Looking Beyond Activity Labels: Mining Context-Aware Resource
Profiles Using Activity Instance Archetypes
Peer-reviewed author version

# Looking Beyond Activity Labels: Mining Context-Aware Resource Profiles using Activity Instance Archetypes

Gerhardus van Hulzen(✉)[1][0000−0001−8962−9515],
Niels Martin[1,2][0000−0003−3279−3853], and
Benoît Depaire[1][0000−0003−4735−0609]

[1] Hasselt University, Research group Business Informatics, 3500 Hasselt, Belgium
{gerard.vanhulzen,niels.martin,benoit.depaire}@uhasselt.be
[2] Research Foundation Flanders (FWO), 1000 Brussels, Belgium

**Abstract.** Efficient resource management is a critical success factor for all businesses. Correct insights into actual *resource profiles*, i.e. groups of resources performing similar activity instances, is important for successful knowledge and (human) resource management. To this end, organisational mining, a subfield of Process Mining, focuses on techniques to extract such resource profiles from event logs. However, existing techniques ignore contextual factors that impact how and by whom an activity is performed. This paper introduces the novel method *ResProMin* to discover *context-aware resource profiles* from event logs. In contrast to the state-of-the-art, this method builds upon the notion of activity instance archetypes, which incorporates the activity instance's context. An evaluation of the method on real-life event logs demonstrates its feasibility and potential to uncover valuable business insights.

**Keywords:** process mining · organisational mining · resource profiles · context-aware process mining

## 1 Introduction

Efficient resource management is a key success factors for all businesses. A comprehensive understanding of the complex relation between resources and activities enables efficient resource allocation and potential cost reductions [5,15]. To this end, process owners first need an objective insight into the *context-aware resource profiles*, i.e. *who* does *what* in *which context*?

*Organisational mining* – a subfield of Process Mining – focuses on discovering organisational structures and social networks within organisations from event logs [17] and addresses this need. Several research efforts focused on discovering resource profiles from event logs [1,2,9,17,20]. However, existing algorithms ignore context, i.e. the circumstances in which the activity was executed, and rely solely on activity labels to mine resource profiles. In real-life settings, this limiting assumption can hide important nuances. For instance, two nurses can

perform the same set of activities, but the patient's health condition might dictate the preference of one nurse over the other. While both nurses are equal based on activity labels, the context reveals that both nurses have a different profile. Consequently, there is a need for mining context-aware resource profiles from an event log.

This paper introduces the method *ResProMin* to generate context-aware resource profiles from event logs. Firstly, the method discovers *activity instance archetypes* reflecting the activity instance's context, i.e. the circumstances under which the activity instance was executed, such as case attributes and variables capturing the system state. Secondly, it assigns resources to these activity instance archetypes in a probabilistic manner, from which it discovers context-aware resource profiles. Not only do these profiles reveal *who* does *what* in *which context*, but it also allows the distinction between specialists and generalists.

The contribution of this work is twofold:

- The design of a novel method for discovering context-aware resource profiles is presented and discussed.
- A demonstration of the method on real-life datasets is presented to evaluate the method's feasibility and the ability to uncover valuable business insights.

An overview of the related work on this topic is provided in Section 2. Section 3 introduces the design of the ResProMin method. Next, the feasibility and value of ResProMin are evaluated in Section 4. Finally, the conclusion and opportunities for future research are discussed in Section 5.

## 2   Related Work

While the field of Process Mining traditionally focused on discovering the control-flow of processes from event logs, the sub-field *organisational mining* is receiving more and more attention [20]. Song and van der Aalst [17] were among the first to explore resource-related topics within a Process Mining context. They focused on discovering organisational structures and social networks from event logs leveraging task-based metrics based on joint activities [17]. These ideas are still used today, for instance, by Camargo *et al.* [4] to discover resource groups that perform similar tasks in their tool *Simod*.

Various resource-related topics have been investigated in Process Mining literature. To describe resource behaviour, Pika *et al.* [14] provide a framework to extract metrics on skills, productivity, utilisation, and collaboration patterns from event logs. Similarly, Nakatumba and van der Aalst [13] describe resource behaviour but specifically focus on the effect of workload on resource performance. Other researchers focused on rule mining to assign resources to tasks. Cabanillas *et al.* [3] developed *RALph Miner*, which is a tool to discover graphical resource-aware process models in which various task assignment rules are incorporated. Schönig *et al.* [16] also focused on finding assignment rules, but from a team perspective.

The most closely related research stream focuses on the discovery of groups of similar resources. In this respect, Jin *et al.* [9] propose an approach to mine resource roles, which are groups of resources that have performed the same tasks in similar volumes. This creates an abstraction layer between the individual resources and activities. A similar approach is proposed by Burattin *et al.* [2], who look at roles from the perspective of the handover of activities. However, they assume that a specific activity cannot belong to multiple groups at the same time. This assumption does often not hold in reality, where employees who possess several skills are not necessarily bound to one group [20].

To the best of our knowledge, there have been only two research efforts on detecting groups of resources that allow such overlapping group membership. Firstly, Appice [1] analysed the progress of communities over time in dynamic social networks while allowing communities to overlap. These communities represent a company's organisational units, and each resource has a certain degree to which it belongs to a particular unit. The second related research effort was conducted by Yang *et al.* [20]. They propose a Model-based Overlapping Clustering (MOC) model. The output of the MOC model is a boolean-valued membership vector, which indicates whether a particular resource belongs to a group or not.

All of the aforementioned papers which discover groups of similar resources rely on the *performer-by-activity matrix* as an input, except for Appice [1], who used a modified Louvain algorithm, and Burattin *et al.* [2], who relied on the notion of the handover of roles. The performer-by-activity matrix counts for each resource – i.e. the "performer" – how often (s)he executed each activity [17]. Although this is an effective and easy way to derive resource profiles, it is limited to only two dimensions: *who* did *what*. Therefore, information such as *when* or *under which circumstances* gets lost. Our paper anticipates upon this limitation by proposing a method to mine context-aware resource profiles.

## 3   Method

This paper introduces the novel method ResProMin to discover context-aware resource profiles from event logs, which consists of three steps (cfr. Figure 1). Firstly, we enrich the event log by adding relevant contextual variables. Secondly, we cluster the enriched event log from the first step to find *activity instance archetypes* using probabilistic model-based clustering and profile these clusters to get an overview of the different archetypes. Finally, we discover resource profiles by calculating, for each resource, the conditional probability that (s)he performs each activity instance archetype. Moreover, we determine whether a resource specialises in his/her work.

**Step 1: Context Enrichment** ResProMin assumes the presence of an event log that minimally describes each event by a case identifier, a timestamp or other attribute that allows temporal ordering of events, an activity label, and a resource identifier. Additionally, it also assumes that each activity instance
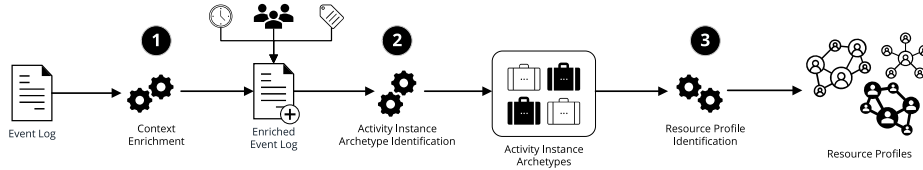
**Fig. 1.** The three steps of ResProMin: (1) context enrichment, (2) activity instance archetype identification, and (3) resource profile identification.

corresponds to a single event in the event log, which is common for most real-life event logs.

The first step adds computed and derived attributes to obtain an enriched event log. This allows us to describe *when* and *under which circumstances* an activity was executed: e.g., weekday, morning or evening shift, case type, activity duration, workload within parts of the process, and many more. The richer the event log, the more interesting patterns can be uncovered. An example of such an enriched event log is shown in Table 1, where each row represents an activity instance with various contextual attributes.

The number and which attributes can be added is different for each event log and mainly depends on the availability of information. However, it is important to consider that Step 2 will apply clustering directly to the enriched event log. Therefore, it is essential to include only attributes which are meaningful in a cluster analysis, e.g. it is best to omit the raw timestamp and use a more aggregated attribute, such as weekday, instead.

**Table 1.** Example of an enriched event log.

| Activity | Resource | Case Procedure | Weekday | Active cases | ... |
|---|---|---|---|---|---|
| Create purchase requisition | Anna | Comprehensive | Wednesday | 15 | ... |
| Amend request for quotation | Mike | Regular | Monday | 22 | ... |
| Send invoice | Jane | Regular | Thursday | 12 | ... |
| Confirm purchase order | Anna | Comprehensive | Friday | 6 | ... |
| Create purchase requisition | John | Regular | Friday | 14 | ... |
| Pay invoice | Jane | Regular | Tuesday | 18 | ... |
| ... | ... | ... | ... | ... | ... |

**Step 2: Activity Instance Archetype Identification** Next, we cluster the enriched event log to find *activity instance archetypes*. Each activity instance archetype comprises a set of activity instances that exhibit a high homogeneity with instances of the same archetype and high heterogeneity with instances in other archetypes. To identify activity instance archetypes, we propose to use Finite Mixture Models, which has the inherent advantage of using probabilities, providing statistical criteria to choose the number of clusters, and allowing the use of variables of different types, such as nominal, discrete, and continuous [19].

A *Finite Mixture Model (FMM)* is a probabilistic model-based clustering technique that allows overlapping clusters [11]. Suppose we have a set of $N$ data observations $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ and assume that the random variable $\mathbf{y}_n$ is distributed according to a mixture of $K$ components. Each component – or cluster – represents an activity instance archetype, is assumed to follow a parametric distribution, and has an assigned weight, i.e. the prior probability of observing cluster $k$, with $k = 1, \ldots, K$. The mixture density function $h$ is given by Equation 1.

$$h(\mathbf{y}_n \mid \vartheta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_n \mid \theta_k), \tag{1}$$

where $f_k(\mathbf{y}_n \mid \theta_k)$ is the $k^{\text{th}}$ component density function with parameter vector $\theta_k$, $\vartheta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ is the vector of all model parameters, and $\pi_k$ is the prior probability, or mixture proportion, which must satisfy $\sum_{k=1}^{K} \pi_k = 1$, where $\forall k : \pi_k > 0$. The parameters of this model ($\vartheta$) can be fitted using the *Expectation-Maximisation (EM)* algorithm, which tries to maximise the log-likelihood [11].

Gaussian distributions are often used in FMMs, which is then called a Gaussian Mixture Model (GMM). GMMs are used in many applications, including biology, physics, medicine, marketing, and economics [6]. However, because variables such as the activity label and resource identifier are nominal, we cannot use Gaussian distributions. Instead, we use multinomial distributions for these variables.

To determine the number of components – i.e. $K$ – we use the *Bayesian Information Criterion (BIC)*, which tries to balance the goodness-of-fit with the model complexity, i.e. it penalises more components harder. One should choose the number of components resulting in the lowest BIC [6], or the point where adding additional components barely improves the BIC [10].

Once the appropriate number of clusters ($K$) is determined, the intra-cluster distributions are used to profile each activity instance archetype using a label and a brief description. This makes it easier to refer to a particular archetype and enhances its recognisability by domain experts.

**Step 3: Resource Profile Identification** Context-aware resource profiles are groups of resources that perform similar activity instances, taking into account contextual information, and, hence, move beyond hierarchical functions or resource groups solely defined using activity label information. To identify these profiles, we first need to calculate the probability that a resource belongs to a particular activity instance archetype based on the intra-cluster resource distribution fitted by the FMM. To this end, we apply Bayes' Theorem:

$P(Cluster = c \mid Resource = r) \propto$
$$P(Resource = r \mid Cluster = c)P(Cluster = c), \tag{2}$$

where $P(Resource = r \mid Cluster = c)$ is the probability of observing resource $r$ in cluster $c$, and $P(Cluster = c)$ is the mixture proportion (also denoted by $\pi_c$).

After calculating these probabilities, we discover the resource profiles and determine whether a resource specialises in his/her work. We do this by first constructing a distance matrix using the Euclidean distance between the probabilities derived from Equation 2. Resources with a smaller "distance" are more closely related than resources with a larger "distance". Next, we cluster this matrix using *Agglomerative Hierarchical Clustering (AHC)* and choose the number of clusters where the *Total Within Sum of Square (WSS)* plot shows an "elbow" pattern [8]. These clusters form our final *resource profiles*.

Additionally, we can also find groups of resources with a similar degree of specialism. First, we transform the table derived using Equation 2, so that the probabilities of belonging to a particular cluster are ordered from left to right, i.e. the first column contains the cluster with the highest probability for a particular resource and the last column the cluster with the lowest probability. Next, we use the same clustering technique used to find the resource profiles. In this way, we can discern "specialists" – i.e. resources which mainly focus on a selective set of activity instance archetypes – from "generalists" – i.e resources who divide their time over more archetypes [5].

## 4   Demonstration

In this demonstration, we will validate whether the application of ResProMin is feasible on real-life data and capable of finding valuable business insights. To this end, we used the publicly available event logs of the 2015 BPI Challenge, which describe the process of building permit applications of five Dutch municipalities [18]. These five logs contain information about the performed activities with the associated resource, as well as other case-related attributes.

Section 4.1 highlights how the three steps in our method are applied. Section 4.2 will discuss the results for municipality 1. Due to space restrictions, the other municipalities' results, along with the code used to fit the FMMs, can be consulted in an online appendix[3]. Section 4.3 will discuss the findings across municipalities and compare whether the same process is organised differently.

### 4.1   Setup

**Step 1: Context Enrichment**  In Step 1, the event log is enriched with contextual factors. Table 2 shows an overview of the attributes used in the cluster model. Some attributes were already present in the event log; others have been derived from existing attributes. For instance, the weekday is derived from the event's timestamp.

The activity attribute contains many distinct activity labels (on average, each log contains over 280 different labels). To obtain interpretable results, and in the absence of domain knowledge to compose meaningful groups of activity labels, we opted to use the "phase" as the activity label. To determine the phase, the

---

[3] https://doi.org/10.5281/zenodo.4606757

code of an activity instance is used, e.g. "01_HOOFD_xxx" refers to an activity in the first phase [18]. It should be stressed that we did not remove any events while abstracting the activity label, e.g. when five different activities of phase 0 were executed, we referred to each of these activities as "phase 0".

**Table 2.** BPIC'15 attributes with description. Attributes with an asterisk (*) have been derived from existing attributes.

| Attribute | Description |
|---|---|
| Phase* | The phase within the process. Derived from the "concept.name" attribute, where the first digit of the last part expresses a phase within the process. A total of nine phases are present: Phase 0–8 |
| Resource | The unique identifier of the resource who executed this activity instance, e.g. "560462" |
| Case Procedure | Either blank (no value), "Regulier" (regular), or "Uitgebreid" (comprehensive) |
| Case Status | Either "G" or "O". We filtered out "T", because this only applied to two cases across all logs |
| Weekday* | Number indicating the day of the week, starting with "1" for Monday. Derived from the timestamp indicating when the activity was completed |
| Case Parts* | The category/ies the application relates to. Derived from the "(case) parts" attribute and transformed into dummy variables. An activity instance applies to at least one category, but multiple categories could be applicable. Some categories were aggregated to limit the number of variables, e.g. everything related to environment was bundled into one dummy "Environment" |

**Step 2: Activity Instance Archetype Identification** To determine the appropriate number of clusters of the FMM, we decided to fit two to ten components on each log as considering even more components would hamper the interpretability. Each model was repeated five times to limit the risk of finding a local optimum. The stability of the results across repetitions confirmed that five runs per component were sufficient. This resulted in a total of 45 models per municipality: nine potential numbers of clusters, each with five repetitions. We fitted the mixture models using the R-package *flexmix* (version 2.3-17) [7] (R version 3.6.1). It took, on average, 3.7 minutes for a model to converge to a solution.

To decide for each municipality which of the 45 models to select, we applied three rules: (1) per number of components, we selected the repetition with the highest log-likelihood, (2) we looked where the BIC-curve showed an "elbow" pattern: adding more clusters would make the model more complex, harder to interpret, and barely improves the model, and (3) no cluster should become smaller than 5% of all observations. This resulted in 7, 8, 7, 6, and 9 clusters for municipality 1–5, respectively. For example, Figure 2 shows the evolution of

the BIC when adding more clusters to the model of municipality 1. An "elbow" pattern can be spotted at seven clusters. The BIC could be slightly improved by adding one additional cluster. However, this would make the second cluster smaller than 5%, violating our third rule.
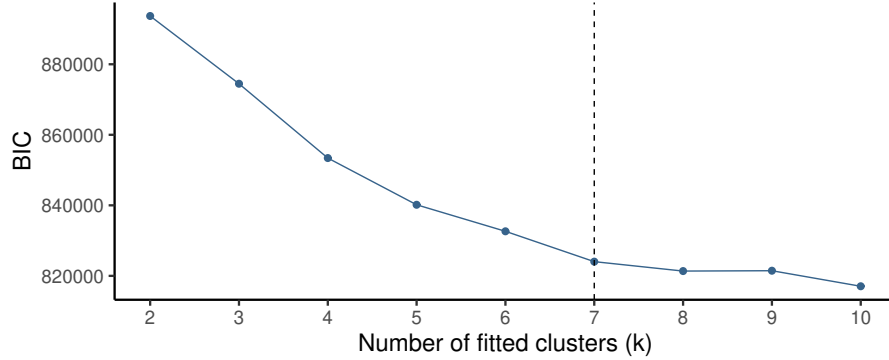


**Fig. 2.** BIC evolution when adding more clusters to the model of municipality 1.

**Step 3: Resource Profile Identification** In the final step, we apply Agglomerative Hierarchical Clustering to discover the context-aware resource profiles and find groups of resources with similar degrees of specialism.

### 4.2    Intra-Municipality Results for Municipality 1

The results of the fitted parameters of the FMM (Step 2) for municipality 1 are shown in Table 3. In Table 3a we see the intra-cluster phase distribution. For instance, cluster 3 mainly (88.27%) contains activities from phase 0, while cluster 4 mainly focuses on phase 4 and 5. If we add up each probability from largest to smallest until we reach a threshold of 70% for each cluster, we could identify the most dominant phases for each activity instance archetype, e.g. in cluster 4 this would be phase 4 and 5.

If we look at the case procedures in Table 3b, we notice that all clusters describe activity instance groups with a "blank" case procedure, except for the first cluster, which is more likely to contain activity instances that required a comprehensive procedure.

Regarding case status in Table 3c, clusters 1, 2, and 7 mainly contain activities with case status "G", whereas cluster 6 is more likely to have an "O" status. The case status in clusters 3, 4, and 5 is evenly spread among "G" and "O".

The probabilities of observing an event on a particular part of the week for each cluster is given in Table 3d. We aggregated the probabilities of Monday, Thursday, and Friday to "Beginning/end of week"; Tuesday and Wednesday to "Midweek"; and Saturday and Sunday to "Weekend". This makes the relation

between the cluster and weekday more distinct. We notice that clusters 1, 2, and 3 are mainly performed during the beginning/end of the week, whereas the others are more spread out over the working week. In addition, it is improbable to observe an event during weekends, which is not surprising within the context of a permit application process.

Table 4 shows the probability of observing a particular case part (or category). As multiple labels might apply to an activity instance, the sum of all labels does not equal 100%, in contrast to the previous attributes. We notice that clusters 3, 4, and 5 predominantly concerns an application related to construction. Cluster 2 is always related to tree felling, 1 always to environment, and 7 predominantly to demolition. In cluster 6, there is not really one category that is mainly related to all activity instances of this archetype. Therefore, we refer to this archetype as a residual archetype.

Table 5 describes the six identified activity instance archetypes based on the insights from Table 3 and Table 4, together with the relative size of the cluster to the entire log. For instance, activity instance archetype 5 ("Other construction cases") is the largest cluster which applies to around a third (34.19%) of all events recorded for municipality 1.

The *input* columns of Table 6 show the result of applying Bayes' Theorem in the third step, i.e. the conditional probability of executing an activity instance from a particular activity instance archetype, given a specific resource. We can look at these probabilities from two different angles. Firstly, we could look for resources that work on the same activity instance archetypes, i.e. resource profiles. We cluster the input columns of Table 6 using AHC into seven clusters as the Total Within Sum of Square plot in Figure 3a shows the typical "elbow" pattern there. Figure 3b shows the resulting resource profiles, which are also labelled in the *output* column in Table 6. For instance, resources "4936828", "560462", and "560950" mainly perform activity instances from activity instance archetype 1. Therefore, we refer to this profile as resources that work on "environmental cases". "Tree felling" (cluster 2) is mainly executed by "560872" and "5726485". However, as tree felling is a relatively small archetype (only 7.83% of the complete event log), these resources likely have to fill their remaining time with other work, such as construction-related activity instances.

Secondly, we could focus on whether a resource is a "specialist" or "generalist". We transformed the input columns of Table 6 so that the probabilities of belonging to a particular cluster are ordered from left to right, i.e. the first column contains the cluster with the highest probability and the last column the cluster with the lowest probability. We use the same clustering technique used for discovering the resource profiles to find six degrees of specialism. For instance, resource "4936828" always works on activity instances from archetype 1, whereas "560999" always works on archetype 6. We could say that they are both specialised in their work, but they do not do the same things. In contrast, resource "560464" more evenly spreads his/her time among clusters 3, 4, 5, and 7. This resource clearly does not specialise in a particular activity instance archetype.

**Table 3.** Intra-cluster distributions for phase, case procedure and status, and weekday variables for municipality 1 (in %).

(a) Intra-cluster phase distribution for municipality 1.

| Phase | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 |
|---|---|---|---|---|---|---|---|
| Phase 0 | 50.89 | 31.43 | 88.27 | 18.89 | 35.47 | 28.19 | 35.56 |
| Phase 1 | 12.64 | 11.87 | 11.35 | 2.95 | 19.50 | 11.62 | 18.91 |
| Phase 2 | 4.23 | 6.03 | <0.01 | 1.62 | 11.61 | 7.65 | 10.66 |
| Phase 3 | 4.12 | 7.00 | <0.01 | 1.44 | 13.04 | 8.86 | 9.20 |
| Phase 4 | 13.83 | 17.22 | <0.01 | 25.73 | 18.97 | 25.59 | 14.83 |
| Phase 5 | 7.72 | 15.19 | 0.02 | 46.04 | 1.19 | 16.28 | 10.23 |
| Phase 6 | 0.79 | <0.01 | <0.01 | 0.05 | 0.05 | 0.08 | <0.01 |
| Phase 7 | 1.82 | 3.45 | 0.11 | 0.93 | 0.08 | 0.40 | 0.52 |
| Phase 8 | 3.95 | 7.81 | 0.25 | 2.34 | 0.09 | 1.33 | 0.09 |

(b) Intra-cluster case procedure distribution for municipality 1.

| Procedure | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 |
|---|---|---|---|---|---|---|---|
| *blank* | 35.98 | 100.00 | 93.62 | 89.41 | 96.65 | 91.85 | 94.32 |
| Regular | <0.01 | <0.01 | 1.55 | 3.44 | 1.50 | 0.88 | 2.26 |
| Comprehensive | 64.02 | <0.01 | 4.84 | 7.15 | 1.85 | 7.27 | 3.42 |

(c) Intra-cluster case status distribution for municipality 1.

| Status | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 |
|---|---|---|---|---|---|---|---|
| G | 81.21 | 89.13 | 47.62 | 48.29 | 50.68 | 26.86 | 85.43 |
| O | 18.79 | 10.87 | 52.38 | 51.71 | 49.32 | 73.14 | 14.57 |

(d) Intra-cluster part of week distribution for municipality 1. The individual probabilities for each weekday can be consulted in the online appendix.

| Part of week | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 |
|---|---|---|---|---|---|---|---|
| Beginning/end of week | 79.79 | 89.76 | 86.34 | 43.30 | 59.13 | 51.58 | 48.45 |
| Midweek | 20.13 | 10.21 | 13.20 | 56.67 | 40.83 | 48.42 | 51.54 |
| Weekend | 0.08 | 0.02 | 0.46 | 0.03 | 0.04 | <0.01 | 0.02 |

**Table 4.** Intra-cluster case part distribution for municipality 1 (in %). Note that unlike the variables in the Table 3, we assume that the values of case part are independent, i.e. an observation may have multiple case parts. Therefore, the summation over case parts does not add up to 100%.

| Case Part | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 |
|---|---|---|---|---|---|---|---|
| Installation | <0.01 | 1.65 | 5.60 | 0.39 | 0.50 | 28.96 | <0.01 |
| Construction | 21.24 | <0.01 | 83.70 | 100.00 | 100.00 | 17.49 | 16.38 |
| Fireproof | 0.74 | <0.01 | 2.96 | 2.12 | 0.40 | 13.12 | <0.01 |
| Flora & Fauna | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| Area protection | 4.81 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| Acting in violation of SPR | 5.40 | <0.01 | 15.43 | 8.43 | 7.81 | 45.98 | 0.49 |
| Tree felling | 0.24 | 100.00 | 1.17 | 1.24 | 0.47 | 3.72 | 3.18 |
| Entrance/exit | <0.01 | <0.01 | 1.54 | 0.77 | 0.76 | 4.95 | <0.01 |
| Environment | 100.00 | <0.01 | 6.04 | 5.76 | 6.38 | 0.84 | <0.01 |
| Monument | <0.01 | <0.01 | 4.41 | 9.17 | 3.91 | 3.97 | 6.99 |
| Advertisement | <0.01 | <0.01 | 1.77 | 1.10 | 0.85 | 6.57 | <0.01 |
| Demolition | 0.30 | <0.01 | 4.16 | 4.39 | 3.04 | 0.01 | 98.42 |

SPR = Spatial Planning Rules

**Table 5.** Activity instance archetypes with descriptions for municipality 1.

| Cluster | Label | Description | Size (%) |
|---|---|---|---|
| 1 | Environmental cases | Mainly occur at the beginning/end of the week, more likely require a comprehensive procedure, with typically a "G" case status | 8.97 |
| 2 | Tree felling cases | Mainly occur at the beginning/end of the week, always have a "blank" procedure, with typically status "G" | 7.83 |
| 7 | Demolition cases | Occur evenly across the week, typically have a "blank" procedure with status "G" | 5.20 |
| 3 | Construction cases in phase 0 | Mainly occur at the beginning/end of the week, typically have a "blank" procedure, and are evenly split between status "G" and "O" | 17.95 |
| 4 | Construction cases in phase 4 and 5 | Occur evenly across the week, typically have a "blank" procedure, and are evenly split between status "G" and "O" | 17.77 |
| 5 | Other construction cases | Occur evenly across the week, typically have a "blank" procedure, and are evenly split between status "G" and "O" | 34.19 |
| 6 | Other cases | Occur evenly across the week, typically have a "blank" procedure with status "O" | 8.09 |
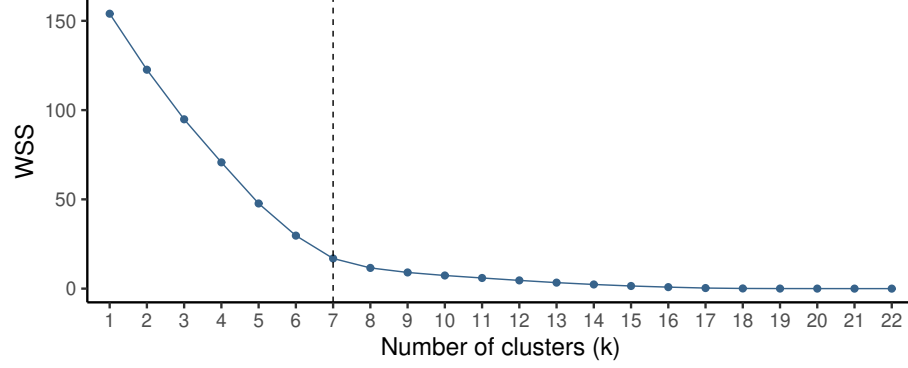
(a) Total Within Sum of Square (WSS) plot to determine the number of clusters. In this case, we choose seven clusters.
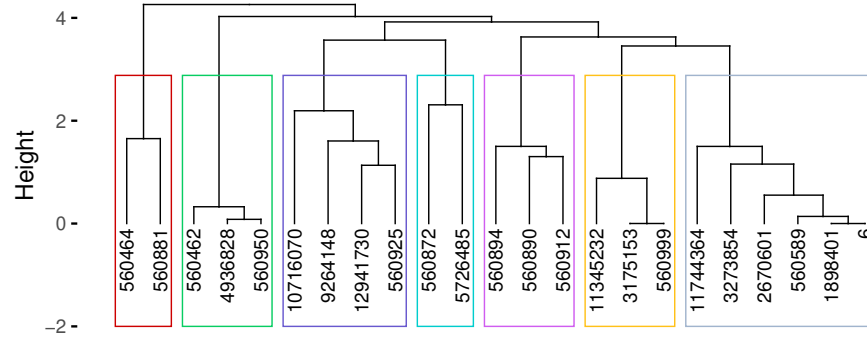


(b) Cluster dendrogram of the resource profiles. From left to right: demolition and other construction cases; environmental cases; construction cases in phase 0; tree felling and construction cases; construction cases in phase 4 and 5, and other cases; other cases; and other construction cases.

**Fig. 3.** Resources working on the same activity instance archetypes in municipality 1.

**Table 6.** Probabilities for each resource to belong to a particular cluster in municipality 1 (in %). The profiles are the results of clustering the table using AHC.

| Resource | Input | | | | | | | Output |
|---|---|---|---|---|---|---|---|---|
| | Clust1 | Clust2 | Clust3 | Clust4 | Clust5 | Clust6 | Clust7 | Profile |
| 10716070 | <0.01 | <0.01 | **63.72** | 35.85 | 0.42 | <0.01 | <0.01 | |
| 12941730 | <0.01 | <0.01 | **100.00** | <0.01 | <0.01 | <0.01 | <0.01 | Construction cases in |
| 560925 | 3.93 | 6.26 | **81.64** | 0.05 | 0.12 | <0.01 | 8.00 | phase 0 |
| 9264148 | 2.22 | 16.91 | **80.87** | <0.01 | <0.01 | <0.01 | <0.01 | |
| 11345232 | <0.01 | <0.01 | 20.00 | <0.01 | <0.01 | **80.00** | <0.01 | |
| 3175153 | <0.01 | <0.01 | 0.03 | <0.01 | <0.01 | **99.97** | <0.01 | Other cases |
| 560999 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | **99.99** | <0.01 | |
| 11744364 | <0.01 | <0.01 | <0.01 | <0.01 | **59.59** | 40.41 | <0.01 | |
| 1898401 | <0.01 | <0.01 | 0.03 | <0.01 | **99.97** | <0.01 | <0.01 | |
| 2670601 | 0.10 | 0.33 | 0.09 | 0.03 | **85.21** | 13.02 | 1.22 | Other construction |
| 3273854 | <0.01 | <0.01 | 2.08 | 0.01 | **80.24** | 5.43 | 12.25 | cases |
| 560589 | <0.01 | <0.01 | <0.01 | 2.59 | **97.40** | <0.01 | <0.01 | |
| 6 | <0.01 | <0.01 | 0.01 | <0.01 | **99.99** | <0.01 | <0.01 | |
| 4936828 | **99.99** | <0.01 | 0.01 | <0.01 | <0.01 | <0.01 | <0.01 | |
| 560462 | **94.98** | 2.72 | 2.30 | <0.01 | <0.01 | <0.01 | <0.01 | Environmental cases |
| 560950 | **98.53** | <0.01 | <0.01 | 1.47 | <0.01 | <0.01 | <0.01 | |
| 560464 | <0.01 | <0.01 | 20.20 | 22.57 | **22.68** | <0.01 | **34.56** | Demolition and other |
| 560881 | 0.29 | 0.54 | 0.37 | <0.01 | **53.19** | 2.29 | **43.31** | construction cases |
| 560872 | 18.03 | **29.06** | **50.86** | <0.01 | <0.01 | <0.01 | 2.05 | Tree felling and |
| 5726485 | <0.01 | **34.17** | <0.01 | 12.54 | **53.29** | <0.01 | <0.01 | construction cases |
| 560890 | 0.17 | <0.01 | <0.01 | **75.95** | <0.01 | **17.83** | 6.06 | Construction cases in |
| 560894 | <0.01 | <0.01 | 4.77 | **44.31** | <0.01 | **42.38** | 8.54 | phase 4 and 5, and |
| 560912 | 5.16 | 2.94 | 18.62 | **54.00** | 4.71 | **12.50** | 2.08 | other cases |

Table 7 tabulates the number of resources for each profile-specialism combination. The degree of specialism is ordered from left – "pure specialist" – to right – "pure generalist". For instance, we notice that environmental cases are only performed by resources with the highest specialisation degree.

### 4.3    Inter-Municipality Results

In the previous subsection, we discussed the finding of applying ResProMin in municipality 1. We found similar patterns in the other municipalities, e.g. all municipalities have an archetype for environmental cases. The construction cases were present as well, but not always with a focus on the same phases. In addition, each municipality has several specialists and generalists.

However, we also found some differences between the municipalities. Firstly, only municipality 1 exhibited the pattern where some activity instance archetypes were mainly performed during either the end or the beginning of the week. Instead, a frequently observed pattern in the other municipalities was a much lower conditional probability to observe a particular activity instance archetype

**Table 7.** Number of resources per profile-specialism combination in municipality 1.

| Profile | Specialism Group | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Construction cases in phase 0 | **3** | **1** | 0 | 0 | 0 | 0 |
| Other cases | **3** | 0 | 0 | 0 | 0 | 0 |
| Other construction cases | **5** | **1** | 0 | 0 | 0 | 0 |
| Environmental cases | **3** | 0 | 0 | 0 | 0 | 0 |
| Demolition and other construction cases | 0 | **1** | 0 | 0 | 0 | **1** |
| Tree felling and construction cases | 0 | 0 | **2** | 0 | 0 | 0 |
| Construction cases in phase 4 and 5, and other cases | **1** | 0 | 0 | **1** | **1** | 0 |

on Friday. In other words, Fridays seemed to be quieter than other weekdays. In municipality 1, Wednesday was often the quieter day. Secondly, the more resources a municipality has – most likely bigger municipalities – the larger the proportion of resources seems to specialise in particular archetypes, as shown in Table 8. This offers face validity to our method as it seems reasonable that when there are more resources to divide the work among, there is more room to specialise. However, an interesting exception is that the smallest municipality (i.e. 4) has the second-highest specialist rate of all municipalities. This might indicate that municipality 4 uses a different way of handling the permit application process. Thirdly, activity instance archetypes requiring a comprehensive procedure, typically related to environment (such as cluster 1 in municipality 1, as described in Table 5) are more likely to have specialised resources involved. Finally, activity instance archetypes that involve predominantly construction-related activity instances are also more likely to have specialised resources involved, albeit less clearly than the comprehensive environment archetype.

**Table 8.** Number and proportion of specialists in each municipality.

| Municipality | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Specialised resources | 18 | 5 | 8 | 7 | 12 |
| Total resources | 23 | 11 | 14 | 10 | 21 |
| Proportion specialised | 78% | 45% | 57% | 70% | 57% |

## 5   Conclusion

In this paper, we extend the existing work on organisational mining by introducing our method, *ResProMin*. In contrast to the state-of-the-art, ResProMin is capable of finding context-aware resource profiles based on the notion of activity instance archetypes. Instead of solely considering activity labels to group resources, ResProMin accommodates contextual information such as case attributes and variables capturing the system state. In addition, our method allows

activities to belong to multiple profiles simultaneously and is capable of discerning specialists from generalists. Our demonstration confirms the feasibility of our method to discover context-aware resource profiles from real-life event logs. This provides rich insights to process owners, which can help them manage their resources better by uncovering, e.g. (potentially implicit) task division patterns.

Besides these contributions, we also acknowledge some limitations of our method. Firstly, estimating a Finite Mixture Model's parameters is a computationally demanding process and suffers from *the curse of dimensionality*. However, this study's focus was on demonstrating whether our method is capable of uncovering meaningful resource-related insights that are valuable in a business context and not on optimising its execution. Moreover, this kind of analysis is typically not performed in real-time, supporting that runtime optimisation will not be the primary goal as long as execution times remain practically feasible. Secondly, we had no access to domain experts in the municipalities to validate and elaborate more on our findings. Nevertheless, our demonstration shows that ResProMin is capable of finding interesting and valuable insights into the prevailing resource profiles.

We identify several directions for future work. Firstly, heuristics could be developed to improve our method's computational efficiency while still obtaining near-optimal solutions. For instance, a quasi-Newton approach could be adopted to accelerate the convergence of the EM algorithm [12]. Secondly, instruments to facilitate the enrichment of an event log with context-related information can be developed. Thirdly, it could be investigated whether different resource-related organisations between municipalities are associated with process performance differences. Finally, we could determine how the insights of ResProMin can be leveraged by models which require fine-grained resource allocation information, such as Business Process Simulation models.

## Acknowledgements

## References

1. Appice, A.: Towards Mining the Organizational Structure of a Dynamic Event Scenario. Journal of Intelligent Information Systems **50**(1), 165–193 (2018)
2. Burattin, A., Sperduti, A., Veluscek, M.: Business Models Enhancement Through Discovery of Roles. In: Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining. pp. 103–110. CIDM'13 (2013)
3. Cabanillas, C., Ackermann, L., Schönig, S., Sturm, C., Mendling, J.: The RALph Miner for Automated Discovery and Verification of Resource-Aware Process Models. Software and Systems Modeling (2020)

4. Camargo, M., Dumas, M., González-Rojas, O.: Automated Discovery of Business Process Simulation Models from Event Logs. Decision Support Systems **134** (2020)
5. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer-Verlag, Berlin, Germany, second edn. (2018)
6. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. No. 692 in Springer Series in Statistics, Springer, New York, NY, USA (2006)
7. Grün, B., Leisch, F.: FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. Journal of Statistical Software **28**(1), 1–35 (2008)
8. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis: A Global Perspective. Pearson Education, Upper Saddle River, NJ, USA, seventh edn. (2010)
9. Jin, T., Wang, J., Wen, L.: Organizational Modeling from Event Logs. In: Proceedings of the 6th International Conference on Grid and Cooperative Computing. pp. 670–675. GCC'07 (2007)
10. Masyn, K.E.: Latent Class Analysis and Finite Mixture Modeling. In: The Oxford Handbook of Quantitative Methods in Psychology, vol. 2: Statistical Analysis, pp. 551–611. Oxford University Press, New York, NY, USA (2013)
11. McLachlan, G.J.: Model-Based Clustering. In: Comprehensive Chemometrics, vol. 2, pp. 655–681. Elsevier, Oxford, UK (2009)
12. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. No. 1345 in Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, second edn. (2007)
13. Nakatumba, J., van der Aalst, W.M.P.: Analyzing Resource Behavior Using Process Mining. In: Proceedings of the Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 43, pp. 69–80 (2010)
14. Pika, A., Leyer, M., Wynn, M.T., Fidge, C.J., ter Hofstede, A.H.M., van der Aalst, W.M.P.: Mining Resource Profiles from Event Logs. ACM Transactions on Management Information Systems **8**(1), 1:1–1:30 (2017)
15. Reijers, H.A., Liman Mansar, S.: Best Practices in Business Process Redesign: An Overview and Qualitative Evaluation of Successful Redesign Heuristics. Omega **33**(4), 283–306 (2005)
16. Schönig, S., Cabanillas, C., Di Ciccio, C., Jablonski, S., Mendling, J.: Mining Team Compositions for Collaborative Work in Business Processes. Software & Systems Modeling **17**(2), 675–693 (2018)
17. Song, M., van der Aalst, W.M.P.: Towards Comprehensive Support for Organizational Mining. Decision Support Systems **46**(1), 300–317 (2008)
18. van Dongen, B.F.: BPI Challenge 2015. 4TU.ResearchData (2015)
19. Vermunt, J.K., Magidson, J.: Latent Class Cluster Analysis. In: Applied Latent Class Analysis, pp. 89–106. Cambridge University Press, Cambridge, UK (2002)
20. Yang, J., Ouyang, C., Pan, M., Yu, Y., ter Hofstede, A.H.M.: Finding the "Liberos": Discover Organizational Models with Overlaps. In: Proceedings of the 16th International Conference on Business Process Management. Lecture Notes in Computer Science, vol. 11080, pp. 339–355 (2018)