrease in the number of processed declarations in August and a relative increase in the previous month – July. The university should consider the peculiarity of these processes and allocate the resources for declarations processing carefully.

However, there is also a difference for the presented processes. There was expressed a vivid peak in the number of declarations processed per day in April, 2018. Perhaps, it is due to the filing of applications for a conference, but there was no such surge a year before. Such a hit also can be explained by long-term non-effective resources allocation in the previous months and the staff members had additional assignments that influenced their productiveness

In order to draw a conclusions about the representativeness of the statistics, it is necessary to analyze the stability of the characteristics that the time-series demonstrates. That is why the Dickey-Fuller test for the stationarity and ACF, PACF analysis was performed.

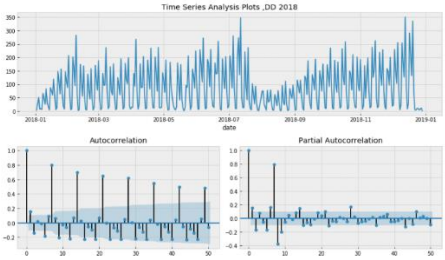

Fig. 6. Time-series, ACF and PACF plots for domestic declarations, 2018

According to the test, p-value = 0 that is less than the acceptable significance level $\alpha = 0{,}01$. The null hypothesis of non-stationarity of the time-series is not accepted with type 1 error probability equal to 0,01. The autocorrelation function clearly showcases the weekly seasonality, that is logical because the number of processed declarations increases in the beginning of the week and decreases to the end.

The same situation is presented in the process of ID processing per days: p-value equals to 0, the time-series is stationary and weekly seasonality is also admitted.

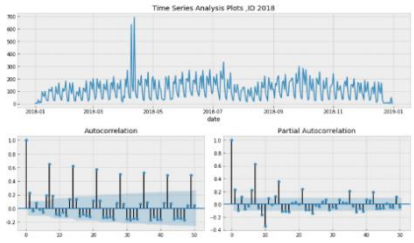

Fig. 7. Time-series, ACF and PACF plots for international declarations, 2018

It was worth checking whether the number of declarations processed per day affects the throughput. As the criterion for the throughput, the number of changed events in terms of one declaration was used. The time-series that has been constructed from the average number of changed events in terms of one case id per day and according to the ADF test is also stationary. Therefore, the Granger casualty test was done, and it showed that changes in terms of the number of processed declarations influence the throughput, because all p-value < 0,05 - acceptable significance level. The full listing of the test results is listed in Fig. 8 and Fig. 9.

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=7.8535  , p=0.0054  , df_denom=355, df_num=1
ssr based chi2 test:   chi2=7.9199  , p=0.0049  , df=1
likelihood ratio test: chi2=7.8336  , p=0.0051  , df=1
parameter F test:         F=7.8535  , p=0.0054  , df_denom=355, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=8.4591  , p=0.0003  , df_denom=352, df_num=2
ssr based chi2 test:   chi2=17.1585 , p=0.0002  , df=2
likelihood ratio test: chi2=16.7589 , p=0.0002  , df=2
parameter F test:         F=8.4591  , p=0.0003  , df_denom=352, df_num=2
```

Fig. 8. Granger casualty test, DD

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=10.6449 , p=0.0012  , df_denom=361, df_num=1
ssr based chi2 test:   chi2=10.7334 , p=0.0011  , df=1
likelihood ratio test: chi2=10.5782 , p=0.0011  , df=1
parameter F test:         F=10.6449 , p=0.0012  , df_denom=361, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=12.6251 , p=0.0000  , df_denom=358, df_num=2
ssr based chi2 test:   chi2=25.6029 , p=0.0000  , df=2
likelihood ratio test: chi2=24.7403 , p=0.0000  , df=2
parameter F test:         F=12.6251 , p=0.0000  , df_denom=358, df_num=2
```

Fig. 9. Granger casualty test, ID

The trend-seasonality decomposition was performed on both (DD and ID) time-series in order to find any hidden patterns and tendencies.

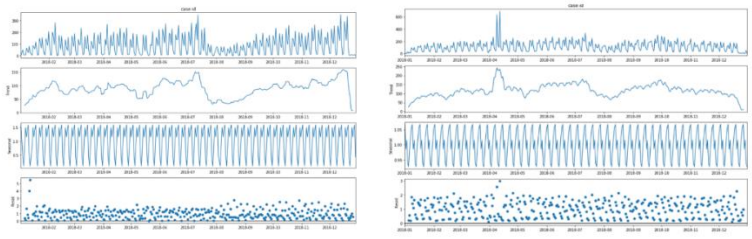

Fig. 10. Seasonal and trend component and variance decomposition for DD (left) and ID (right)

There was no trend component found in the DD data and in the ID data, but only weekly seasonality was recorded. The volatility is low enough and there are several surges. In general, such a behavior is typical for stationary processes as confirmed by the tests above. That means that both processes demonstrate constant characteristics and the data can be used for prediction modelling. The TU/e management can use the data in the forecasting of the capacity prediction and elimination of critical periods with lack of resources.

2.5 Clustering analysis as a tool for epmhasis of common features and deviations

In order to perform a more detailed analysis, a clustering analysis on the basis of python library *sklearn* is provided. To visualize the results, python libraries *plotly* and pm4py were used as the visualization tools. The presence of an event in each trace was used as a feature for clustering. The clustering analysis was divided into two stages.

Firstly, the feature space was projected onto a plane using t-SNE technique. This chosen technic showed more shaped clusters, whereas PCA and SVD dimensionality reduction techniques showed a more sparse projection.

After the dimension reduction based on the obtained components clustering was performed by the method, that takes into account the density of the cluster of points - DBSCAN. This method was chosen in order to find the most common examples of event sequences in traces.

The event-based clustering is justified by the fact that considering all the traces in tables in general does not give a sense of patterns that are hidden inside. Moreover, DBSCAN clustering can help to find large anomalies [9].

Common traces selection

The clustering analysis showed that during the domestic declaration processing one type of trace covers almost 50% of possible passes with minimal deviations. There are few points on the plane, but in fact a large number of traces may be hidden behind the projection of a single point.

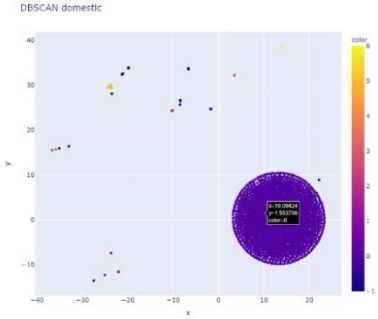

Fig. 11. Results of the clustering analysis, domestic declarations (DBSCAN, t-SNE projection)

This cluster showcases the most popular version of the declaration processing, its trace is reflected on Fig.11. The graph was plotted with the heuristic miner algorithm that reflects only common cases and escapes according to the frequency threshold not important events:

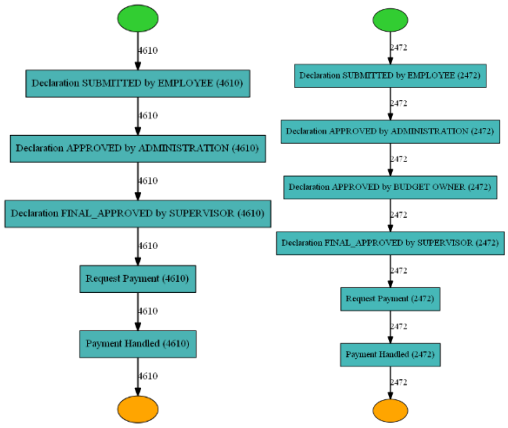

Fig. 12. The most popular processes graphs DD (left), ID (right), 2018

The next most popular cluster includes only one more event - budget owner approval, it is the additional but not obligatory option for some cases. This trace type also aligns the process flow described in rules and is not considered as a deviation. Another feature that also deserves attention is that there are 100 cases in which employees simply saved declarations and did not send them further. Such a case occurrence means that 100 employees did not manage to fill the declarations, the university should pay more attention for clear process flow description.

The existence of TU/e employees that submitted a request for payment instead of a travel declaration and double payments for declarations

The analysis of the results of the performed cluterization for permission log covered two questions simultaneously: the events outlined in red rectangles in the Fig. 13 highlight cases when the TU/e employees submitted a request for payment instead of a travel declaration the events outlined in orange rectangle highlight thecases of double payment for some declarations.

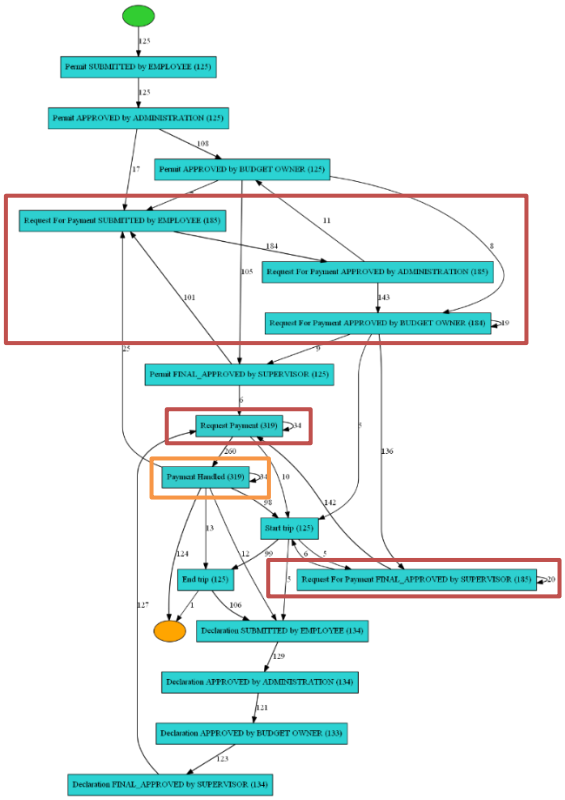

Fig. 13. Graph of deviations in TP log

Similarly, in the clustering of international declarations, it was found that some declarations have skipped approval stages. It means that some employees have started trips without approval of responsible persons.

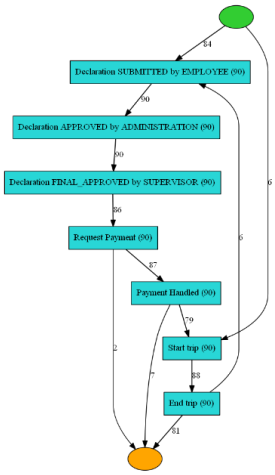

Fig. 14. Graph representation of a cluster in ID withought necessary travel permissions

The university should apply restrictions on the process of documents approvals for international trips. Such a case as highlighted in the Fig. 14 proves that employees can submit declarations and go on an international trip without any obligator permissions.

Cluster analysis also highlighted the bottlenecks - the loops – when the declarations were rejected by Administration and then resubmitted by employee over and over again. Possibly due to the opacity of the declaration process for employees.

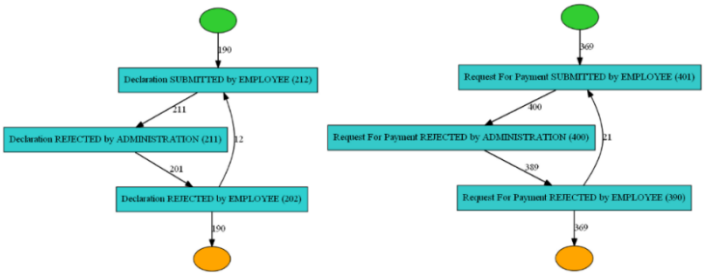

Fig. 15. Cyclic graphs from clusters in international declarations (right) and requests for payment logs (left)

To address this problem, the university is encouraged to create a publicly accessible infographics in the form of a decision tree.

2.6 Differences between case attributes: departments, projects etc.

Initially, each event log has several keys. The latter can be case id, declaration numbers, id requests for payment, permits, and others. To answer this question, the

previous model was used, in which the trace attribute “case: concept: name” is used as a common unique key. The event log traces were concatenated, using this key. Thus, a single file containing global attributes was obtained. The completeness of the description of this file is ensured by analyzing the attributes vector: Organizationalentity, Activity, Project, Budgetnumber, case:concept:name. Fig 16 provides an analysis of departments, projects and budgets.

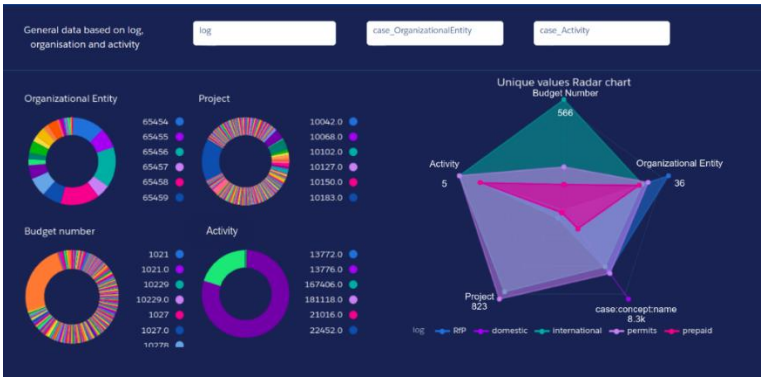

Fig. 16. The distribution of traces in the general event log by attributes Organizational Entity, Activity, Project, Budget Number, case:concept:name

The four piechart on the left show the part of the total number of traces for the current attribute for the current filters. The radar chart on the right shows the number of unique values for each analyzed attribute for each log.

The figure clearly shows that the largest number of departments (36) falls on the RfP. At the same time, 20 departments are common to all event logs, except for DD. TP and ID are very similar in all attributes except BudgetNumber, and ID has the greatest variety of budgets.

On the radar chart on the right, RfP and PTC are close, but the Rfp file has a larger amount of data and more variety of cases. As for DD, there is no division between departments or projects for domestic declarations, and all cases fall into a single budget.

On the radar chart DD has a set of values only for the case: concept: name attribute, however, it can be seen that the DD process is the most demanded, since it has the largest number of cases among all event logs.

The color distribution of pie chart shows that lots of traces fall into various departments, budgets, projects. Despite this, there are also popular attribute values. The vast majority of traces are split into two activity values. The situation with the rest becomes a little more certain when choosing event logs and fixing attributes. Einstein Analytics Studio can also filter values by log, OrganizationalEntity, Activity, but the corresponding figures are not presented here.

It was found out that: almost all PTC traces fall into one Project; Activity separates PTC, RfP and ID, TP; RfP does not have a Budget, while PTC has about

100 ones, most of the ID and TP belong to less than 10 departments - the same situation with PTC and TP. In ID, among others, there is one of the most popular projects with code 426, which corresponds to almost half of ID budgets and 464 traces. The most sense makes considering the distribution by department. Table 5 below shows the most popular departments and their percentage of the total.

Table 5. The distribution of traces in event logs by department. The values in the files are encoded as 654 **. The last two digits are used. The share less than 5% is not taken into account.

Fraction	*RfP*	*PTC*	*TP*	*ID*
15 – 20 %	-	61	56, 58	56, 58
10 – 15 %	58, 61, 69	54, 58, 69	54, 55, 59, 60	54, 55
5 - 10 %	54, 56, 62, 63, 65, 68, 82	56, 62, 63, 65, 68	57, 64, 66	57, 59, 60, 64, 66

Table 5 shows that there are common departments for logs: 54, 56, 58, that also have the most cases. As for departments accordance, TP basically covers ID, RfP overlaps PTC. RfP contains most of all departments, but the rest have 9 departments. It is also worth noting that the distribution of Rfp and PTC by department is completely identical when considering costs rather than the number of traces. Also, there are only a few requests for payment in the data, the amount of which is 3 orders of magnitude higher than all the others. These unique cases fall into Department 68 and Budget 1327 from PTC or Rfp id 166647 from RfP. The rest of the data is very scattered, there are rare cases, as can be seen in Fig. 17, that cannot be ordered. Therefore, the previously described technique of dimensionality reduction based on trace signs was used. Further, clustering was done using the KMeans method. The results are shown in Figures 17 - 19. The more obvious cases of DD and Rfp are not presented here.

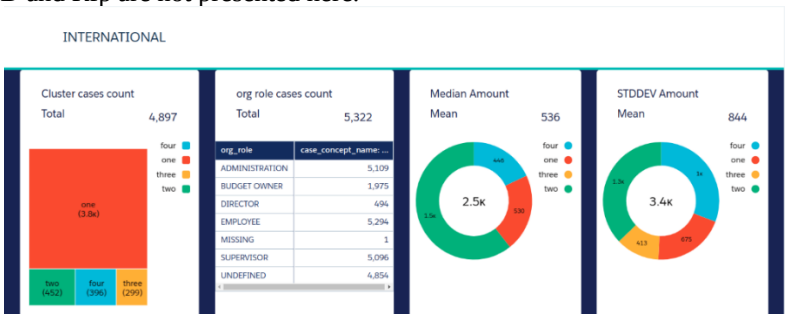

Fig. 17. International declarations log Amounts analysis on KMeans clusters

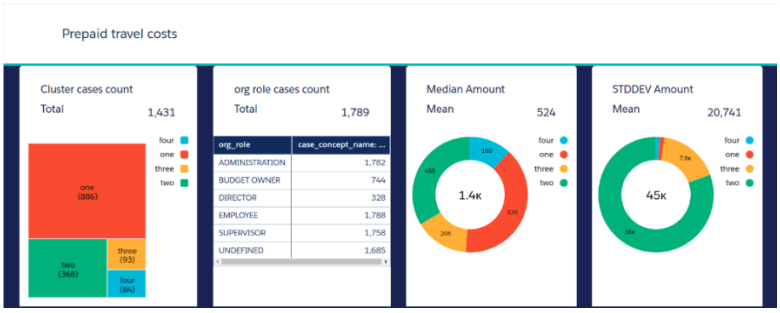

Fig. 18. Prepaid travel costs log Amounts analysis on KMeans clusters

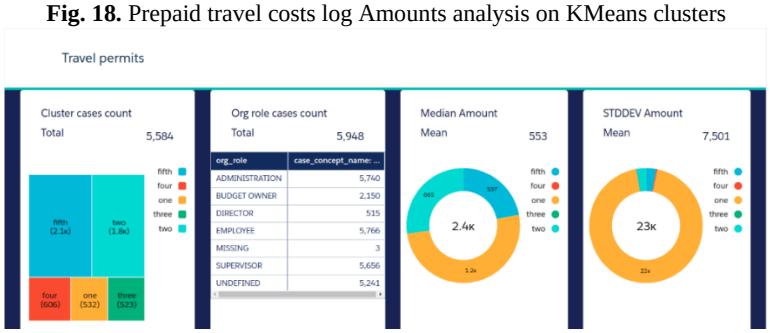

Fig. 19. Travel permits log Amounts analysis on KMeans clusters

The above figures show from left to right: the size and proportion of the resulting clusters, distribution of the amount of traces by responsible roles, where UNDEFINED is responsible for the submission event to the system, the mean of Requested Amount depending on the cluster, the average, the standard deviation of Requested Amount depending on the cluster, their average. There is an obvious difference for the resulting clusters, and it is relevant to use case perspective together with organizational perspective. Event paths vary greatly, and processes have been simplified as much as possible using the heuristics miner algorithm.

Having such structures for each cluster, it became possible to come to the following results: in prepaid travel costs log there are clusters in which the process necessarily goes through the stage of permit, and there are clusters where it does not. There is a difference between the clusters in the passage of the declaration through the administration, supervisor, director. The latter sometimes is not completely separated, but there is a certain pattern of separation. This construction led to the following results in explaining the diversity in Figures 17 to 19:

- Cluster *"three"* in DD: Declaration was firstly rejected, then payment was handled
- Cluster *"three"* in ID: Permit passed, however declaration was rejected, sometimes even trip didn't happen, cluster "two": Permit events include DIRECTOR role, cluster "three" operation was cancelled
- Cluster *"one"* in PTC: Permit events include BUDGET OWNER AND

DIRECTOR role, in cluster *"two"* does not include DIRECTOR, in clusters *"three"* and *"four"* no Permit events at all, in cluster *"four"* operation was cancelled

- Cluster *"one"* in TP: Permit events include BUDGET OWNER AND DIRECTOR role, Declaration events include BUDGET OWNER, in clusters *"three"* and *"four"* operation was cancelled, in cluster "fifth" only ADMINISTRATION and SUPERVISOR are present
- Cluster *"two" in* RfP: all were rejected, in clusters *"three"* and *"four"* were rejected, then payment was handled though

When cancellations occur, amounts are always different and usually small. The situation with RfP is different, the cancellations there are very similar to the second cluster in the PTC. Cases requiring the Director's decision are, on average, costly, and in this case the standard deviation is often very high. Perhaps, the greatest cost required the attention of the director. On the other hand, this is not observed for the PTC (so obvious, concerning ID), that can be attributed to an unsuccessful coincidence with other events, since in TP, in addition to the Permit event, there are Request for payment event, Declaration event. If DIRECTOR or BUDGET OWNER is in TP log, it is not always in other logs. So, such clustering does not consider absolutely all the details.

2.7 Bottlenecks analysis

To understand the data structure a comprehensive analysis of the available information of the traces was conducted, the case perspective approach was used.

To identify bottlenecks in the processes, it is necessary to have an overview of the timing of the steps in the particular process.[5] It can be done using Disco – a process mining toolkit that uses log files to filter and select the needed process parameters. Firstly, the DD log was analyzed. In order to find the bottlenecks, it is necessary to filter out only those cases that duration exceeds the average duration of cases in this process significantly, that is 7.3 days.[4] For this purpose, there was created a filter, based on the case duration.

As the result, the process graph was obtained, it includes 16% of the initial number of cases. A time representation of the resulting graph is in Fig.20.

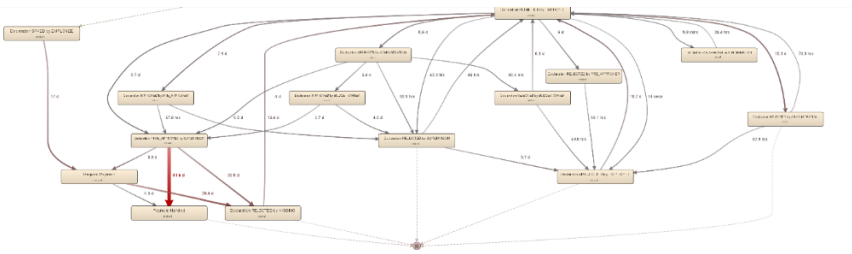

Fig. 20. Process graph for Domestic Declarations

The Fig.20 shows that the longest transition is from Declaration FINAL_APPROVED by SUPERVISOR to Payment handled steps. However, the overall picture of the process shows that only 7 cases go this way. Therefore, accelerating this transition will not bring much improvement for the process in common.

At the next stage, it was supposed that the bottlenecks in the process can be sets of paths leading from point A to point B. To test this assumption, the analysis of the broader picture of the process was undertaken. As the result, bottlenecks were identified in the transitions between Declaration SUBMITTED by EMPLOYEE and Declaration REJECTED by EMPLOYEE or Declaration SUBMITTED by EMPLOYEE and Declaration REJECTED by ADMINISTRATION. The number of cases passing along these two paths is about 10% of the total number of cases, so speeding up the transitions between these steps can significantly accelerate the whole process.

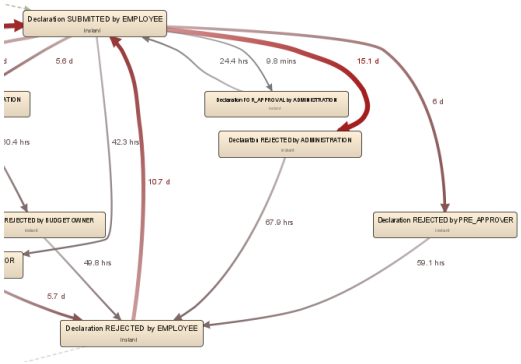

Fig.21. Highlighted bottlenecks

Similarly, the ID log was analyzed. 19% of the total number of cases were filtered out and detailed analysis revealed a problem in the transition from Payment Handled to Start Trip steps. About 3% of the total number of cases are involved in this transition, with an average duration of 51.2 weeks, which is an order of magnitude longer than the duration of other transitions in the process. This delay means that the payment is not always made immediately before the employee's trip, namely, in 3% of cases, it is made several months before the trip. In a less detailed process analysis, no significant bottlenecks were found that could be improved to improve the overall process.

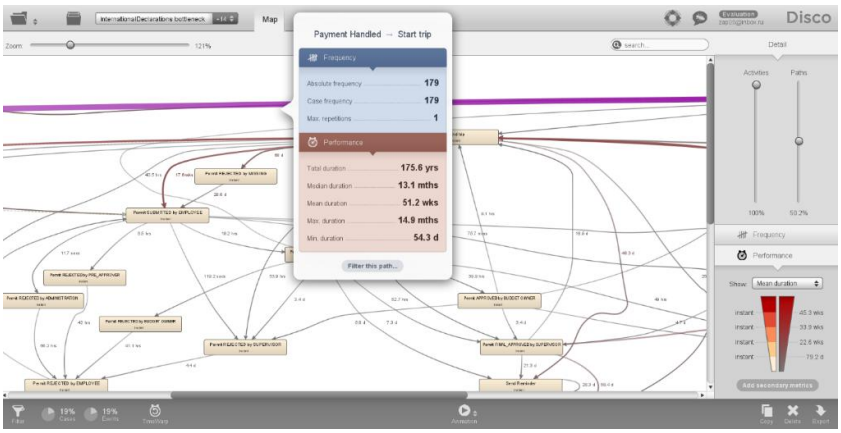

Fig. 22. Process graph for ID

A detailed analysis of the PT log revealed that the bottleneck of this process is the Request for Payment SUBMITTED by EMPLOYEE event. The transitions between this event and Permit FINAL_APPROVED by SUPERVISOR, Permit FINAL_APPROVED by DIRECTOR events are 7% and 16% of the total number of cases respectively, and the average duration of these transitions in filtered cases is up to 2 times longer than in cases before filtering. Moreover, the average transition time from Permit FINAL_APPROVED by DIRECTOR to Request for Payment SUBMITTED by EMPLOYEE is almost equivalent to the average execution time of all filtered cases. This suggests that improving this transition can significantly improve the speed of the entire process. Less detailed analysis did not reveal additional bottlenecks in this process.

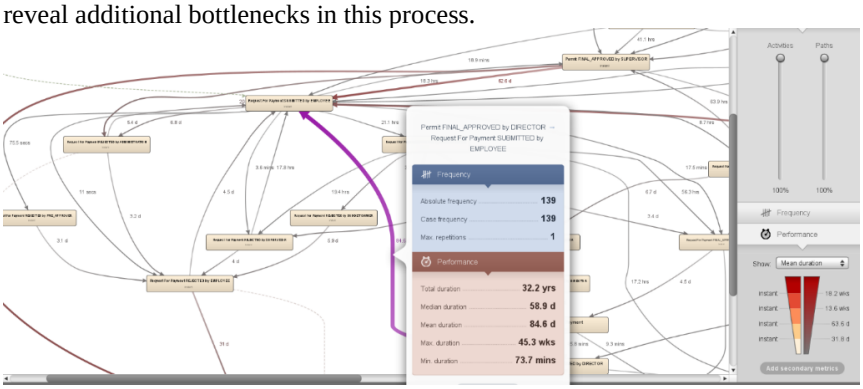

Fig. 23. Process graph for Prepaid Travel Cost.

The Request for Payment log was analyzed. No bottlenecks were found in this log for both types of analysis. Extreme duration cases are associated with anomalies and do not provide information for improving the overall process. Separately, the transition between the Request for Payment FINAL_APPROVED by SUPERVISOR and Request Payment events can be highlighted. From time to time,

a large number of cases accumulate there. This conclusion is based on simulating the process using Disco toolkit. Perhaps, on this step there are not enough resources to process all incoming requests in time. The problem is confirmed in Fig. 24.

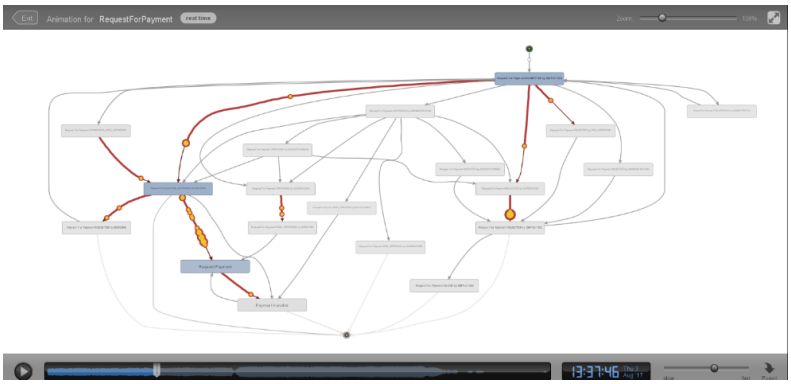

Fig. 24. Process flow problem in Request for Payment

During the TP log analysis from both aspects no bottlenecks were found in the log. Extreme duration cases are associated with anomalies and do not provide information for improving the overall process. In general, it can be seen that travel permits are granted much earlier than the start of the trip, namely, 1-2 months before. Possibly, requiring employees to submit requests no earlier than 1 month before travel would reduce the likelihood that the approval stage would be overloaded with unreasonably early applied requests.

2.8 Declarations allocation among projects

38% of cases in the IDs dated from 2018 belong to the project marked as "UNKNOWN". It means that almost half of the declarations are not project-related. For the remaining 62% cases the distribution of the number of declarations by project is as follows:

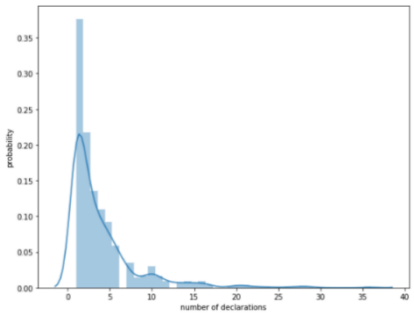

Fig. 25. The distplot for international declarations allocation on projects

Most often there are projects with 1-2 declarations, much less often cases of 3+ declarations per project. However, there are projects with 20-36 travel declarations. Within the travel permits log 43.3% cases similarly belong to the project "UNKNOWN". Considering the remaining 56.7% cases, the distribution of the number of declarations by project is as follows:

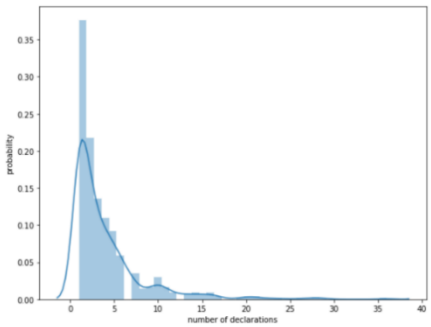

Fig. 26. The distplot for domestic declarations allocation on projects

The graph shows an absolutely similar trend, that has been found for the international declarations.

Similar to the previous event logs, the prepaid travel costs log has 40.5% cases belonging to the project "UNKNOWN", and the remaining cases represent the following distribution:

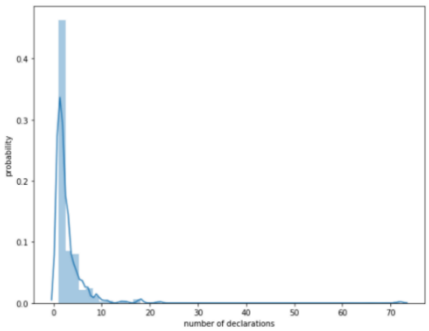

Fig. 27. The distplot for prepaid travel costs declarations allocation on projects

In general, the trend continues, with most projects having one case, but less than two or three cases per project.

Thus, one labeled project most often has 1-2 international declarations, 1-2 travel permits and 1 request for prepaid costs.

2.9 Travel declarations rejection and further resubmission

The ID and TP logs were analyzed. In order to find out which declarations were

rejected due to a long submission after the end of the trip, it is necessary to filter out cases in which the gap between the “end trip” event and the “declaration saved by employee” event is 2 or more months. This result can be achieved by applying the Follower filter in Disco Toolkit.

After filtering the ID log, it turned out that 104 declarations were submitted 2 months or more after the end of the trip, of which 21 were rejected and 100 were approved. Thus, this log contains 4 declarations, that were ultimately rejected due to late submission.

As for the differences between departments, the department with id 65454 takes 20% of “late” declarations and 50% of finally rejected declarations. The department with id 65458 takes 50% of “late” declarations, however, all submitted declarations were accepted. Some departments do not have “late” declarations, submitted after the deadline.

The TP log was filtered in the same way. In this log, 353 declarations were submitted later than two months after the end of the trip. This is about 4% of the total number of cases in the log. Of these 353 declarations, 134 were eventually rejected, that is 38% of the «later» declarations. The Organization Entity field is empty almost for all of these declarations, so it is difficult to talk about the difference between departments.

2.10 Analysis of travel declaratons corrections

Firstly, the DomesticDeclarations log was analyzed. Here 1019 declarations have been corrected. Of these, 38 were rejected even after corrections.

As for international trips (ID log), 21% of declarations have been corrected. Of these, only one declaration was rejected after the correction, while the others were approved.

In the PTC log, 8% of declarations have been corrected. Of these, 6 were rejected even after corrections.

The RfP log also contains 8% of declarations that have been firstly rejected. After resubmission, 31 of them failed to be approved again.

The TP log has a correction cycle (loop) in 21% of cases. However, all corrected declarations have been successfully approved.

2.11 Cases that were not approved by budget holders and rerouted to supervisers

The TP log was analyzed. It is assumed that the declaration is automatically rerouted to the supervisor 7 days after any of the approval or the rejection judgement on the previous steps except the budget owner step. In order to provide such filtering, two Follower filters must be used in Disco Toolkit. In the TP log, 5% of cases bypass the budget owner judgement. Moreover, all these cases were approved.

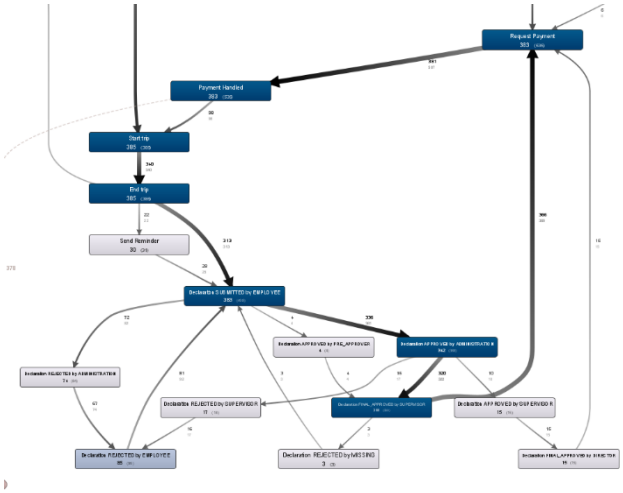

Fig. 28. A part of the graph for the PermitLog log

The ID log was analyzed in the same way. Here, 354 cases, or 6% of the total, were automatically rerouted to the supervisor. As in the previous log, all declarations were approved.

The PTC log was analyzed. In this log only 2% of declarations were not approved by budget owner in time (7 days).

In the RfP log only 304 cases meet the required conditions, that is 4% of the total number of cases. There is a logic - in the logs that are directly related to payments, declarations cannot be finally approved without a budget owner approval step.

Most of the cases covered by this question are found in the DD log. There are 7325 such declarations, that is 69% of the total number of domestic declarations.

3 Conclustion

Begining from the high level and divig deeper in details we provided the step by step analysis for the provided logged data from the Eindhoven University of Technology (TU/e) to be able to stress bottlenecks and deviations in the current process of travel arranging and costs reimbursement. According to the gained results, the logging system, that now exists in the university *allows* its users to follow wrong process flows and allow submitting the declaration papers even with the time and structural mistakes. Such an approach leads to double cost payments, wrong tracks of document submission for TU/e employees who followed track for non-TU/e ones by mistakes, extremely long time processing of documents, cycles in rejections and further resubmissions, sendings of obligatory papers after the

expiration date of the reimbursement, realization of international trips without the required permissions and simply confused process flow step following. All the listed cases lead to creation of the bottlenecks due to extra load on staff members. There should be a complex treatment applied for such situation:

- Technical improvements: appliance of the restrictions for disabling the ability to violate the stated process flow, including not only step subsequence, but also timing limints
- Creation an infographics using decision tree idea according to the needed action confirmation

The clarification of the papers submission with the help of system restrictions will reduce the huge amount of deviations that now exists in TP, ID, PTC processes and will help to eliminate small outliers in DD and RfP processes.

References

1. David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, Qingmai Wang, Anomaly Detection in Online Social Networks. In: International Journal of Innovative Science and Research Technology, ISSN No:-2456-2165, vol. 3.
2. Ferreira, Diogo R.: A Primer on Process Mining. Practical Skills with Python ang Graphviz
3. Springer Verlag , van der Aalst W.M.P.: Process Mining: Data Science in Action (2016)
4. Reinkemeyer, Lars. (2020). Process Mining in Action. 10.1007/978-3-030-40172-6.
5. Syring, Anja & Tax, Niek & Aalst, Wil. (2020). Evaluating Conformance Measures in Process Mining using Conformance Propositions (Extended version)
6. Weijters, A. & Aalst, Wil & Medeiros, Alves. (2006). Process Mining with the Heuristics Miner-algorithm.
7. Dogan, Onur & Öztayşi, Başar & Fernandez-Llatas, Carlos. (2019). Segmentation of indoor customer paths using intuitionistic fuzzy clustering: Process mining visualization. Journal of Intelligent and Fuzzy Systems. 38. 675-684. 10.3233/JIFS-179440.
8. Minseok Song, Christion W. Gunther, Wil Van der Aalst. (2008). Trace clustering in process mining.