Using Sequential Pattern Mining and Social Network Analysis to Identify Similarities, Differences and Evolving Behaviour in Event Logs

Scott Buffett and Bruno Emond

National Research Council Canada
{scott.buffett,bruno.emond}@nrc-cnrc.gc.ca

Abstract. This paper reports on findings resulting from our research conducted for the 2015 Business Process Intelligence Challenge (BPIC), an annual competition in which participants are tasked with conducting process mining-related analysis on a real-life dataset. This year's data was provided by 5 Dutch municipalities, and contained activity pertaining to their building permit application process. Questions regarding the underlying process posed to the participants centered around identifying differences in control flow among municipalities, the responsible factors for longer processing times, and differences in the various roles of employees involved, to name a few. Our approach to addressing these questions involved the application of methods from the field of sequential pattern mining, an area of research that identifies frequently occurring sequences of events in potentially large databases. In particular, sequence classification is used to identify statistically significant differences in control flow among municipalities. Also, value-based sequential pattern mining is used to identify patterns in control flow that are linked to 1) high/low throughput times, in order to identify similarities and differences among the five municipalities, and 2) earlier/later process instances, in order to examine how municipalities' underlying process may have changed over time, and how these changes may be similar among municipalities. We also employ traditional methods from the field of process mining to shed light on the the social network-related aspects of the data, such as how the roles of employees differ among municipalities in terms of task similarity.

1 Introduction

This paper reports on findings resulting from our research conducted for the 2015 Business Process Intelligence Challenge (BPIC). The BPIC is an annual competition in which participants are tasked with conducting process mining-related analysis on a real-life dataset. This year's data was provided by 5 Dutch municipalities, and contained activity pertaining to the process of building permit applications. Each record in the dataset described an action taken by an employee in processing a particular application, and contained such informa-

tion as the case ID, employee ID, action taken and a timestamp, among others. A number of questions about the underlying process were then posed to the participants, asking for insight into the differences in control flow among municipalities, the responsible factors for longer processing times, and differences in various roles of the employees involved, to name a few. This paper describes our findings for a number of these questions, and constitutes our entry into the competition.

Our approach to this competition consisted of methods from the field of sequential pattern mining, which is an area of research that is very relevant and complementary to existing techniques from process mining. The goal of sequential pattern mining is to identify temporal patterns, or *sequences*, of activity in the data that may be of interest. Such potential interest could be due the pattern's high frequency, its ability to explain various phenomena, or its potential predictive ability. In this paper, we utilize sequential pattern mining to identify patterns that are commonly associated with certain aspects of the data from a number of different dimensions, such as data related to instances that occur earlier or later in time, or those that take longer to process. We also utilize traditional process mining techniques and technology to address the social network-centric questions posed in the challenge

2 Competition Guidelines and Methods

2.1 The 2015 Business Process Intelligence Challenge

The 2015 Business Process Intelligence Challenge (BPIC'15) [1] involved the analysis of data generated from the building permit application process in 5 (unnamed) municipalities located in the Netherlands, each provided in a separate dataset. Each record described a single step taken by an employee in the process, and consisted of the case ID, the activity performed (given by its ID as well as its description in both English and Dutch), the time the action was completed, the ID of the employee responsible for the action as well as that for a monitoring employee, the cost associated with the case, and many other fields.

The questions posed by the process owner were given as follows:

1. What are the roles of the people involved in the various stages of the process and how do these roles differ across municipalities?
2. What are the possible points for improvement on the organizational structure for each of the municipalities?
3. The employees of two of the five municipalities have physically moved into the same location recently. Did this lead to a change in the processes and if so, what is different?
4. Some of the procedures will be outsourced from 2018, i.e. they will be removed from the process and the applicant needs to have these activities performed by an external party before submitting the application. What will be the effect of this on the organizational structures in the five municipalities?

5. Where are differences in throughput times between the municipalities and how can these be explained?
6. What are the differences in control flow between the municipalities?

The following sections will present our answers to these six questions. Questions number one, two and four will be answered using social network analysis, while questions number three, five and six will rely on sequential pattern analysis.

3 Social Network Analysis

This section aims at answering questions 1, 2, and 4 of the Business Process Intelligence Challenge. The answers are based on both process model discovery, and similar tasks social network analysis. Given that the logs did not contain information about roles played by staff in the organizations, similar tasks metric appeared especially relevant to infer these organizational roles. Similar tasks assume that people doing similar things have stronger relationships than people doing different things. The similar task metric is not linked to the specific flows of hand-over work, subcontracting, or working together, which are more centered on sequences of activities within a case.

3.1 Methodology

The analysis conducted prior to answering questions 1, 2, and 4 consisted of the following steps:

1. Data contained in the CSV files was processed with a Common Lisp program to produce XES log files. During this process, the event activity codes were changed from labels, such as '01_BB_010', to the English descriptions for those codes (ex. 'register submission date request'). The data for all municipalities contained 5,647 cases, and 262,628 events.
2. Only frequent starting, ending, and middle events were kept in each of the municipalities' respective logs. ProM's module 'Filter Log using Simple Heuristics' was used to execute this selection with a 80% selection threshold.
3. Process models for each of the municipalities were generated using the inductive miner algorithm (ProM's 'Visual Inductive Miner'), using only the 90% most frequent sequences. Figures 1, 2, 3, 4, and 5 presents the process models, larger versions are also available in an appendix. The activities identified in these process models were used later as the basis to make resource-activity matrices.
4. Social network analysis was performed using a similar task metric on resource degree correlation coefficients. The degree of a node being the number of nodes that are connected to it (both in and out connections). The social networl analysis graph layout used the Fruchterman-Reingold force-directed algorithm, because it created more space between node clusters to facilitate visual inspection or social networks.

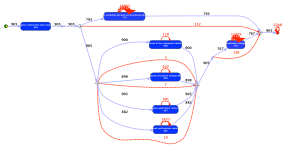

Fig. 1. PM1 - See Figure 22.

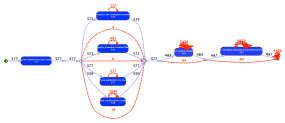

Fig. 2. PM2 - See Figure 24.

Fig. 3. PM3 - See Figure 26.

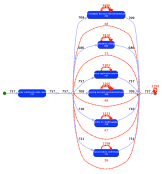

Fig. 4. PM4 - See Figure 28.

Fig. 5. PM5 - See Figure 30.

5. The visual inspection of social networks allowed to identify clusters of resources working on similar tasks. Figures 6, 7, 8, 9, and 10 present the social networks of each municipality. Larger version of the graphs are available in an appendix. Clusters are labeled by the letters A, B, and C. In addition, high frequency task executors are identified by blue dots. The notion of high and low frequency task executors is explained below.

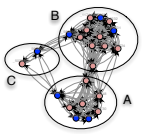

Fig. 6. SN1 - See Figure 23.

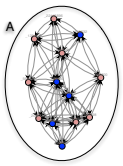

Fig. 7. SN2 - See Figure 25.

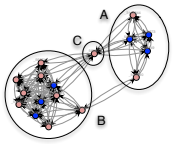

Fig. 8. SN3 - See Figure 27.

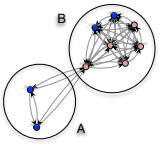

Fig. 9. SN4 - See Figure 29.

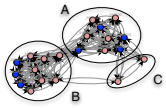

Fig. 10. SN5 - See Figure 31.

6. Finally, resource-activity matrices were created for each municipality using role clusters identified by the social analysis, and activities referred by the discovered process models (step #3 above). The resource-activity matrices will be presented in the next section. People performing individual tasks in a proportion of over 10% of the total number of task executions were labeled as high frequency task executors for the task at hand, and others labeled as low frequency task executors. The purpose of this distinction was to facilitate linking tasks to cluster roles.

3.2 Question 1: Roles in processes

Question 1: *What are the roles of the people involved in the various stages of the process and how do these roles differ across municipalities?*

Roles can be characterized as a set of tasks people perform. Using discovered process models, and social network analysis of similar tasks, we constructed resource-activity matrices for each of the municipalities which allows to describe the clusters in terms of task descriptions.

Figure 11 shows the resource-activity matrix for the municipality 1 using the three clusters identified in the social network of Figure 6. The role of cluster A involves the following activities: register submission date request, phase application received, send confirmation receipt, enter senddate decision environmental permit, and forward to the competent authority. The other main role (cluster B) involves procedure change, and regular procedure without MER. The cluster C is particular, as it is associated to only one activity, enter senddate decision environmental permit.

No resource-matrix was produced for the municipality 2. The social network analysis (Figure 7) did not reveal any obvious clusters, and it appears that all resources from this municipality perform similar tasks.

Figure 12 shows the resource-activity matrix for the municipality 3 using the three clusters identified in the social network of Figure 8. The role of cluster A involves the following activities: register submission date request, phase application received, send confirmation receipt, and enter senddate decision environmental permit. The other main role (cluster B) involves procedure change, and grounds for refusal. The cluster C is distinctive by the low frequency its activities .

Figure 13 shows the resource-activity matrix for the municipality 4 using the two clusters identified in the social network of Figure 9. The role of cluster A involves the following activities: register submission date request, phase

Activities	A Sum	%	B Sum	%	C Sum	%	Total Sum
register submission date request	**1214**	**66.27%**	**585**	**31.93%**	**33**	**1.80%**	**1832**
High Frequencies	976	65.55%	513	34.45%		0.00%	1489
Low Frequencies	238	69.39%	72	20.99%	33	9.62%	343
phase application received	**1188**	**64.85%**	**612**	**33.41%**	**32**	**1.75%**	**1832**
High Frequencies	951	65.41%	503	34.59%		0.00%	1454
Low Frequencies	237	62.70%	109	28.84%	32	8.47%	378
send confirmation receipt	**2024**	**62.92%**	**1151**	**35.78%**	**42**	**1.31%**	**3217**
High Frequencies	1557	63.04%	913	36.96%		0.00%	2470
Low Frequencies	467	62.52%	238	31.86%	42	5.62%	747
enter senddate decision environmental permit	**1474**	**58.42%**	**232**	**9.20%**	**817**	**32.38%**	**2523**
High Frequencies	1292	59.00%	129	5.89%	769	35.11%	2190
Low Frequencies	182	54.65%	103	30.93%	48	14.41%	333
forward to the competent authority	**917**	**50.41%**	**879**	**48.32%**	**23**	**1.26%**	**1819**
High Frequencies	602	54.93%	494	45.07%		0.00%	1096
Low Frequencies	315	43.57%	385	53.25%	23	3.18%	723
procedure change	**58**	**1.93%**	**2889**	**96.36%**	**51**	**1.70%**	**2998**
High Frequencies		0.00%	2627	100.00%		0.00%	2627
Low Frequencies	58	15.63%	262	70.62%	51	13.75%	371
regular procedure without MER	**781**	**43.01%**	**1010**	**55.62%**	**25**	**1.38%**	**1816**
High Frequencies	502	37.46%	838	62.54%		0.00%	1340
Low Frequencies	279	58.61%	172	36.13%	25	5.25%	476
Grand Total	**7656**	**47.74%**	**7358**	**45.88%**	**1023**	**6.38%**	**16037**

Fig. 11. Resource-activity matrix using process model PM1 (Figure 1) and social network SN1 (Figure 6).

Activities	A Sum	%	B Sum	%	C Sum	%	Total Sur
register submission date request	**1050**	**97.04%**	**31**	**2.87%**	**1**	**0.09%**	**1082**
High Frequencies	1049	100.00%		0.00%		0.00%	1049
Low Frequencies	1	3.03%	31	93.94%	1	3.03%	33
phase application received	**1041**	**96.21%**	**40**	**3.70%**	**1**	**0.09%**	**1082**
High Frequencies	1041	100.00%		0.00%		0.00%	1041
Low Frequencies		0.00%	40	97.56%	1	2.44%	41
enter senddate decision environmental permit	**1444**	**96.20%**	**57**	**3.80%**		**0.00%**	**1501**
High Frequencies	1442	100.00%		0.00%		0.00%	1442
Low Frequencies	2	3.39%	57	96.61%		0.00%	59
send confirmation receipt	**1777**	**91.17%**	**170**	**8.72%**	**2**	**0.10%**	**1949**
High Frequencies	1777	100.00%		0.00%		0.00%	1777
Low Frequencies		0.00%	170	98.84%	2	1.16%	172
grounds for refusal	**259**	**24.98%**	**778**	**75.02%**		**0.00%**	**1037**
High Frequencies	105	13.71%	661	86.29%		0.00%	766
Low Frequencies	154	56.83%	117	43.17%		0.00%	271
procedure change	**404**	**22.86%**	**1363**	**77.14%**		**0.00%**	**1767**
High Frequencies		0.00%	1281	100.00%		0.00%	1281
Low Frequencies	404	83.13%	82	16.87%		0.00%	486
Grand Total	**5975**	**70.98%**	**2439**	**28.97%**	**4**	**0.05%**	**8418**

Fig. 12. Resource-activity matrix using process model PM3 (Figure 3) and social network SN3 (Figure 8).

application received, send confirmation receipt, and enter senddate decision environmental permit. The other main role (cluster B) involves procedure change, and article 34 WABO applies. No cluster C was identified by the social analysis.

Figure 14 shows the resource-activity matrix for the municipality 5 using the two clusters identified in the social network of Figure 10. The municipality 1 and 5 are very similar in terms of the distribution of activities in roles. Like the

	A		B		Total Sum
Activities	Sum	%	Sum	%	
▼ enter senddate decision environmental permit	997	91.72%	90	8.28%	1087
High Frequencies	997	100.00%		0.00%	997
Low Frequencies		0.00%	90	100.00%	90
▼ register submission date request	677	88.85%	85	11.15%	762
High Frequencies	677	100.00%		0.00%	677
Low Frequencies		0.00%	85	100.00%	85
▼ phase application received	670	87.93%	92	12.07%	762
High Frequencies	670	100.00%		0.00%	670
Low Frequencies		0.00%	92	100.00%	92
▼ send procedure confirmation	623	84.19%	117	15.81%	740
High Frequencies	623	100.00%		0.00%	623
Low Frequencies		0.00%	117	100.00%	117
▼ article 34 WABO applies	147	20.62%	566	79.38%	713
High Frequencies	101	17.38%	480	82.62%	581
Low Frequencies	46	34.85%	86	65.15%	132
▼ procedure change	489	37.67%	809	62.33%	1298
High Frequencies	489	42.19%	670	57.81%	1159
Low Frequencies		0.00%	139	100.00%	139
Grand Total	3603	67.20%	1759	32.80%	5362

Fig. 13. Resource-activity matrix using process model PM4 (Figure 4) and social network SN4 (Figure 9).

municipality 1, municipality 5 has the role of cluster A involving the following activities: register submission date request, phase application received, send confirmation receipt, enter senddate decision environmental permit, and forward to the competent authority. Likewise the other main role (cluster B) involves procedure change, and regular procedure without MER. However, contrary to the municipality 1, there is no focus on a simgle activity in cluster C.

In conclusion, the most similar municipalities in regard to the distribution of activities in roles are the municipality 1 and 5. With the exception of municipality 2, which does not seem to have a clear separation of roles, the process mining and social analysis indicate that one role across all the municipalities can be defined by the four activities of 1) register submission date request, 2) phase application received, 3) send confirmation receipt, and 4) enter senddate decision environmental permit, while the other role involves only one activity, procedure change.

3.3 Question 2: Organizational improvements

Question 2: *What are the possible points for improvement on the organizational structure for each of the municipalities?*

To answer this question, we are assuming that the municipalities' goal is to conform to the generic process model given in Figure 15, which was discovered using a combined log for all the municipalities. The main interesting feature of this all municipalities process model is that there is a clear end activity process state.

Figure 16 presents abstract process models for the municipalities using the identified roles in the previous section (cluster of people executing similar tasks)

	A		B		C		Total Sum
Activities	**Sum**	**%**	**Sum**	**%**	**Sum**	**%**	
send confirmation receipt	**1412**	**87.38%**	**170**	**10.52%**	**34**	**2.10%**	**1616**
High Frequencies	1067	100.00%		0.00%		0.00%	1067
Low Frequencies	345	62.84%	170	30.97%	34	6.19%	549
enter senddate decision environmental permit	**1221**	**93.63%**	**37**	**2.84%**	**46**	**3.53%**	**1304**
High Frequencies	1050	100.00%		0.00%		0.00%	1050
Low Frequencies	171	67.32%	37	14.57%	46	18.11%	254
register submission date request	**845**	**90.96%**	**55**	**5.92%**	**29**	**3.12%**	**929**
High Frequencies	670	100.00%		0.00%		0.00%	670
Low Frequencies	175	67.57%	55	21.24%	29	11.20%	259
phase application received	**827**	**89.02%**	**75**	**8.07%**	**27**	**2.91%**	**929**
High Frequencies	652	100.00%		0.00%		0.00%	652
Low Frequencies	175	63.18%	75	27.08%	27	9.75%	277
forward to the competent authority	**614**	**66.67%**	**296**	**32.14%**	**11**	**1.19%**	**921**
High Frequencies	360	100.00%		0.00%		0.00%	360
Low Frequencies	254	45.28%	296	52.76%	11	1.96%	561
procedure change	**27**	**1.76%**	**1503**	**98.11%**	**2**	**0.13%**	**1532**
High Frequencies		0.00%	1391	100.00%		0.00%	1391
Low Frequencies	27	19.15%	112	79.43%	2	1.42%	141
regular procedure without MER	**479**	**52.07%**	**430**	**46.74%**	**11**	**1.20%**	**920**
High Frequencies	260	41.94%	360	58.06%		0.00%	620
Low Frequencies	219	73.00%	70	23.33%	11	3.67%	300
Grand Total	**5425**	**66.56%**	**2566**	**31.48%**	**160**	**1.96%**	**8151**

Fig. 14. Resource-activity matrix using process model PM5 (Figure 5) and social network SN5 (Figure 10).

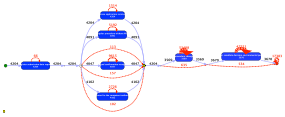

Fig. 15. PM-All - See Figure 21.

instead of the specific process model activities. In those abstract role process models, *A* refers to cluster A, *B* refers to cluster B, *As* refers to the group of people in cluster A executing the starting event 'register submission date request', and *Ae* refers to to the group of people in cluster B executing the ending event 'enter senddate decision environmental permit'.

Given our analysis, and our simple assumption regarding improvement goals, Figure 16 suggests that no improvement is required for the municipalities 2 and 5 as they conform to the target process model.

Municipalities 1 and 3 share a similar process model where the targeted end state is not well sequentially positioned. One area of improvement for the municipality 1 might be to examine the impact that some individuals in the organization can have on the process by performing only one activity at a high frequency, the end state 'enter senddate decision environmental permit', as it is revealed in the municipality 1 resource-activity matrix (Figure 11). It is not clear what could cause the misalignment of the municipality 3 process model, it could be related to the high frequency of the 'grounds for refusals' activity.

Finally, given our level of analysis, the municipality 4 seems to be requiring the most improvement to its processes to conform to the target process model.

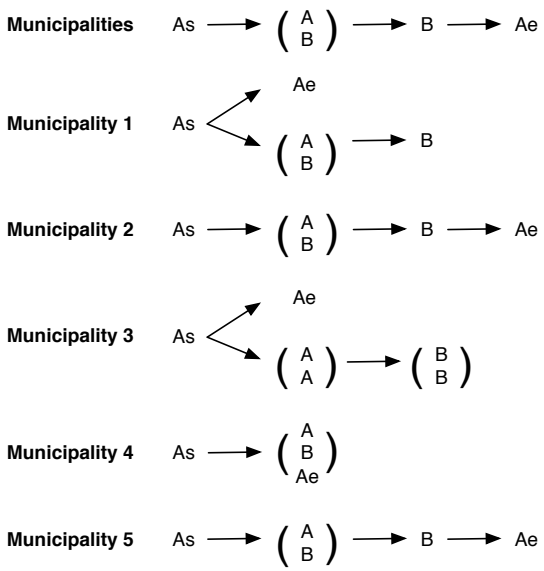

Fig. 16. Abstract role process models using roles (similar tasks), where *A* refers to cluster A, *B* refers to cluster B, *As* refers to the group of people in cluster A executing the starting event 'register submission date request', and *Ae* refers to to the group of people in cluster B executing the ending event 'enter senddate decision environmental permit'

The process model of municipality 4 is not well defined and indicates that the municipality has problems reaching the process end state.

3.4 Question 4: Organizational impact of outsourcing procedures

Question 4: *Some of the procedures will be outsourced from 2018, i.e. they will be removed from the process and the applicant needs to have these activities performed by an external party before submitting the application. What will be the effect of this on the organizational structures in the five municipalities?*

To answer this question, we assess the impact of outsourcing procedures on both municipalities process models, and social networks. From the process model perspective, organizations with well separated activity sequences related to organizational roles should be less impacted by outsourcing procedures than organizations with ill-defined process sequences. For example, the generic model for all municipalities identifies some discrete sequence elements that could be outsourced. If one assumes that the municipalities are to initiate and terminate processes, then they should keep activities related to the role of cluster A, and possibly outsource the activities related to cluster B. A good candidate for outsourcing in the generic model would be the 'procedure change' associated to the cluster B staff role.

From the perspective of the social network of similar tasks, organizations with well separated task roles will suffer less impact from outsourcing, than organizations where roles are not well separated. One important condition to the success of outsourcing procedures is the decoupling of activities, so that a set of activities associated to a role could be outsourced without disturbing the cohesion of activities performed by staffs within the organization.

Table 1. Impact of outsourcing procedures for municipalities.

Municipality	Processes	Social	Overall
1	Medium	Low	Medium
2	Low	High	High
3	Medium	Low	Medium
4	High	Low	High
5	Low	Low	Low

Table 1 summarizes our analysis. The impact of outsourcing on municipalities 1 and 3 should be medium because of the need to slightly improve their processes as it was discussed in the answer to question #2. The impact on municipality 2 should be high in spite of the fact that its processes conform to the target process model, because of the lack of identified clusters of similar tasks (Figure 7). The impact on municipality 4 would be high but for different reasons. The organization has well defined roles, but its processes are not well sequenced, which would have a high impact on the organization processes cohesion. Finally, the municipality 5 would be the best candidate for outsourcing procedures because its processes are well sequenced, and the organization has well defined and separated roles.

4 Sequential Pattern Analysis

Sequential pattern analysis techniques were used to address the three questions that could be approached with temporal data mining-based techniques, namely questions 3, 5 and 6:

Question 3: *The employees of two of the five municipalities have physically moved into the same location recently. Did this lead to a change in the processes and if so, what is different?*

The process owner did not explicitly specify which 2 of the the 5 municipalities moved. Thus our approach to this particular question was to identify key changes in the processes over time for each municipality, and then perform a comparison for each pair to determine if there was any significant increase in similarities. This gave us a hint as to which two may have moved together. The similarities

that became apparent over time then gave us the answer to the question of how the process may have changed.

Question 5: *Where are differences in throughput times between the municipalities and how can these be explained?*

We used sequence mining to identify patterns in the data that tended to be explanatory of high/low throughput times, and presented these patterns.

Question 6: *What are the differences in control flow between the municipalities?*

We utilized feature selection methods from the area of sequence classification to identify patterns that were unique to each municipality with statistical significance with respect to high frequency.

Before addressing these questions, some brief background on the sequential pattern analysis techniques is given.

4.1 Sequential Pattern Mining

Sequential pattern mining (SPM) [2, 5] is a research discipline within the field of data mining that focuses on identifying frequently occurring sequences of objects or events. Let I be a set of *items*, and S be a set of *input sequences*, where each $s \in S$ consists of an ordered list of *itemsets*, or sets of items from I, also known as *transactions*. A sequence $\langle a_1 a_2 \dots a_n \rangle$ is said to be *contained* in another sequence $\langle b_1 b_2 \dots b_m \rangle$ if there exist integers $i_1, i_2, \dots, i_n$ with $i_1 < i_2 < \dots < i_n$ such that $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$. A sequence $s \in S$ *supports* a sequence s' if s' is contained in s. The support $sup(s')$ for a sequence s' given a set S of input sequences is the percentage of sequences in S that support s', and is equal to $sup(s') = |\{s \in S | s \text{ supports } s'\}| / S$. A sequence s' is deemed a *sequential pattern* if $sup(s')$ is greater than some pre-specified minimum support. Such a pattern with a total cardinality of its itemsets summing to n is referred to as an *n-sequence* or *n-pattern*. A sequential pattern s' is a *maximal sequential pattern* in S' if $\forall s'' \in S'$ where $s'' \neq s'$, s'' does not contain s'. The general goal of sequential pattern mining is then to identify the set S' that contains all (and only those) sequences that are deemed sequential patterns according to the above. In some cases, the set consisting of only maximal sequential patterns is preferred.

To illustrate, consider the example set S of sequences given in the first column of Fig. 17. This will be used as a running example throughout the paper. Using a minimum support of 0.4, the set of all sequential patterns is demonstrated, specified by 1-sequences, 2-sequences and 3-sequences.

Note that, while this table gives all of the frequent sequential patterns, only $\langle \{a\}, \{d\} \rangle$, $\langle \{c\}, \{c\} \rangle$, $\langle \{b, c\}, \{d\} \rangle$, $\langle \{b\}, \{d\}, \{e\} \rangle$ and $\langle \{c\}, \{d\}, \{e\} \rangle$ are *maximal* sequential patterns.

Sequence Database	1-seq	2-seq	3-seq
$\langle\{b,c\},\{c,d\},\{e\}\rangle$	$\langle\{a\}\rangle$	$\langle\{a\},\{d\}\rangle$	$\langle\{b,c\},\{d\}\rangle$
$\langle\{a,c\},\{b,c\},\{d\}\rangle$	$\langle\{b\}\rangle$	$\langle\{b,c\}\rangle$	$\langle\{b\},\{d\},\{e\}\rangle$
$\langle\{c\},\{e\}\rangle$	$\langle\{c\}\rangle$	$\langle\{b\},\{d\}\rangle$	$\langle\{c\},\{d\},\{e\}\rangle$
$\langle\{c\},\{d\},\{e,f\}\rangle$	$\langle\{d\}\rangle$	$\langle\{b\},\{e\}\rangle$	
$\langle\{a,b\},\{d\},\{e\}\rangle$	$\langle\{e\}\rangle$	$\langle\{c\},\{c\}\rangle$	
		$\langle\{c\},\{d\}\rangle$	
		$\langle\{c\},\{e\}\rangle$	
		$\langle\{d\},\{e\}\rangle$	

Fig. 17. Example sequence database with mined sequential patterns using a minimum support of 0.4.

4.2 Value-Based Sequential Pattern Mining

The value-based sequential pattern mining problem uses the same set of inputs as the classical SPM problem, with the addition of a function $v : S \to \mathbb{R}$, where $v(s)$ indicates the specified value of s. Informally, the goal of the problem is then to identify sequential patterns that are frequently supported by sequences with high value and, in particular, are likely to contribute to increased sequence value when present. Thus it is not simply enough to find sequential patterns in high-valued sequences, but to find those sequential patterns that are *responsible*, at least in part, for the increased value. That is, input sequences containing these sequential patterns will be likely to be ranked higher according to value than those without, all else equal, and these sequential patterns do not simply contain sub-patterns that are at least as beneficial. Such sequential patterns are referred to as *value-influential*.

Formally, let S be the set of input sequences with value function $v : S \to \mathbb{R}$, let $r(s)$ be the rank of sequence $s \in S$, (where smaller values for rank are associated with higher value, e.g. for n sequences, the best sequence has rank 1 while the worst has rank n) and let $ar(sp, S)$ and $ar'(sp, S)$ be the *average rank* of sequences in S that contain the sequential pattern sp and of those that do not contain sp, respectively. The sequential pattern sp is deemed *value-influential* if both (1) $ar(sp, S) < ar'(sp, S)$ and (2) there is no sequential pattern sp' such that sp' is contained in sp and $ar(sp', S) \leq ar(sp, S)$. Thus the existence of an influential sequential pattern will likely improve the ranking of the sequence that contains it, and this improvement is not due simply to a sequential pattern contained inside it.

As an illustration, consider the example database given in Fig. 17, complete with example values and corresponding ranks as given in Fig. 18. Fig. 19 gives the average rank for each sequential pattern. Note that any sequential pattern with average rank less than 3 (i.e. the median rank) will satisfy condition (1) $ar(sp, S) < ar'(sp, S)$. Condition (2) eliminates all sequential patterns with average rank greater than or equal to that of any of its subsequences. For example, $\langle\{c\},\{d\},\{e\}\rangle$ would not be deemed influential since it has an average rank of 2.5, while $\langle\{c\},\{d\}\rangle$ has an average rank of 2.33, which suggests that $\langle\{c\},\{d\},\{e\}\rangle$

is only associated with higher-valued input sequences since it contains the more influential sequential pattern $\langle\{c\},\{d\}\rangle$. In all, the resulting influential sequential patterns are $\langle\{b\}\rangle$, $\langle\{c\}\rangle$, $\langle\{b,c\}\rangle$, $\langle\{c\},\{c\}\rangle$ and $\langle\{c\},\{d\}\rangle$ (depicted in the rightmost column of Fig. 18).

Sequence Database	Value	Rank	Influential
$\langle\{b,c\},\{c,d\},\{e\}\rangle$	9	1	$\langle\{b\}\rangle$
$\langle\{a,c\},\{b,c\},\{d\}\rangle$	6	2	$\langle\{c\}\rangle$
$\langle\{c\},\{e\}\rangle$	4	3	$\langle\{b,c\}\rangle$
$\langle\{c\},\{d\},\{e,f\}\rangle$	2	4	$\langle\{c\},\{c\}\rangle$
$\langle\{a,b\},\{d\},\{e\}\rangle$	1	5	$\langle\{c\},\{d\}\rangle$

Fig. 18. Example sequence database with example input sequence values and associated ranks, and the value-influential sequential patterns that result.

1-sequences	2-sequences	3-sequences
$\langle\{a\}\rangle$ 3.5	$\langle\{a\},\{d\}\rangle$ 3.5	$\langle\{b,c\},\{d\}\rangle$ 1.5
$\langle\{b\}\rangle$ 2.67	$\langle\{b,c\}\rangle$ 1.5	$\langle\{b\},\{d\},\{e\}\rangle$ 3
$\langle\{c\}\rangle$ 2.5	$\langle\{b\},\{d\}\rangle$ 2.67	$\langle\{c\},\{d\},\{e\}\rangle$ 2.5
$\langle\{d\}\rangle$ 3	$\langle\{b\},\{e\}\rangle$ 3	
$\langle\{e\}\rangle$ 3.25	$\langle\{c\},\{c\}\rangle$ 1.5	
	$\langle\{c\},\{d\}\rangle$ 2.33	
	$\langle\{c\},\{e\}\rangle$ 2.67	
	$\langle\{d\},\{e\}\rangle$ 3.33	

Fig. 19. Sequential patterns with average ranks used to obtain the set of value-influential sequential patterns in Fig. 18.

4.3 Question 3: Change in Processes

This section provides an analysis of the changes that occur in each municipality's process over time, with the objective of addressing question 3, i.e. to determine whether there is evidence of two municipalities physically moving into the same location, and whether this led to a change in processes.

Methodology The methodology for this particular aspect of the study was to perform value-based sequential pattern mining *with respect to case completion time* on the event log for each municipality. Case completion time is taken as the finish time (given under `Complete Timestamp`) for the final activity executed in the case. This mining is then done twice for each log: once where a higher value is associated with earlier completion time, and one where higher value is associated with later completion time. The result will thus consist of two sets

of activity flows for each municipality: one that tends to occur earlier in the process, and one that tends to occur later in the process. This will allow us to easily assess how each pair of processes may have converged/diverged over time.

The data was analyzed as follows. As a preprocessing step for each log, each case was labeled with a normalized value according to completion time of its final action. Thus for the early completion scenario, the case with earliest completion time took the value 1, down to the case that completed last, which was labeled by 0. The opposite was done for the later completion scenario. Value-based sequential pattern mining was then performed on each log with the parameters given in Table 2. We refer to the sequential patterns that meet these thresholds as *high-impact* sequential patterns. Thus a high-impact pattern for the early completion scenario would be a pattern that appeared early in the process with high frequency, and a high-impact pattern for the later completion scenario would be a pattern that appeared later on in the process with high frequency.

Parameter	Explanation
$min\_support = 0.1$	Any sequential pattern returned must appear in at least 0.1 of the cases in the log
$min\_avg\_value = 0.55$	A pattern's value (i.e. the average value of cases in which the pattern appears) must be at least 0.55, thus a significantly higher value than the median (0.5)
$min\_value\_increase = 0.1$	A pattern's value must be at least 0.1 higher than any of its sub-patterns, thus avoiding the inclusion of numerous super-patterns of high valued patterns that do not provide much benefit to the analysis

Table 2. Parameters for value-based sequential pattern mining

Once the high-impact patterns have been established for each of the early and late completion analyses, a set of common high-impact sequential patterns among all five municipalities is presented to show where the universal similarities lie. A sequential pattern is considered to be common if both 1) it is found to be of high impact for some municipality, and 2) for all other municipalities, it has no more than 10% less value and at least 1/3 the support.

Once the commonalities have been established, an examination of high-impact control sequences for each specific municipality is then conducted. For each municipality, a sample of the more high-impact control sequences is presented. An analysis is then conducted that assesses the impact of the municipality's high-impact patterns in each of the other municipalities, and a statistical comparison is given to determine where the key similarities and differences lie. Note that the high-impact patterns that are identified as common among the five municipalities are discarded, since they have little bearing on this part of the analysis.

The process of assessing the impact of a high-impact sequential pattern in each of the other municipalities is achieved by computing the *relative impact*

of the pattern in each municipality. Specifically, for a sequential pattern s that has high impact in municipality i, the relative impact of s with respect to i in municipality j is equal to the value of s in j, multiplied by a discount factor d where

$$d = \min\{1, \frac{support_j(s)}{support_i(s)}\}$$

Thus if s has a value v_j in j, but only half the support that it has in i, its value v_j will be reduced by half to give the relative impact. Note that d can be no greater than 1, and thus its value can not increase simply due to high frequency.

The key differences between the frequent sequential patterns appearing early/late in the process for a municipality can thus be seen by identifying the high-impact patterns that have low relative impact in other municipalities. For each municipality, a sample of such patterns are listed. Additionally, a statistical analysis is conducted to assess the correlation of the values of all high-impact sequential patterns with their relative impact values in the other municipalities, in order to determine which municipalities are are generally more similar/different. A final analysis is then made to assess which municipalities may be converging or diverging over the period of the study.

Early Completion There were 60 high-impact sequential patterns found to commonly occur more often early in the log across all five municipalities. These 60 sequential patterns were thus omitted from the subsequent analysis of high-impact sequential patterns for individual municipalities. The list below gives a sample of the highest impact patterns in terms of average value across municipalities. Note that the sequence value of 0.833 can be interpreted as the level of "earliness" on a scale of 0-1. Thus the average completion time for `[[reception through OLO]]` is earlier than 88.3% of the finish times in the log, or in other words, about 11.7% of the way into the log. The second value provided for each pattern indicates the average support.

- `[[reception through OLO]]` 0.883 0.193
- `[[registrer date of publishing received request]]` 0.875 0.180
- `[[WAW permit aspect], [terminate on request]]` 0.874 0.098
- `[[WAW permit aspect], [activities regular procedure]]` 0.870 0.142
- `[[activities regular procedure], [phase advice known]]` 0.855 0.088
- `[[activities regular procedure], [assessment of content completed]]` 0.854 0.093
- `[[completed subcases content], [activities regular procedure]]` 0.852 0.094
- `[[suspension ground applicable], [article 34 WABO applies]]` 0.846 0.082
- `[[terminate on request], [extend procedure term]]` 0.846 0.133
- `[[phase advice known], [suspension ground applicable]]` 0.844 0.160

Tables 3 to 7 show the top outliers for each of the five municipalities (labeled M1-M5), where an outlier in this context is considered to be a high-impact

sequential pattern that has low relative impact in another municipality. For each pattern, the average value (top) and support (bottom) are given for each municipality, and the cells that stray far from the values for municipality in question are shaded. The final row in each table then gives the mean square error (MSE) for the relative impact for each municipality when compared to the impact for the municipality in question, thus giving a some perspective on the similarity of the municipality's process to each of the others.

For example, for municipality M1 in Table 3 we can see that M4 has a very low relative support for the sequence `[enter senddate acknowledgement, send confirmation receipt]` compared to M1, meaning that it is only executed about 1/7 of the time. Similarly, `[enter senddate acknowledgement, regular procedure without MER]` occurs never or almost never (i.e. less than 1% of the time) in M4 (and thus the average weight does not exist or is otherwise unknown). The action `[investigate BAG objects]` also appears infrequently in M3, M4 and M5, and also has an extremely low weight in M5, meaning that it only occurred very late in the process. The final row indicates that M2 is by far the most simalar to M1, with a MSE of just 0.01, while M4 is the least similar.

Sequence	**M1**	**M2**	**M3**	**M4**	**M5**
[enter senddate acknowledgement],	0.84	0.75	0.81	0.88	0.83
[send confirmation receipt]	0.14	0.10	0.17	0.02	0.13
[enter senddate acknowledgement],	0.67	0.60	0.48	-	0.65
[regular procedure without MER]	0.12	0.20	0.19	0.00	0.18
[investigate BAG objects]	0.62	0.57	-	-	0.03
	0.14	0.10	0.00	0.00	0.03
Mean standard error	**-**	**0.01**	**0.05**	**0.11**	**0.07**

Table 3. Outliers for M1's high-impact sequential patterns for early completion

Table 4 shows that there a number of high-impact patterns in municipality 2 that appear very infrequently in municipalities 2-4. As a result, municipality 5 is by far the most similar.

Table 5 shows two high-impact patterns for municipality 3 that appear infrequently in municipality 4, as well as the `[no permit needed or only notification needed]` activity that has very low value for M5 and thus appears in the *later* stages of that process, and with much higher frequency.

Table 6 gives an example of a high-impact pattern (`[enter senddate acknowledgement]` that appears to be an outlier for this particular municipality, when compared to the others, showing that it appears almost all the time, both early and late, in each of the other municipalities, while only appearing very early in the process for municipality 4.

Finally, Table 7 shows that municipality 5 has a number of high-impact patterns that have low weight and/or frequency in other municipalities, in particular municipality 1. Municipality 2 is clearly the most similar.

Sequence	M2	M1	M3	M4	M5
[objection lodged against decision,	0.86	0.79	-	0.85	0.82
close case]	0.14	0.03	0.00	0.04	0.25
[enter date publication decision environmental permit,	0.80	0.70	-	0.72	0.77
set phase: phase permitting irrevocable]	0.23	0.04	0.00	0.08	0.28
[enter date publication decision environmental permit,	0.76	0.69	-	-	0.75
register deadline]	0.15	0.03	0.00	0.00	0.21
[register deadline,	0.70	0.71	-	0.78	0.76
close case]	0.28	0.04	0.00	0.03	0.24
[objection lodged against decision,	0.70	-	-	-	0.72
register deadline]	0.20	0.00	0.00	0.00	0.18
Mean standard error	**-**	**0.17**	**0.23**	**0.18**	**0.06**

Table 4. Outliers for M2's high-impact sequential patterns for early completion

Sequence	M3	M1	M2	M4	M5
[register submission date request,	0.84	0.85	0.73	0.89	0.84
enter senddate acknowledgement]	0.21	0.19	0.22	0.07	0.19
[enter senddate acknowledgement,	0.81	0.84	0.75	0.88	0.83
send confirmation receipt]	0.17	0.14	0.10	0.02	0.13
[no permit needed or only notification needed]	0.71	0.70	0.76	0.74	0.34
	0.14	0.12	0.11	0.15	0.47
Mean standard error	**-**	**0.01**	**0.03**	**0.05**	**0.02**

Table 5. Outliers for M3's high-impact sequential patterns for early completion

Sequence	M4	M1	M2	M3	M5
[enter senddate acknowledgement]	0.85	0.51	0.49	0.50	0.49
	0.12	0.88	0.94	0.93	0.93
[generating decision environmental permit,	0.83	0.77	0.81	0.82	0.82
enter senddate decision environmental permit]	0.10	0.03	0.09	0.09	0.14
[phase application received,	0.75	0.79	0.79	0.78	0.60
forward to the competent authority]	0.38	0.18	0.20	0.07	0.16
[transcript decision environmental permit to stakeholders,	0.74	0.74	0.71	-	0.69
enter senddate decision environmental permit,	0.21	0.01	0.11	0.00	0.15
objection lodged against decision]					
Mean standard error	**-**	**0.07**	**0.04**	**0.08**	**0.04**

Table 6. Outliers for M4's high-impact sequential patterns for early completion

Sequence	M5	M1	M2	M3	M4
[phase application received, enter senddate procedure confirmation]	0.91	0.90	0.87	0.90	0.82
	0.10	0.04	0.03	0.06	0.02
[enter senddate decision environmental permit, transcript decision environmental permit to stakeholders]	0.83	0.84	0.83	0.89	0.83
	0.18	0.04	0.07	0.08	0.03
[inform BAG administrator, regular procedure without MER]	0.68	0.70	0.72	0.69	0.74
	0.22	0.02	0.08	0.11	0.16
[objection lodged against decision]	0.60	0.42	0.53	0.63	0.56
	0.61	0.11	0.62	0.03	0.60
[set phase: phase permitting irrevocable]	0.60	0.40	0.53	0.59	0.45
	0.65	0.13	0.74	0.03	0.43
[close case]	0.59	0.49	0.52	0.62	0.72
	0.69	0.14	0.74	0.04	0.09
Mean standard error	-	**0.19**	**0.07**	**0.19**	**0.18**

Table 7. Outliers for M5's high-impact sequential patterns for early completion

Later Completion As was the case with the early completion dataset, there were 60 high-impact sequential patterns found to commonly occur more later on in the log across all five municipalities. These 60 sequential patterns were thus omitted from the subsequent analysis of high-impact sequential patterns for individual municipalities. The list below gives a sample of the highest impact patterns in terms of average value across municipalities.

- `[[create subcases content]]` 0.903 0.118
- `[[procedure change], [ask stakeholders views]]` 0.903 0.108
- `[[publish], [forward to the competent authority]]` 0.900 0.111
- `[[start WABOprocedure], [publish]]` 0.895 0.125
- `[[create procedure confirmation], [publish]]` 0.885 0.167
- `[[create letter requesting missing data]]` 0.828 0.099
- `[[enter senddate decision], [record date of decision environmental permit]]` 0.821 0.077
- `[[phase decision sent], [record date of decision environmental permit]]` 0.813 0.089
- `[[publish]]` 0.763 0.246
- `[[create procedure confirmation]]` 0.760 0.456

Tables 8 to 12 show the top outliers for each of the five municipalities. Table 8 shows that municipality 1 has a number of high-impact patterns appearing later in the process that are virtually non-existent for many of the other municipalities. Of particular note is the pattern where `send confirmation receipt` is repeated, which happens commonly later in the process for all municipalities but M4. Overall, municipalities 2 and 5 appear to be significantly more similar to municipality 1 than municipalities 3 and 4.

Table 9 shows that municipality 2 has a number of high-impact patterns appearing later in the process that are virtually non-existent for municipalities 1, 3 and 4. As a result, municipality 5 is by far the most similar.

Sequence	M1	M2	M3	M4	M5
[publish,	0.91	-	-	-	0.96
[create publication document]	0.14	0.00	0.00	0.00	0.05
[forward to the competent authority,	0.78	0.62	0.67	-	0.58
procedure change]	0.11	0.02	0.10	0.00	0.01
[registration date publication,	0.71	-	-	0.44	0.94
regular procedure without MER]	0.37	0.00	0.00	0.03	0.03
[set phase: phase permitting irrevocable]	0.60	0.47	0.41	0.55	0.40
	0.13	0.74	0.03	0.43	0.65
[send confirmation receipt,	0.59	0.59	0.62	0.28	0.61
send confirmation receipt]	0.75	0.73	0.73	0.04	0.67
Mean standard error	**-**	**0.17**	**0.28**	**0.32**	**0.15**

Table 8. Outliers for M1's high-impact sequential patterns for late completion

Sequence	M2	M1	M3	M4	M5
[publish,	0.90	-	-	-	0.88
enter senddate acknowledgement]	0.13	0.00	0.00	0.00	0.07
[regular procedure without MER,	0.82	0.75	0.82	-	0.81
enter senddate acknowledgement]	0.24	0.29	0.01	0.00	0.25
[close case,	0.75	-	-	-	0.74
set phase: phase permitting irrevocable]	0.26	0.00	0.00	0.00	0.15
[stop all running subcases 2b]	0.73	0.77	0.76	0.77	0.71
	0.33	0.08	0.02	0.08	0.23
[close case,	0.71	-	-	-	0.47
register deadline]	0.18	0.00	0.00	0.00	0.01
Mean standard error	**-**	**0.23**	**0.35**	**0.34**	**0.06**

Table 9. Outliers for M2's high-impact sequential patterns for late completion

Table 10 depicts some deviations for municipality 3. Municipality 2 has a clear edge in terms of similarity.

Sequence	M3	M1	M2	M4	M5
[phase application received],	0.91	0.64	0.68	-	-
enter senddate acknowledgement,	0.11	0.06	0.02	0.00	0.00
create procedure confirmation]					
[send letter in progress,	0.90	0.91	0.91	0.92	0.90
phase advice known]	0.11	0.08	0.02	0.09	0.03
[forward to the competent authority	0.73	0.80	0.73	0.77	0.75
request complete]	0.17	0.02	0.02	0.05	0.04
[regular procedure without MER	0.67	0.75	0.83	-	0.65
procedure change]	0.25	0.24	0.07	0.00	0.06
Mean standard error	**-**	**0.10**	**0.18**	**0.21**	**0.19**

Table 10. Outliers for M3's high-impact sequential patterns for late completion

Table 11 shows that there some extreme deviations in high-impact patterns for municipality 4 from all other municipalities. This gives a clear indication this municipality's process has experienced a significant deviation over time from the others.

Sequence	**M4**	**M1**	**M2**	**M3**	**M5**
[procedure change,	0.90	-	-	-	-
enter senddate procedure confirmation]	0.11	0.00	0.00	0.00	0.00
[send confirmation receipt,	0.90	-	-	-	-
send procedure confirmation,	0.11	0.00	0.00	0.00	0.00
publish]					
[procedure change,	0.74	0.82	0.56	0.73	0.75
forward to the competent authority]	0.34	0.05	0.03	0.12	0.03
[regular procedure without MER,	0.73	0.79	0.63	0.63	0.59
phase application receptive]	0.23	0.10	0.02	0.10	0.02
[calculate provisional charges]	0.66	-	0.67	-	0.64
	0.35	0.00	0.29	0.00	0.36
Mean standard error	**-**	**0.32**	**0.31**	**0.31**	**0.30**

Table 11. Outliers for M4's high-impact sequential patterns for late completion

Finally, Table 12 shows that municipality 5 has also witnessed a significant deviation from other municipalities, except for municipality 2, where clear similarities continue to be demonstrated.

Sequence	**M5**	**M1**	**M2**	**M3**	**M4**
[create monitoring case oversight]	0.89	-	0.89	-	-
	0.12	0.00	0.12	0.00	0.00
[forward to the competent authority	0.89	0.76	0.88	0.82	-
enter senddate acknowledgement]	0.15	0.29	0.14	0.01	0.00
[phase archived case,	0.80	0.78	0.76	-	-
set phase: phase permitting irrevocable]	0.11	0.01	0.24	0.00	0.00
[suspension ground applicable,	0.77	-	-	-	-
no permit needed or only notification needed]	0.16	0.00	0.00	0.00	0.00
[no permit needed or only notification needed]	0.66	0.30	0.24	0.29	0.26
	0.47	0.12	0.11	0.14	0.15
Mean standard error	**-**	**0.20**	**0.05**	**0.30**	**0.28**

Table 12. Outliers for M5's high-impact sequential patterns for late completion

Table 13 summarizes the results of the study. For each municipality listed on the left hand side, the mean square error associated with the weights of the municipality's high-impact patterns for early completion (top value) and later completion (bottom value), as compared to the relative impact for each of the other municipalities, is given. For example, consider the high-impact patterns for M1. If we compare the weight of these patterns with the relative impact of those patterns in M2, we see a mean square error of 0.014 when looking at

the early completion patterns, and 0.177 when looking at the later completion patterns. This indicates that there may have been a significant deviation in the two municipalities' processes.

Municipality	M1	M2	M3	M4	M5
M1	-	0.014	0.052	0.106	0.071
	-	0.177	0.278	0.316	0.147
M2	0.168	-	0.230	0.180	0.057
	0.229	-	0.345	0.339	0.064
M3	0.009	0.027	-	0.047	0.024
	0.104	0.178	-	0.210	0.189
M4	0.067	0.039	0.078	-	0.044
	0.321	0.310	0.308	-	0.296
M5	0.194	0.072	0.194	0.182	-
	0.199	0.050	0.296	0.282	-

Table 13. Comparison of differences between municipalities for early vs late case completion

We draw two main conclusions from this data:

1. *There has been a significant deviation in municipalities' processes over time.* The mean square error has increased from early to later completion across the board, and in some cases, quite dramatically, with an average increase of 339%. Municipalities 3 and 4 have shown the most deviation in terms of average pct increase (M3, 668%) and average absolute increase (M4, 0.252).
2. *Municipalities 2 and 5 have converged.* When compared with the global increase of 339%, the (M2, M5) error only increased by 11% from 0.057 to 0.064. Moreover, the (M5, M2) error actually *decreased* from 0.072 to 0.05. Thus, when adjusted for the global error, a significant decrease has been observed in the error, suggesting convergence in the processes.

Figure 20 depicts the impact values for M5 over time (indicated by the smooth curve), where the left half shows the value of high-impact patterns for early completion and right half shows the value for later completion, accompanied by the relative impact for each pattern for M2. Outside of a small number of outliers on the right-hand side, one can clearly see the processes converge over time, particularly in the very late stages. Thus there is sufficient evidence to conclude that the employees from the two municipalities may have begun to work together, and it has indeed had an effect on their processes. Tables 9 and 12 identify a number of such patterns that are responsible for this convergence.

4.4 Question 5: Differences in Throughput Times

Similar to that done above, this section approaches the analysis of municipalities' throughput time differences by identifying high-impact sequential patterns,

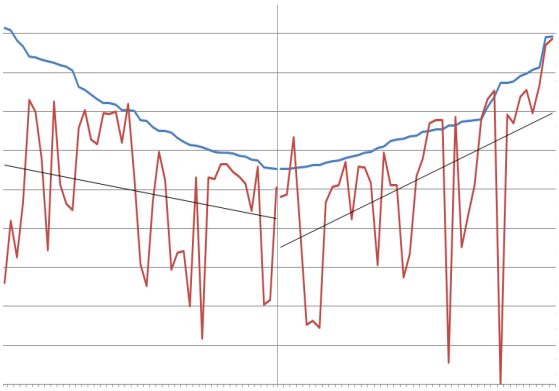

Fig. 20. Convergence of processes for municipalities 2 and 5.

where such a pattern in this context is deemed to be one that is often associated with high throughput-time cases. This is done for each municipality, and a few select high-impact patterns that have low relative impact in the other municipalities are highlighted. These are depicted below in Tables 15 to 19.

Table 14 shows the average throughput time by municipality. As one can clearly see, there are some obvious discrepancies. In particular, M2 has an average throughput time significantly longer than the other municipalities, especially when compared with M3 where there is a 157% increase.

Municipality	Avg Throughput Time (Days)
M1	95.7
M2	160.1
M3	62.2
M4	116.8
M5	98.3

Table 14. Average throughput time

4.5 Question 6: Differences in Control Flow Between the Municipalities

To identify the key differences in control flow, we took the approach of identifying sequential patterns that are most unique to a particular municipality. To

Sequence	M1	M2	M3	M4	M5
[set phase: phase permitting irrevocable]	0.79	0.56	0.72	0.66	0.58
	0.13	0.74	0.03	0.43	0.65
[objection lodged against decision]	0.76	0.51	0.67	0.60	0.55
	0.11	0.62	0.03	0.60	0.61
[register deadline]	0.70	0.55	0.64	0.42	0.56
	0.16	0.75	0.04	0.09	0.66

Table 15. Outliers for M1's high-impact sequential patterns for high throughput time

Sequence	M2	M1	M3	M4	M5
[appeal logded]	0.81	0.96	0.94	0.85	0.90
	0.15	0.03	0.01	0.07	0.06
[publish, subcases completeness completed]	0.76	0.84	0.89	0.76	0.72
	0.10	0.02	0.04	0.02	0.05
[create publication document]	0.60	0.51	0.55	0.64	0.41
	0.21	0.47	0.21	0.27	0.23

Table 16. Outliers for M2's high-impact sequential patterns for high throughput time

Sequence	M3	M1	M2	M4	M5
[applicant is stakeholder, forward to the competent authority]	0.60	-	0.45	0.53	0.54
	0.10	0.00	0.02	0.03	0.01
[create subcases completeness, regular procedure without MER]	0.60	0.53	0.41	0.19	0.53
	0.12	0.13	0.02	0.10	0.01
[article 34 WABO applies, grounds for refusal]	0.58	0.58	0.65	0.58	0.73
	0.10	0.07	0.05	0.04	0.03
[enter senddate acknowledgement, send procedure confirmation]	0.57	0.49	0.44	-	0.69
	0.11	0.24	0.04	0.00	0.06

Table 17. Outliers for M3's high-impact sequential patterns for high throughput time

Sequence	M4	M1	M2	M3	M5
[set phase: phase permitting irrevocable]	0.66	0.79	0.56	0.72	0.58
	0.43	0.13	0.74	0.03	0.65
[calculate final charges]	0.62	-	0.51	-	0.42
	0.32	0.00	0.22	0.00	0.27
[read publication date field]	0.60	0.75	0.55	0.75	0.48
	0.32	0.10	0.25	0.03	0.20

Table 18. Outliers for M4's high-impact sequential patterns for high throughput time

Sequence	M5	M1	M2	M3	M4
[resume completeness subcases]	0.64	0.72	0.72	0.72	0.73
	0.17	0.05	0.09	0.05	0.02
[generating decision environmental permit,	0.62	0.57	0.63	0.62	0.61
enter senddate decision environmental permit]	0.14	0.03	0.09	0.09	0.10

Table 19. Outliers for M5's high-impact sequential patterns for high throughput time

accomplish this, a sequence classification [3, 4, 6] approach was taken to identify those sequential patterns that have the strongest ability to identify to which "class" (i.e. municipality in this case) a sequence of activities most likely belongs. For example, if such a sequential pattern is identified for municipality 1, then this means that any sequence of activities observed that contains this sequential pattern is most likely to come from municipality 1. All such sequential patterns are required to pass a chi-squared test for significance.

Tables 20 to 24 present the top two most significant unique sequential patterns for each municipality, according to the chi-squared statistic. Supports for each municipality are also indicated. For example, in Table 20, the first row indicates that this sequential pattern appeared in 34.5% of cases in municipality 1, but in only 0.7%, 0.1%, 2.4%, and 3.7% of cases in municipalities 2 to 5, respectively.

For a detailed statistical analysis on the comparison across municipalities, we direct the reader to section 4.3.

Sequence	M1	M2	M3	M4	M5
[register submission date request, phase application received, start WABOprocedure, registration date publication, forward to the competent authority]	0.345	0.007	0.001	0.024	0.037
[send confirmation receipt, send confirmation receipt, create procedure confirmation, registration date publication, create subcases completeness]	0.312	0.011	0.001	0.000	0.054

Table 20. Sequential patterns unique to municipality 1

5 Conclusions

This paper reported on findings resulting from our research conducted for the 2015 Business Process Intelligence Challenge (BPIC), an annual competition in

Sequence	M2	M1	M3	M4	M5
[enter senddate decision environmental permit, register deadline]	0.683	0.098	0.013	0.014	0.518
[close case]	0.738	0.142	0.038	0.093	0.689

Table 21. Sequential patterns unique to municipality 2

Sequence	M3	M1	M2	M4	M5
[OLO messaging active, phase application received, applicant is stakeholder, forward to the competent authority, regular procedure without MER]	0.448	0.001	0.105	0.002	0.245
[phase application received, applicant is stakeholder, enter senddate acknowledgement, forward to the competent authority, regular procedure without MER]	0.471	0.017	0.115	0.005	0.253

Table 22. Sequential patterns unique to municipality 3

Sequence	M4	M1	M2	M3	M5
[register submission date request, phase application received, send confirmation receipt, send procedure confirmation, create subcases completeness]	0.480	0.003	0.034	0.016	0.004
[OLO messaging active, phase application received, send procedure confirmation, applicant is stakeholder, regular procedure without MER]	0.360	0.002	0.020	0.009	0.003

Table 23. Sequential patterns unique to municipality 4

Sequence	M5	M1	M2	M3	M4
[OLO messaging active, phase application received, send confirmation receipt, applicant is stakeholder, terminate on request]	0.442	0.005	0.338	0.015	0.047
[no permit needed or only notification needed, record date of decision environmental permit]	0.282	0.000	0.001	0.004	0.000

Table 24. Sequential patterns unique to municipality 5

which participants are tasked with conducting process mining-related analyses on a real-life dataset. This year's data was provided by 5 Dutch municipalities, and contained activity pertaining to their building permit application process. A number of interesting connections and patterns were identified, often showing that there are many major discrepancies in the processes of the various municipalities. We hope our findings can be utilized to identify where these difference may cause problems, and ultimately lead to better practices and more efficient processes.

References

1. BPIC 2015. The 2015 business processing intelligence challenge (bpic). http://www.win.tue.nl/bpi/2015/challenge. Date accessed: Jul 12, 2015, 2015.
2. Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.
3. Neal Lesh, Mohammed J Zaki, and Mitsunori Ogihara. Mining features for sequence classification. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 342–346. ACM, 1999.
4. Neal Lesh, Mohammed J Zaki, and M Oglhara. Scalable feature mining for sequential data. *Intelligent Systems and their Applications, IEEE*, 15(2):48–56, 2000.
5. Carl H Mooney and John F Roddick. Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR)*, 45(2):19, 2013.
6. Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010.

A Annex 1: Social Analysis Figures

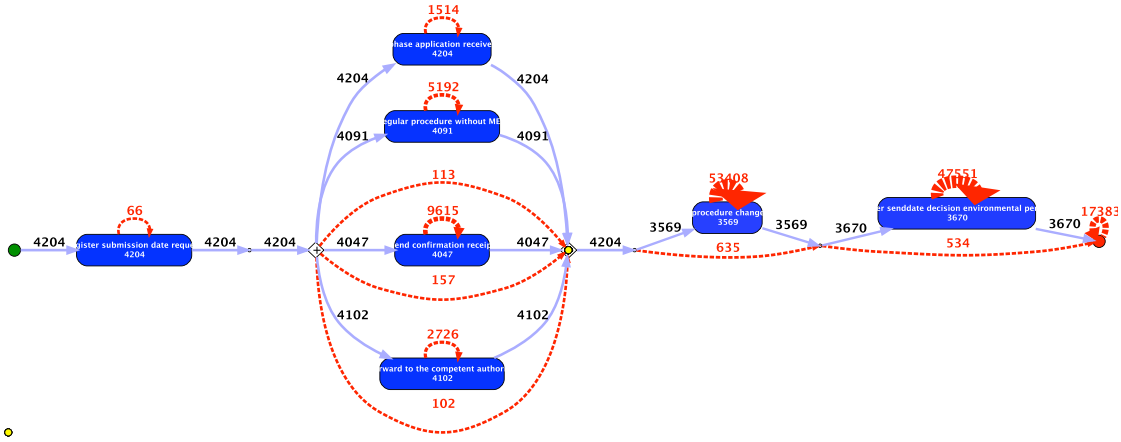

Fig. 21. All municipalities: process model.

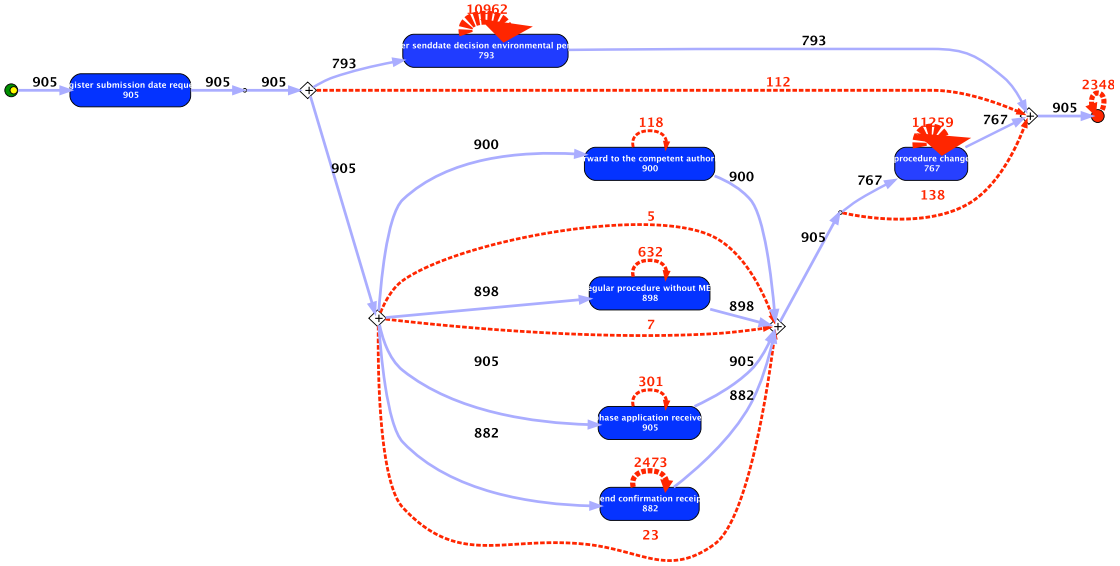

Fig. 22. Municipality 1: process model.

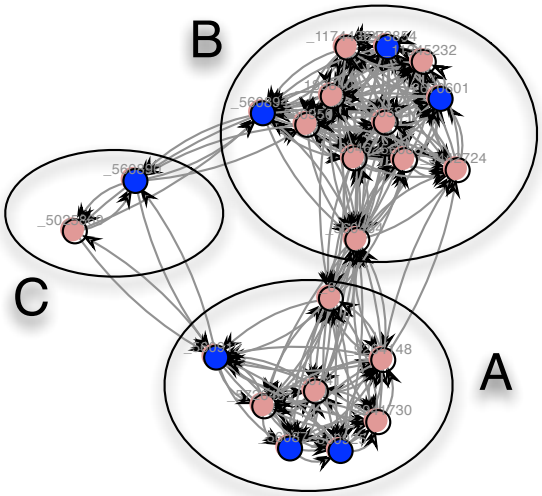

Fig. 23. Municipality 1: Clusters of similar tasks.

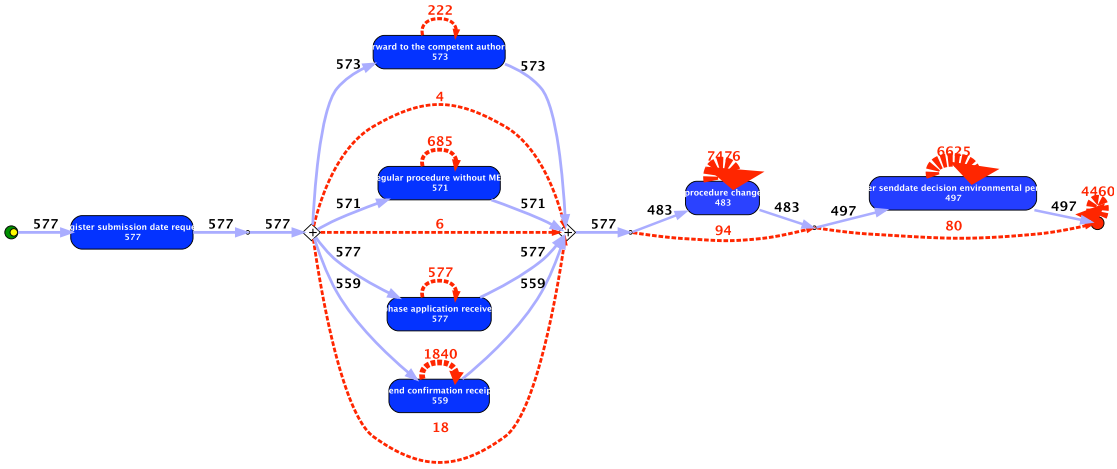

Fig. 24. Municipality 2: process model.

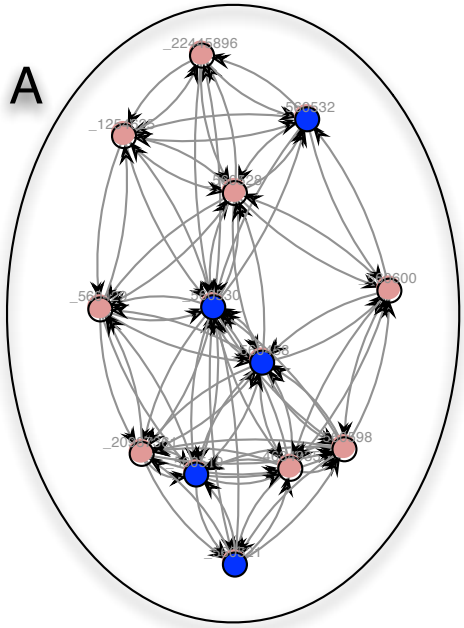

Fig. 25. Municipality 2: Clusters of similar tasks.

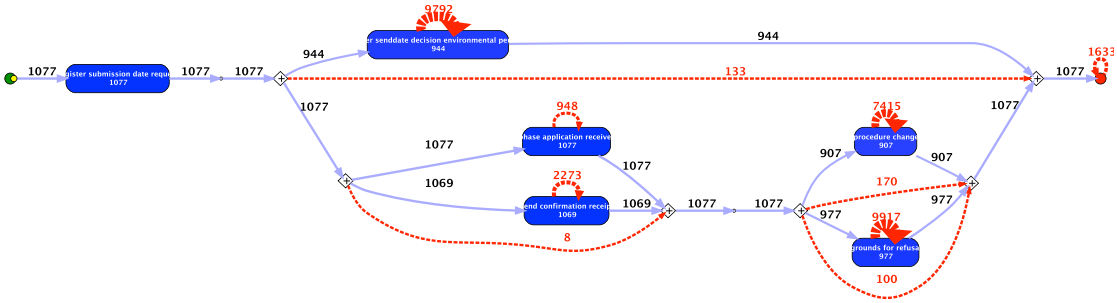

Fig. 26. Municipality 3: process model.

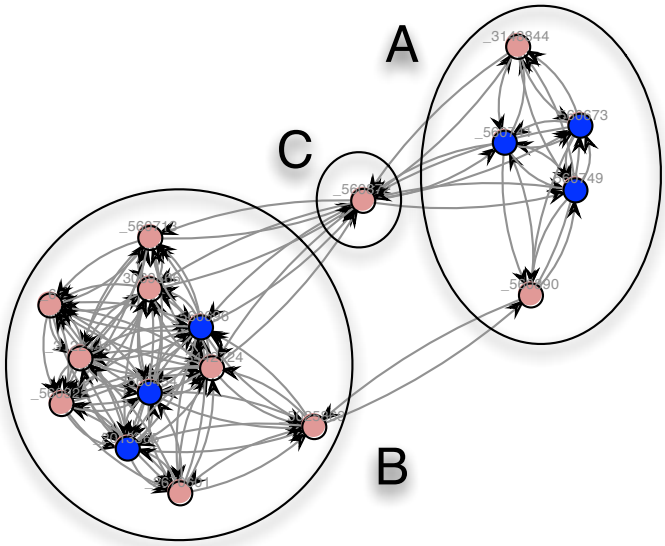

Fig. 27. Municipality 3: Clusters of similar tasks.

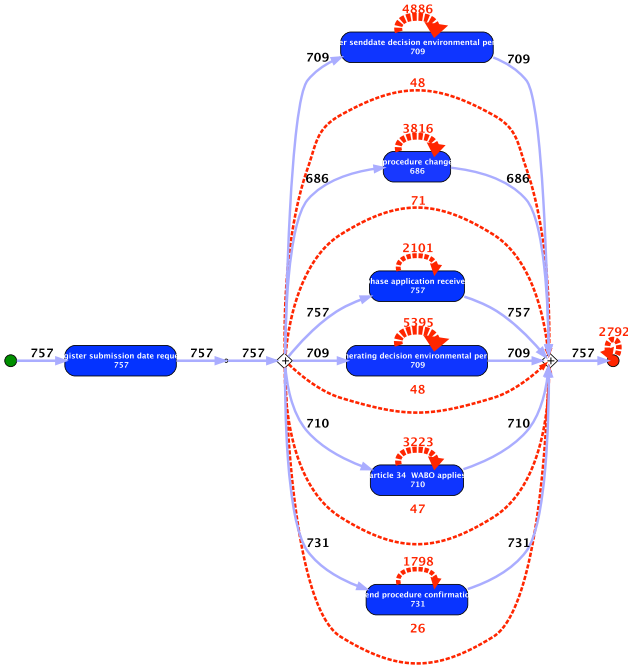

Fig. 28. Municipality 4: process model.

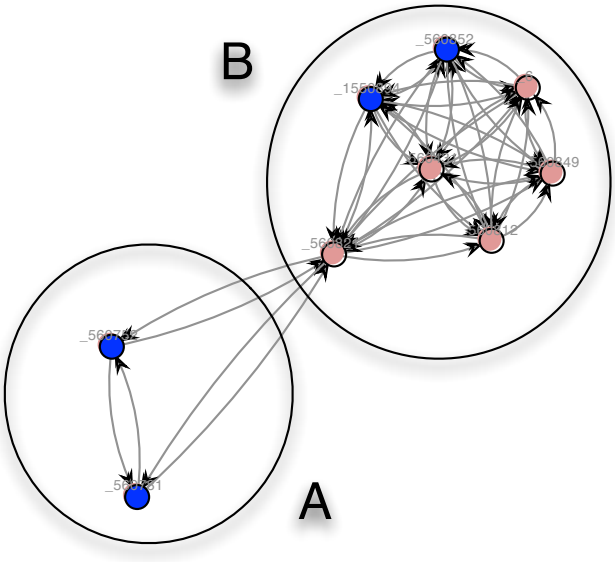

Fig. 29. Municipality 4: Clusters of similar tasks.

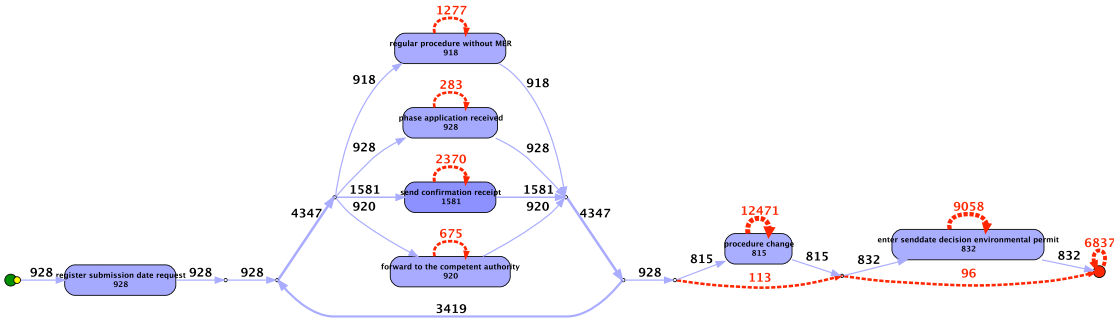

Fig. 30. Municipality 5: process model.

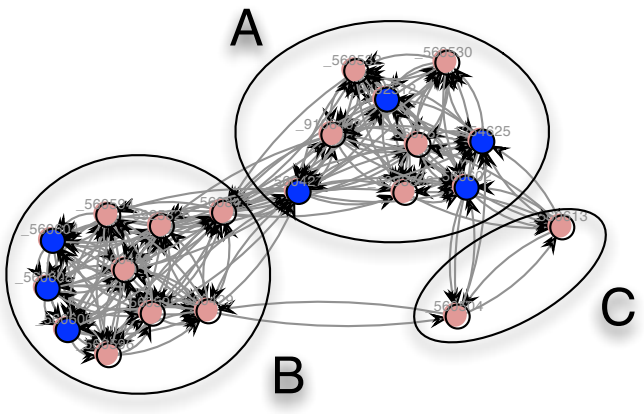

Fig. 31. Municipality 5: Clusters of similar tasks.