Process Mining: Diving Into Travel Data

Mikhail Poruchikov, Pavel Katkov

Sberbank of Russia, Samara, Russia
mporuchikov@gmail.com, katkov.p.i@gmail.com

Abstract. Every day companies all over the world try to improve efficiency of their business processes. In this report we analyze travel data supplied by Eindhoven University of Technology. Due to specific nature of data, represented as event logs, we used process mining, a technique that allows to discover real behavior of a system, as well as more traditional statistical analysis. To build as-is model of a process we used PM4Py, open source process mining platform. During research many outliers and bottlenecks were found. In this report we describe results of research and make suggestions that allow to make the process of managing travels at Eindhoven University of Technology more efficient.

Keywords: Business Process Intelligence, Process Mining, Process Discovery, Event Logs, PM4Py.

1 Introduction

BPI Challenge is an annual competition in process analysis. In year 2020 the challenge's goal is to explore real-life event logs of processes of Eindhoven University of Technology. Process owners are interested to get answers to many questions like "what is the throughput of a travel declaration from submission (or closing) to paying?", "is there are difference in throughput between national and international trips?". The rest of the questions can be found at the event website [1]. Event logs are represented as log files in XES format [2].

Process mining is a set of techniques for business process analysis. The most important approach in process mining is process discovery, a technique that allows to create process model from event logs.

A lot of specific tools for process mining are available at the moment, for instance, Python library PM4Py [3].

2 Data Exploration

2.1 Quick Look at Data

An event log for domestic declarations is a 56437x10 table. It contains information about 10500 activities. The number of activities varies from 1 to 24 for each travel declaration. Fig.1 shows that most domestic declarations consist of 5 activities.

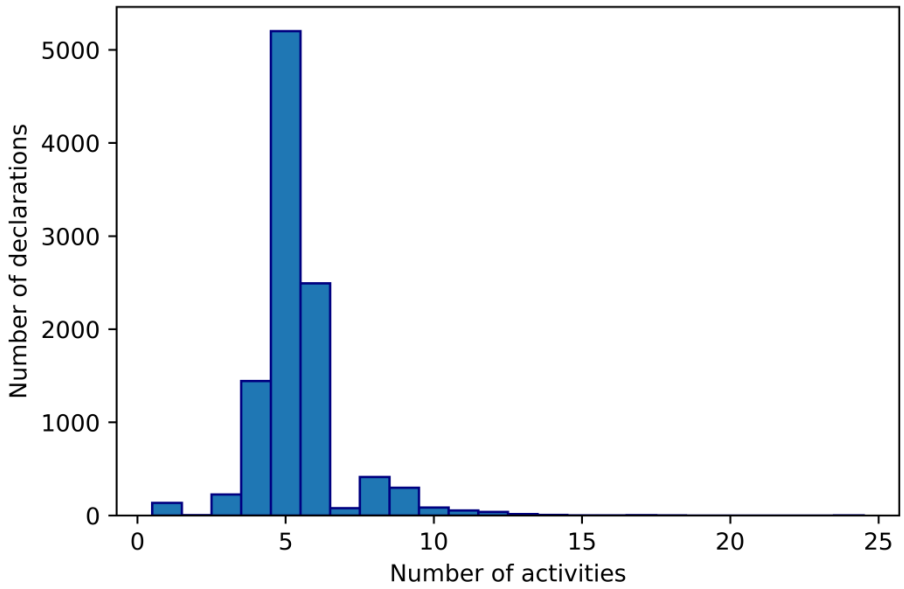

Fig.1. Histogram of number of activities in domestic declarations.

134 declarations contain just one activity.

An event log for international declarations is a 72151x23 table. It contains information about 6449 activities. The number of activities varies from 3 to 27 for each travel declaration. Fig.2 shows that most international declarations consist of 10 activities.

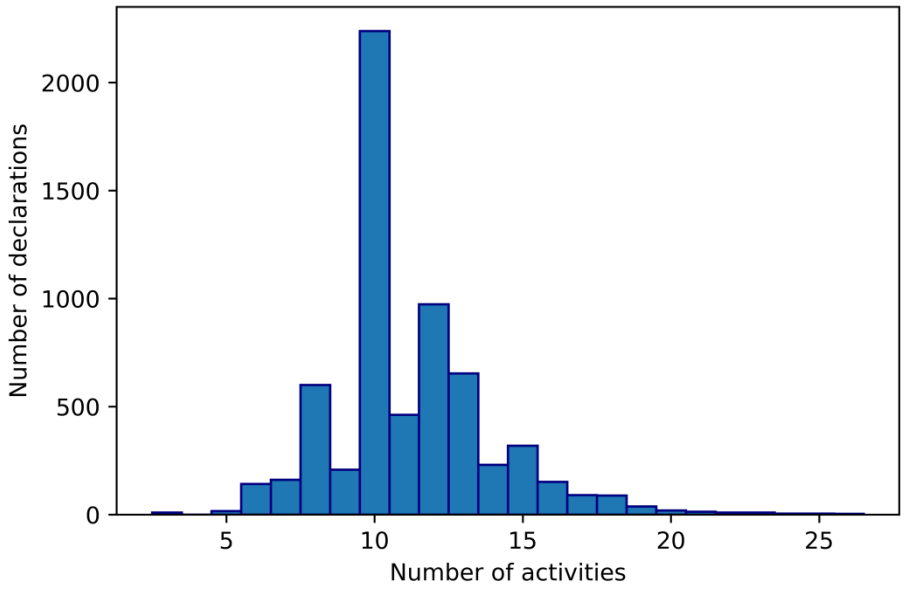

Fig.2. Histogram of number of activities in international declarations.

2.2 Throughput Comparison

To compare throughput of processes for domestic and international we calculated several statistics (Table 1).

Table 1. Statistics of throughput for domestic and international declarations.

Statistics	Domestic travels	International travels
Mean duration, days	10336	86.452
Standard deviation, days	11.674	78.347
Min. duration, days	0.001	6.730
Max. duration, days	469.236	742.000
Median duration, days	7.348	66.042

Distribution of throughput for domestic and international declaration is shown as box plots at Fig.3.

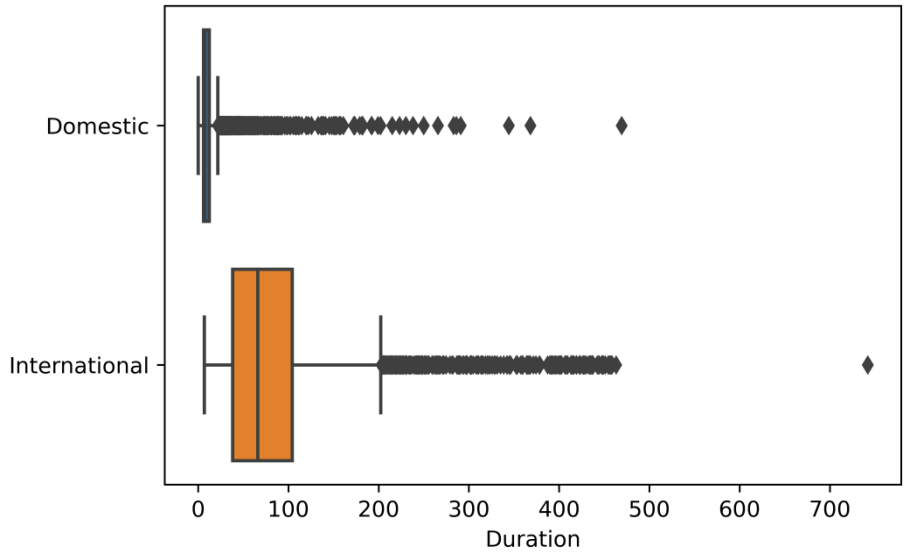

Fig.3. Box plot of throughput for domestic and international declarations.

According to the box plot, throughput for domestic and international declarations differs.

To provide strict proof for the previous statement statistical test should be performed, for instance, Welch's t-test. It does not require samples to have equal size or similar variance. It is implemented in scipy Python library [4]. Null hypothesis H0 is that samples have identical average values. Alternative hypothesis is that samples have different average values. With significance level of 0.05 Welch's test shows t-statistic of -75.5 and p-value of 0.0. Thus null hypothesis is rejected in favor of the alternative one. Therefore there is statistically significant difference between throughput of domestic and international declarations.

2.3 Process drift

We also to find a drift in processes for both domestic and international declarations. By drift we mean changing of case duration in time. Fig. 4 shows duration for each case started in particular time. Every case is denoted by separate dot. Trend in case duration is denoted by the line.

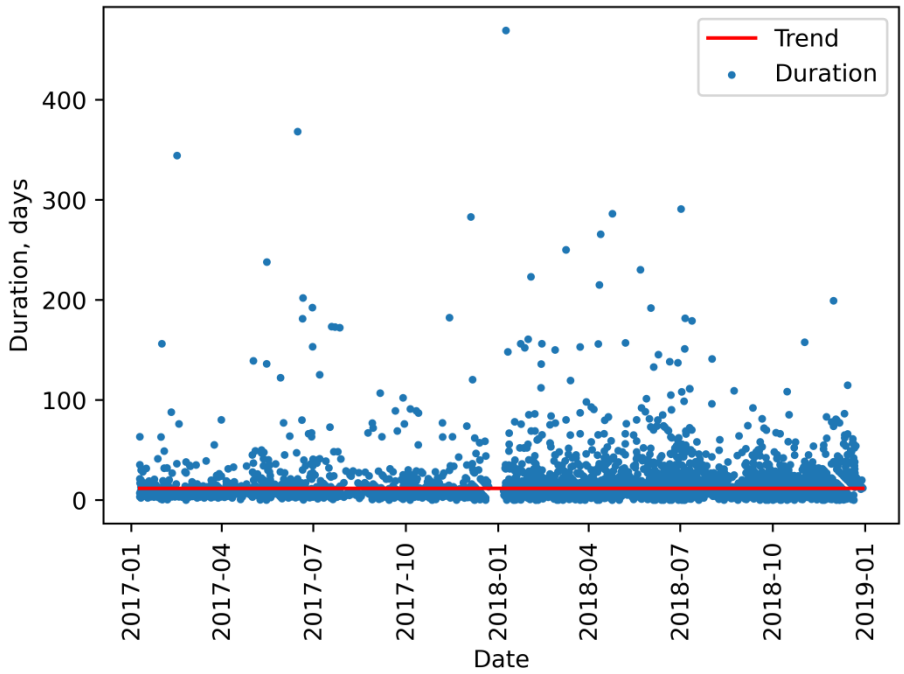

Fig.4. Trend in throughput for domestic declarations.

According to calculations, line slope is very close to zero, thus the line is horizontal. We can conclude there is no drift in process of domestic declarations.

However the similar research for international declarations shows that case durations tend to decrease (Fig.5).

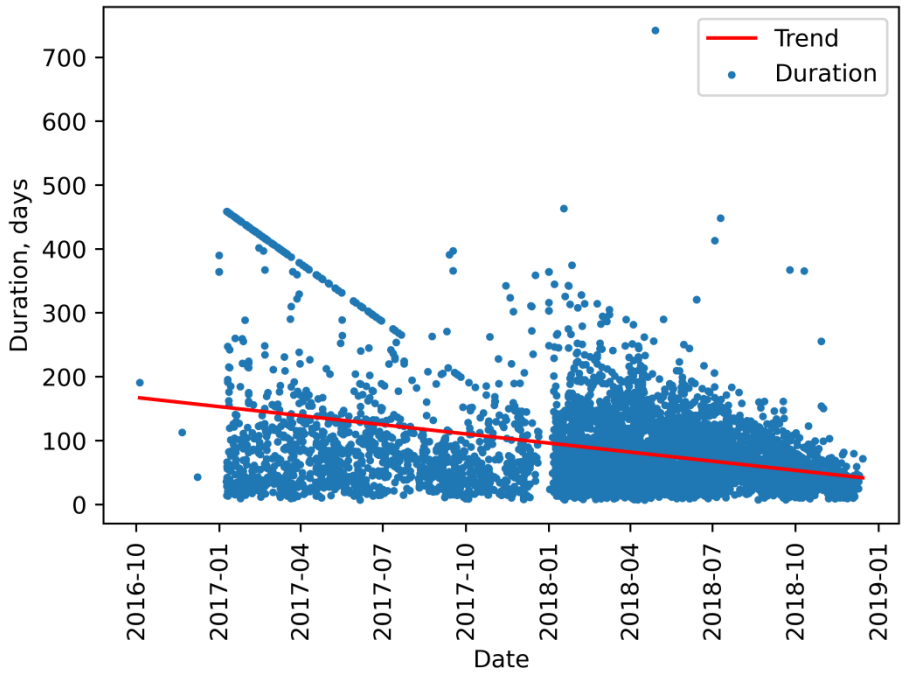

Fig.5. Trend in throughput for international declarations.

Moreover, there are unexplainable linear and triangular patterns on the diagram.

2.4 Process Discovery

To visualize the process we used Directly-Follows Graphs implemented in PM4Py library (Fig. 6).

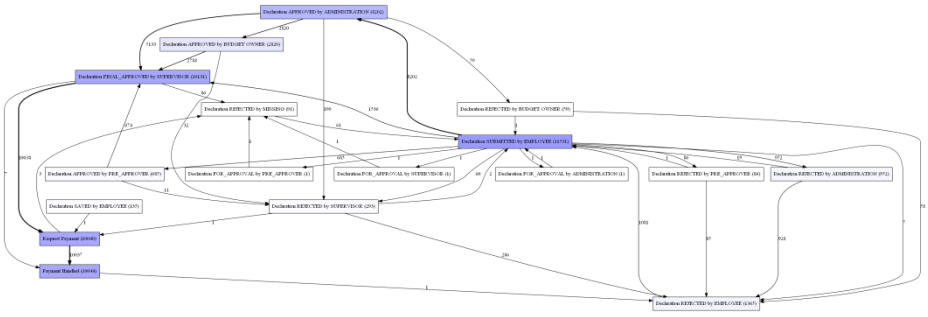

Fig.6. Directly-follow graph for domestic declarations.

3 Results

During our research we came to the following conclusions:

1. The processes for domestic and international declarations have significant difference.
2. Throughput of domestic declarations is much less than of international declarations.
3. There is a trend of drift decreasing case duration in process of international declarations.

To conduct research we used several Python libraries. The most important of them are scipy and PM4Py. Particular Jupyter notebook is available at github repository https://github.com/mporuchikov/BPIC-2020.

References

1. BPI Challenge – Process Mining Conference 2020, https://icpmconference.org/2020/bpi-challenge, last accessed 2020/08/22.
2. van Dongen, Boudewijn (2020): BPI Challenge 2020. 4TU.ResearchData. Collection. https://doi.org/10.4121/uuid:52fb97d4-4588-43c9-9d04-3604d4613b51.
3. PM4Py - Process Mining for Python, https://pm4py.fit.fraunhofer.de, last accessed 2020/08/22.
4. scipy.stats.ttest_ind – SciPy v1.5.2 Reference Guide, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html, last accessed 2020/08/22.