Profiling Event Logs to Configure Risk Indicators for Process Delays

Anastasiia Pika[1], Wil M.P. van der Aalst[2,1], Colin J. Fidge[1], Arthur H.M. ter Hofstede[1,2], and Moe T. Wynn[1]

[1] Queensland University of Technology, Brisbane, Australia
{a.pika,c.fidge,a.terhofstede,m.wynn}@qut.edu.au
[2] Eindhoven University of Technology, Eindhoven, The Netherlands
w.m.p.v.d.aalst@tue.nl

Abstract. Risk identification is one of the most challenging stages in the risk management process. Conventional risk management approaches provide little guidance and companies often rely on the knowledge of experts for risk identification. In this paper we demonstrate how risk indicators can be used to predict process delays via a method for configuring so-called Process Risk Indicators (PRIs). The method learns suitable configurations from past process behaviour recorded in event logs. To validate the approach we have implemented it as a plug-in of the ProM process mining framework and have conducted experiments using various data sets from a major insurance company.

Keywords: process risk indicators, process mining, risk identification.

1 Introduction

Managing risks is one of the top priorities in corporate and government organisations[1]. According to ISO Guide 73:2009, risk is the "effect of uncertainty on objectives" where an effect is "a deviation from the expected — positive and/or negative" [6]. Risk identification is an essential starting point for risk management. It is defined as a "process of finding, recognizing and describing risks" [6]. Although many risk management approaches provide high-level guidance about risk management strategy, they do not provide any tools to operationalize this strategy [12,15]. Standard ISO 31000 specifies that "risk identification can involve historical data" [15], however it does not provide any further guidelines on how to use historical data.

Managing business processes is another important concern of an organisation. Business processes are exposed to different risks. For instance, a process may not be finished within the time-frame defined by a service level agreement, it may produce low-quality results, or it may exceed its budget. We refer to risks that threaten the achievement of process goals as process-related. Most organisations

[1] http://www.gartner.com/id=1957716, Gartner report "CEO Survey 2012: CIOs Must Link Risk Management and Compliance to Business Priorities"

C. Salinesi, M.C. Norrie, and O. Pastor (Eds.): CAiSE 2013, LNCS 7908, pp. 465–481, 2013.

use information systems supporting their operational business processes. Often these systems also record information about process executions in event logs. Our belief is that this information can be exploited for the identification of process-related risks.

In our preliminary work [14] we introduced the idea of using *Process Risk Indicators* (PRIs) to predict whether a deadline transgression is likely to happen or not. For example, if an activity is repeated multiple times for a case, then the likelihood of delay is significantly higher. We also introduced a method for instantiating these indicators from event logs based on statistical techniques for outlier detection. However, our initial experiments showed that further work is required to properly calibrate the indicators to reduce the number of "false positives", i.e., cases that are predicted to be late but in the end are not. In this paper we present a novel method for configuration of PRIs that uses information about outcomes from cases executed in the past. The method aligns these indicators with the specifics of a particular process to minimize the number of false positives. We demonstrate the feasibility of the proposed approach using case studies with data sets from an Australian insurance company (Suncorp).

The remainder of the paper is organized as follows. Section 2 discusses related work. The general approach to process risk identification is presented in Section 3 followed by a description of eight PRIs. We then show how to configure these PRIs using information about the outcomes of cases in the past. Section 4 discusses our implementation in ProM and reports on our experimental results. Section 5 concludes the paper.

2 Related and Previous Work

Few approaches exist that aim to identify and/or assess process risks [7,8,21]. Wickboldt et al. proposed a framework that uses process execution data for risk assessment [21]. Risk assessment modules of the framework use information about risk events reported during past activity executions. Our approach also predicts future risks based on past behaviours, but it does not require risk-related information to be explicitly stored. Jallow et al. [7] proposed a framework for identification of operational process risks. However, estimation of the probabilities and impacts associated with risk-related activities was assumed to be done by experts. Our approach avoids subjective opinions and learns such values from historic event data. Jans et al. [8] proposed using process mining for the identification of one particular type of risk (transactional fraud risk) and showed that available process mining tools can help auditors detect fraud. By contrast, our approach focuses on quantifiable values such as delays or product quality and it emphasises automatable techniques for risk identification that can be used for run-time operational support [16].

Van Dongen et al. proposed an approach for predicting the remaining cycle time of a case by applying non-parametric regression and using case data as predictor variables [20]. The approach for predicting remaining process time proposed by van der Aalst et al. [18] is based on building an annotated transition

system and estimating the average remaining time of cases that visited the same state previously. In contrast, our approach predicts the likelihood of case delay rather than the remaining execution time.

Grigori et al. presented a set of integrated tools that help manage process execution quality supporting such features as analysis and prediction [3]. In other work they propose a method for exception analysis, prediction, and prevention [4]. A common feature of these approaches is that it is the responsibility of users to specify what process properties (conditions, exceptions etc.) they would like to analyse. Our approach does not require such input and is based on a set of risk indicators.

In our own earlier work we introduced the idea of using Process Risk Indicators for predicting case delays and proposed a method for instantiation of the indicators from event logs [14]. The method is based on statistical techniques for outlier detection. It used a simple analysis which assumed that process behaviours have normal distributions with fixed thresholds being sufficient to identify "risky" behaviours. Our initial experiments revealed that risk indicators can be used to predict case delays [14], but further work is required to properly calibrate the indicators to reduce the number of false positives. In this paper we present a method for configuration of risk indicators for process delays that significantly improves precision of case delays predictions.

3 Risk Identification Method

3.1 Approach

Our goal is to develop a method that can identify the risk of delay for running cases with a high degree of precision. Our method analyses characteristics of a current case, compares them with characteristics of similar cases executed in the past and predicts a case delay if a "risky" behaviour is detected. Our overall approach consists of three major steps: (1) define Process Risk Indicators; (2) configure PRIs; (3) identify the presence of PRI instances in a current case.

First, we need to identify which behaviour of a process can be considered "risky". In our initial work we introduced the use of Process Risk Indicators (PRIs) for predicting case delays. We defined a PRI as "a pattern observable in an event log whose presence indicates a higher likelihood of some process-related risk" [14]. For example, an unusually large number of activity repetitions per case may indicate a likely case delay or low-quality output because there seems to be a problem processing this case.

In our preliminary work we also introduced a method for identifying the presence of a PRI based on the "sample standard deviations" approach for outlier detection [14]. For each PRI we defined cut-off thresholds as $\overline{x} + 2s$. Observations whose values are higher than this value were considered outliers. A limitation of the method is the assumption that some particular process behaviour follows a normal distribution (e.g., activity repetitions in a case) which may not be valid in many cases. We also assumed that atypical behaviour of a process can be

considered "risky", e.g. when some activity in a case has an atypically long duration it signals a higher likelihood of the case delay. However, while conducting initial experiments we learned that though atypical behaviour is often associated with case delays it is not always "risky". For example, if a process contains an automated activity which typically takes a very small amount of time compared to the total time that cases take, then variations to the execution time of such an activity, even relatively large ones, do not affect the case duration.

To overcome these weaknesses of our initial work we present here a method for configuration of indicators so that the specific characteristics of a particular process are taken into account. We again use cut-off thresholds to identify "risky" behaviours, however we introduce a way of learning the threshold values by using information about outcomes of cases in the past. The method allows us to identify atypical process behaviour that has been often associated with case delays in the past rather than assuming *any* outlier indicates a risk.

3.2 Process Risk Indicators (PRIs)

A PRI is a pattern that signals an increased likelihood of some process-related risk and which can be identified by analysing an event log. In our previous work [14] we introduced the idea of using Process Risk Indicators to identify the risk of case delay. For the purpose of this paper we use several indicators that can be discovered using basic event logs, information about case outcomes and process models, all of which were available to us in our industrial case study. Below we define eight Process Risk Indicators for process delays.

PRI 1: Atypical activity execution time. The duration of an activity significantly exceeds its typical duration. An activity may take more time than usual due to human factors: an employee executing the activity may be inexperienced or occupied with many other tasks. Fatigue is a common factor that may cause a delay. Another reason can be a complex or exceptional case that requires additional investigation/learning. Activity delay is also often caused by a third party's involvement—reducing the number of contacts with third parties is one of Business Process Re-engineering's best practices [11].

PRI 2: Atypical waiting time. An activity has not been started for an atypically long period of time. One possible explanation for long waiting times is a lack of available resources. Another possible reason is the "too hard basket" syndrome, i.e., the situation where no one is willing to start an activity as it is perceived to be too challenging or time consuming. Also, some employees tend to process certain tasks in batches, which may increase a particular task's waiting time. A typical example is an approval task. Removing batch-processing is another of the BPR best practices [11], as is reducing waiting times because these often occupy 95% of the throughput time of a case [9].

PRI 3: Multiple activity repetitions. The number of times an activity is repeated in a case significantly exceeds its usual value. It may be necessary

to repeat an activity if previous attempts fail. This can happen due to third party involvement, e.g., not receiving an expected service from subcontractors or failure to provide required information by a client. Employees may also need to repeat a task because of inexperience or complex case requirements.

PRI 4: Presence of a "risky" activity. A case contains a "risky" activity. An activity is considered "risky" if the majority of the cases that contained this activity in the past have been delayed. Execution of a "risky" activity may be related to a case's specifics. For example, consultation with an expert or a manager may be required for an exceptionally complex case.

PRI 5: Multiple resource involvement. More resources are involved in a case than usually. One possible reason for such a situation is the so-called "hot potato" phenomenon where a case is forwarded between different resources because nobody is willing to take charge of it. Atypically high resource involvement can also be needed for a very complex case. Reducing the number of parties involved in a case is another of the BPR best practices [11]. Balasubramanian et al. name frequent hand-overs of work between people in a process as one of the factors that can lead to time overruns [2].

PRI 6: Atypical sub-process duration. The sum of activity duration and its waiting time in a case (referred to here as a sub-process) is significantly higher than its typical value. We introduce this indicator to be able to work with event logs that only record "complete" events for activities, as is often the case for real event logs. This indicator tackles the same issues as PRIs 1 and 2.

PRI 7: High resource workload. An activity has been assigned to or started by a resource with a high workload. The workload of a resource at a point in time is the number of items that were started by or assigned to the resource but not yet completed. High resource workload is often mentioned in the literature as a reason for such risks as time overruns or low-quality outputs [5,13].

PRI 8: Use of a "risky" resource. An activity has been assigned to or started by a"risky" resource. A "risky" resource for some activity is the one that was often involved in execution of this activity in delayed cases. Some human resources may be incompetent or inexperienced when it comes to the execution of some activities in a process. It is important to use recent data for identification of this PRI as the qualification levels and experience of resources will change over time. Another reason for a resource to be considered risky is a tendency to postpone execution of certain activities, e.g., approval tasks.

3.3 Configuring Process Risk Indicators

Our method for configuration of indicators requires information about known outcomes from cases that happened in the past, i.e., whether they were delayed or completed in time. We aim to find for the PRIs the values of parameters that

could predict delays with a required degree of precision in the past. If we cannot detect such values for an indicator then it is not used for a particular process.

An input parameter to our method is a desired precision level. Precision is the fraction of cases predicted to be delayed that are actually delayed. Increasing precision is usually done at the expense of decreasing recall, which is defined as the fraction of delayed cases that can be successfully predicted against the actually delayed cases. If a user deals with a critical process, he may prefer monitoring alerts with lower precision levels in order to increase recall, while for a non-critical process he may want to check only those alerts that indicate a very high likelihood of a case delay.

For each relevant process behaviour (e.g., the number of activity repetitions in a case) we look for the smallest value that allows distinguishing between *delayed* and *in time* cases with a required degree of precision. This value is used as a cut-off threshold. In order to define this threshold we need to check the effectiveness of various candidate values. However, there could be a wide range of these. Analysing past logs can be time consuming, so in order to reduce the search space we learn cut-off thresholds for the PRIs by checking only those values from a pool of selected candidates. We use the following heuristic to define candidate values. First, we discard those values lower than the mean $\overline{x}$ (which gives us a measure of central tendency). We then include those values calculated as $\overline{x} + n * s$, where s is the standard deviation (as a measure of statistical dispersion), and n is in the range of 0 to 10 with an increment of 0.25 (these values were used for the experiments, they are input parameters). We do not necessarily assume a normal distribution. Nevertheless, these conventional statistical measures provide a natural starting point for searching for thresholds. We then check all values from the defined pool of candidates.

We are interested in indicators that can predict delays during a case's execution. Therefore, while learning parameters of PRIs from past execution data, our method considers only those events that happened before a deadline, i.e., we discard activities that have been started after the deadline has been missed.

As an example of the calculation, consider PRI 5 " Multiple resource involvement". PRI 5 is a case-based PRI, i.e., it can have only one value per case and we define one cut-off threshold. In order to identify and use PRI 5 the following steps are performed:

1. Define candidate values T for the cut-off threshold t:
   (a) Identify average number of resources involved in a case before deadline ($\overline{x}$) and standard deviation s of the population.
   (b) $T \triangleq \{\overline{x} + n * s \mid n \in \{0, 0.25, 0.50, ..., 10\}\}$
2. Define the cut-off threshold t:
   For each $t_i \in T$:
   (a) Collect a subset C_{true} of the training set comprising all cases that are *delayed* and whose number of resources involved before the deadline is higher than t_i;
   (b) Collect a subset C_{false} of the training set comprising all cases that are *in time* and whose number of resources is higher than t_i;

(c) $p_i = \begin{cases} |C_{true}|/(|C_{true}| + |C_{false}|), & \text{if } (|C_{true}| + |C_{false}|) > 0 \\ 0 & \text{otherwise.} \end{cases}$

Here $t = \arg\min_{t_i \in T} (p_i > p)$, where p is the desired precision level.

3. Check the number of resources involved in the current case and alert a likelihood of a case delay if the number is higher than the value of the learned threshold t.

For activity-based PRIs such as PRI 1 ("Atypical activity execution time"), PRI 2 ("Atypical waiting time"), PRI 3 ("Multiple activity repetitions") and PRI 6 ("Atypical sub-process duration") a similar procedure is repeated for each activity to learn proper thresholds. A case can have multiple instances of an activity-based PRI, e.g., several activities may be delayed or repeated. We consider that there is a chance of a case delay if the case contains at least one instance of an activity-based PRI. For resource-based PRI 7 "High resource workload" we learn appropriate values for cut-off thresholds for each resource. If in a current case an activity is assigned to or started by a resource with a high workload (defined by the learned threshold), a case delay is more likely.

PRIs 4 and 8 do not follow the general procedure described above. These are examples of indicators that can only be identified using information about the outcomes of cases in the past. To identify PRI 4 "Presence of a risky activity" we check if there exists an activity that is executed mainly in delayed cases. For PRI 8 we check for each pair "activity-resource" if some resource's involvement in the execution of an activity mainly occurs in delayed cases. Then we check if a current case contains a "risky" activity or if an activity is assigned to a "risky" resource. Identification of such behaviour signals increased likelihood of case delay.

4 Validation Using Real Event Logs

4.1 Experimental Setup

To estimate the quality of case delay predictions by our method we use hold-out cross-validation [10]. This is a commonly used statistical practice that implies partitioning of data into two subsets, where one subset is used for initial learning (a training set), and the results are validated using the other subset (a test set). To facilitate validation of our approach we have implemented a plug-in of the process mining framework ProM 6[2]. The plug-in takes as an input two event logs. It uses one log as a training set to configure the PRIs, then it analyses cases in the other log (a test set) to identify occurrences of these PRIs. An input parameter is the expected case duration. Cases that take longer than this value are considered to be delayed. If any of the indicators is found in a case it is predicted to be delayed. We compare predicted case delays with the actual case durations and evaluate the performance of the process risk identification method by estimating the values of "precision" and "recall". These metrics

[2] `http://www.promtools.org/prom6/`

are often used in different machine learning areas to estimate performance of prediction techniques. Precision is calculated as the fraction of cases correctly predicted to be delayed against the total number of cases predicted to be delayed. Recall is calculated as the fraction of delayed cases that are successfully predicted against the number of cases that are actually delayed. These values are calculated separately for each indicator to evaluate their individual performance. We also calculate the values of precision and recall for all indicators combined to evaluate their cumulative performance.

We used two different approaches to splitting data into a training set and a test set. In one approach, we split event logs randomly, such that 75% of cases were put into a training set and 25% of cases in a test set (referred to later as a "random" split). In the other approach, cases that were completed during one period of time (four months) were put into a training set while cases that were started within the next period (two months) were put into the test set (referred to later as a "time" split). As our approach is based on learning from past execution data it is important to use large data sets for training, therefore we decided to put more data in the training set while still having enough data in the test set for meaningful validation.

Before applying our method for risk identification it is important to perform data pre-processing. Processes tend to evolve over time. To avoid learning from outdated information recent data should be used. For our experiments we picked cases that were active over the same period of six months. The algorithm should use only completed cases to properly configure PRIs, therefore partial traces representing running process instances should be filtered out. The results of any process mining algorithm depend on input data, therefore the quality of event log data is crucial [1]. For example, if event log data contains mislabelled activities, the performance of the algorithm may be affected, therefore it is important to clean event log first (e.g., filtering out mislabelled events). It is also important to separately analyse cases that are executed in different contexts that affect their durations. For example, the expected case duration may depend on the type of customer ("gold" versus "silver") or type of service ("premium" versus "normal"). If such execution contexts are known, event log data should be first split and cases that are executed in different contexts should be analysed separately.

4.2 Data Properties and Preprocessing

We evaluated our approach using two data sets from *Suncorp*, a large Australian insurance company. Both data sets represent insurance claim processes from different organisational units, referred to here as data set A and data set B. Both event logs provided by Suncorp contained only completed cases. Data set B contains cases from five departments and was split into five sets (referred to here as B1–B5) which were used in separate experiments. Each data set (A, B1–B5) was split into a training set and a test set. The training set was used by the algorithm for learning the cut-off thresholds. Cases in the test set were used for evaluating the performance of the PRIs.

We first cleaned up the data sets by filtering out cases with activities that appear only once in the whole log. In most cases, such activities were not really unique though their label was. Typically this was a consequence of combining an activity's name with the name of the employee who executed that activity. We used original unfiltered data sets to more accurately estimate resource workloads (required for PRI 7).

To more accurately estimate waiting times (for PRIs 2 and 6) we used process models. We first identified the pre-set of an activity, i.e. the set of activities that can directly precede a given activity. We then calculated the waiting time for the activity as the difference between its "start" time and the "complete" time of the last activity from its pre-set preceding it in the case. Since we did not have process models, we instead used process mining to discover them from the event logs. First we filtered the logs so that they contained only cases representing mainstream process behaviour and used these filtered logs to discover process models represented by Petri nets with one of the ProM process mining plug-ins [19]. For data set A we used 95% of the cases representing the most frequent process variants. Data sets B1–B5 proved to have a large variety of process variants. For these data sets only those cases were used for process discovery that share the same process variant with at least four other cases. These filtered logs were only used for process discovery and not in the experiments.

Suncorp's business analysts provided us with indications about what they feel should be the usual case durations for different departments. However, while analysing the event logs we realized that these expectations are not realistic as more than 50% of cases have durations higher than expected in four out of six data sets. For these data sets we therefore learned the values for typical case durations such that at least 50% of cases in a set are completed in time. These values were used in the experiments. Figure 1 shows as an example the distribution of case throughput times for data set B4. Only cases highlighted in blue are completed in time if we consider the value provided by the company's business analysts to be the typical case duration. It is very likely that the behaviour of a process is different when an explicit due date exists and is communicated to workers. However, this should not affect the performance of our method since process behaviour is still consistent across training and test data sets.

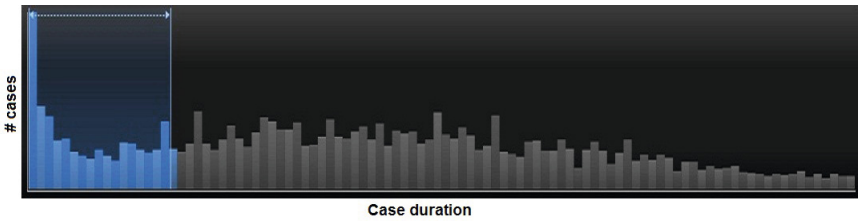


Fig. 1. Durations of cases in data set B4 (more than 50% of cases have durations higher than expected by the company)

Data set A has some nice properties which make it suitable for our experiments: a significant number of cases, steady case arrival rates and similar case duration distributions over time (low variability). Figure 2 shows some basic properties of data set A.

	SLA	Number of traces in the training set	Number of delayed cases in the training set	% of delayed cases in the training set	Number of traces in the test set	Number of delayed cases in the test set	% of delayed cases in the test set
Set A, random split	14	26903	9839	37%	8967	3311	37%
Set A, time split	14	26719	10764	40%	14500	6424	44%

Legend: SLA – usual case duration in days

Fig. 2. Properties of data set A

For data sets B1–B5 additional filtering was required. We were informed by Suncorp that cases with claim amounts higher than a certain value are considered "complex" and that it is normal for them to have long durations. We filtered the event logs for our experiments and only used "simple" cases that are expected to be completed within a certain time period. We found a large number of process variants in these sets. High variability of the processes can be explained by the fact that process models used by Suncorp are not prescriptive and are only used as guidance. High process variability may decrease precision of delay predictions for two PRIs that use information about the order of activities (PRI 2 "Atypical waiting time" and PRI 6 "Atypical sub-process duration"). The performance of other PRIs is not expected to be affected since they do not rely on the order of activity executions. Also case arrival rates, case durations, and mean active and waiting times were found to change over time. All these characteristics of the process may have influenced the results of the experiments. Figure 3 depicts basic characteristics of these five data sets.

	SLA	Number of traces in the training set	Number of delayed cases in the training set	% of delayed cases in the training set	Number of traces in the test set	Number of delayed cases in the test set	% of delayed cases in the test set
B1	30	3467	1195	34%	1155	416	36%
B2	45	354	176	50%	118	58	49%
B3	30	4166	1411	34%	1388	472	34%
B4	70	2339	1045	45%	779	359	46%
B5	120	1861	908	49%	620	322	52%

Legend: SLA – usual case duration in days

Fig. 3. Properties of data sets B1–B5

4.3 Performance of the PRIs

We first conducted our experiments with data set A. Figure 4 depicts the results of the experiments conducted with event log A using a random split and Figure 5 depicts results of the experiments using a time split. An input parameter for the algorithm is the "desired precision level". When we learn a cut-off threshold for an indicator we pick the minimum value of the threshold that allowed predicting

case delays in a training set with a given precision level. We conducted experiments for three precision levels: 95%, 90% and 80%. The columns represent results for individual PRIs. The last column represents the cumulative result for all indicators: a case is predicted to be delayed if *any* of the indicators is found in the case. For a desired precision level the first two rows represent the number of True Positives (TP) and the number of False Positives (FP) produced. These predictions are produced before expiry of the deadline. The next two rows are the number of False Negatives (FN) and the number of True Negatives (TN). $TP + FP$ is the number of cases predicted to be delayed. The *precision* is calculated as the fraction $\frac{TP}{TP+FP}$. $TP + FN$ is the number of cases actually delayed and can be used to compute the *recall* which is the fraction of delayed cases that are successfully predicted and the actually delayed cases, i.e., $\frac{TP}{TP+FN}$. Figures 4 and 5 show both precision and recall values for the test sets.

		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
95%	TP	172	527	0	0	0	504	0	192	961
	FP	11	23	0	0	0	62	0	4	86
	FN	3139	2784	3311	3311	3311	2807	3311	3119	2350
	TN	5645	5633	5656	5656	5656	5594	5656	5652	5570
	precision	94%	96%	NA	NA	NA	89%	NA	98%	92%
	recall	5%	16%	NA	NA	NA	15%	NA	6%	29%
		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
90%	TP	714	1054	0	169	0	664	0	198	1972
	FP	63	107	0	21	0	90	0	5	216
	FN	2597	2257	3311	3142	3311	2647	3311	3113	1339
	TN	5593	5549	5656	5635	5656	5566	5656	5651	5440
	precision	92%	91%	NA	89%	NA	88%	NA	98%	90%
	recall	22%	32%	NA	5%	NA	20%	NA	6%	60%
		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
80%	TP	741	1066	5	169	0	1285	0	233	2169
	FP	75	111	1	21	0	312	0	14	414
	FN	2570	2245	3306	3142	3311	2026	3311	3078	1142
	TN	5581	5545	5655	5635	5656	5344	5656	5642	5242
	precision	91%	91%	83%	89%	NA	80%	NA	94%	84%
	recall	22%	32%	0.2%	5%	NA	39%	NA	7%	66%

Fig. 4. Performance of the PRIs in data set A. "Random" split experiment

The results of the experiments for the two different types of event log split were comparable in terms of the indicators' performance. Most predictions in both cases came from PRIs 1, 2 and 6. Some delays were indicated by PRIs 4 and 8. Poorly performing indicators for this data set were PRIs 3, 5 and 7. In the vast majority of cases it was only possible to identify PRIs 3 ("Multiple activity repetitions") and 5 ("Multiple resource involvement") after the deadline was missed. One of the reasons for the poor performance of PRI 7 ("High resource workload") for this log may be the fact that we do not have all data for the process (incomplete cases were filtered out). We also assumed that resources are involved full-time in this one particular process which may not be true. Figures 4 and 5 also demonstrate the number of delays that can be predicted with these indicators for different precision levels.

		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
95%	TP	1525	677	6	18	3	520	1	32	2314
	FP	163	27	2	3	1	45	1	0	215
	FN	4899	5747	6418	6406	6421	5904	6423	6392	4110
	TN	7913	8049	8074	8073	8075	8031	8075	8076	7861
	precision	90%	96%	75%	86%	75%	92%	50%	100%	91%
	recall	24%	11%	0.1%	0.3%	0.05%	8%	0.1%	0.5%	36%
		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
90%	TP	1866	684	6	43	3	973	1	53	2802
	FP	486	27	2	9	1	115	1	0	573
	FN	4558	5740	6418	6381	6421	5451	6423	6371	3622
	TN	7590	8049	8074	8067	8075	7961	8075	8076	7503
	precision	79%	96%	75%	83%	75%	89%	50%	100%	83%
	recall	29%	11%	0.1%	1%	0.05%	15%	0.1%	1%	44%
		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
80%	TP	2220	1670	11	888	3	1826	1	58	4566
	FP	885	133	2	150	1	397	1	5	1319
	FN	4204	4754	6413	5536	6421	4598	6423	6366	1858
	TN	7191	7943	8074	7926	8075	7679	8075	8071	6757
	precision	71%	93%	85%	86%	75%	82%	50%	92%	78%
	recall	35%	26%	0.2%	14%	0.05%	28%	0.1%	1%	71%

Fig. 5. Performance of the PRIs in data set A. "Time" split experiment

In the "random" split experiment it can be observed that lowering the desired precision level leads to a decrease in precision and an increase in recall. While this can also be observed in the "time" split experiment the decrease of precision is more pronounced while the increase in recall is less.

We have also applied to data set A the risk identification algorithm without configuring the PRIs using a "random" 75/25% split. The results are depicted in Figure 6. For PRIs 1, 2, 3, 5, 6 and 7 the cut-off thresholds were defined as $\overline{x} + 2 * s$, i.e., we assume normal distributions and use a 95% confidence interval. We did not use PRIs 4 and 8 in this experiment as they can only be learned using information about the outcomes of past cases. Precision levels for all indicators were significantly lower than the corresponding values from our previous experiment where we configured the PRIs (depicted in Figure 4). *This confirms that proper configuration of indicators is an essential step in the risk identification method.*

Then we conducted the experiments with data sets B1–B5. Figure 7 depicts the results of the experiments for five departments in data sets B1-B5. We have used a random 75/25% split and 90% as the value for the desired precision level.

	PRI 1	PRI 2	PRI 3	PRI 5	PRI 6	PRI 7	Total
TP	936	153	98	1043	490	83	1856
FP	508	72	283	1966	334	359	2510
FN	2375	3158	3213	2268	2821	3228	1455
TN	5148	5584	5373	3690	5322	5297	3146
precision	65%	68%	26%	35%	59%	19%	43%
recall	28%	5%	3%	32%	15%	3%	56%

Fig. 6. Performance of the PRIs without configurations in data set A. "Random" split experiment

		PRI 1	PRI 2	PRI 3	PRI 4	PRI 5	PRI 6	PRI 7	PRI 8	Total
B1	TP	126	67	3	0	0	78	17	42	199
	FP	29	15	3	0	0	15	5	5	48
	FN	290	349	413	416	416	338	399	374	217
	TN	710	724	736	739	739	724	734	734	691
	precision	81%	82%	50%	NA	NA	84%	77%	89%	81%
	recall	30%	16%	1%	NA	NA	19%	4%	10%	48%
B2	TP	10	34	13	0	2	34	5	12	51
	FP	1	2	1	0	0	2	5	1	10
	FN	48	24	45	58	56	24	53	46	7
	TN	59	58	59	60	60	58	55	59	50
	precision	91%	94%	93%	NA	100%	94%	50%	92%	84%
	recall	17%	59%	22%	NA	3%	59%	9%	21%	88%
B3	TP	191	168	10	0	2	269	14	30	329
	FP	43	26	4	0	0	44	8	5	80
	FN	281	304	462	472	470	203	458	442	143
	TN	873	890	912	916	916	872	908	911	836
	precision	82%	87%	71%	NA	100%	86%	64%	86%	80%
	recall	40%	36%	2%	NA	0%	57%	3%	6%	70%
B4	TP	120	35	2	0	2	44	4	6	171
	FP	10	8	2	0	1	12	5	1	31
	FN	239	324	357	359	357	315	355	353	188
	TN	410	412	418	420	419	408	415	419	389
	precision	92%	81%	50%	NA	67%	79%	44%	86%	85%
	recall	33%	10%	1%	NA	1%	12%	1%	2%	48%
B5	TP	117	44	1	0	0	144	10	116	250
	FP	14	11	0	0	0	14	4	7	39
	FN	205	278	321	322	322	178	312	206	72
	TN	284	287	298	298	298	284	294	291	259
	precision	89%	80%	100%	NA	NA	91%	71%	94%	87%
	recall	36%	14%	0%	NA	NA	45%	3%	36%	78%

Fig. 7. Performance of the PRIs in data sets B1-B5. "Random" split experiment

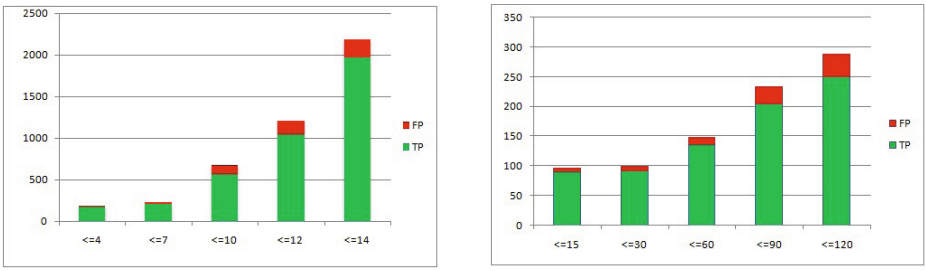
PRIs 1, 2, 6 and 8 demonstrated a good performance for all departments, and a few delays were predicted with PRIs 3, 5 and 7. PRI 4 ("Presence of a risky activity") did not predict any delays for these data sets because no single activity was a strong predictor of delays in these logs.

4.4 Moment of Delay Prediction

We also evaluated the ability to predict delays early during a case's execution which is obviously a highly desirable capability. In order to do so we checked how many true positive and false positive predictions (coming from any of the indicators) were generated before a given point in time during a case's execution, to find the earliest point when we can identify risks. Since the event logs available to us do not have "assign" events recorded, we consider the time of the "start" event for an activity to be the discovery time for PRIs 3, 4, 5, 7 and 8, e.g., when an activity has been started by a "risky" resource (PRI 8), or by a resource with a high workload (PRI 7). The earliest time when we can observe PRI 1 ("Atypical activity duration") is the time of the "start" event of an activity plus the value of PRI 1's threshold for this activity. For example, if an activity is not completed

within three days (the threshold value) after it has been started there is a higher likelihood of the case delay, i.e., at this point we can already predict delay. The earliest time when PRI 2 ("Atypical waiting time") can be observed is either the time of the "complete" event of an activity plus the maximum of its successors' PRI 2 thresholds or the time of the "start" event of the next activity if it has been started earlier and its wait duration is higher than its PRI 2 threshold. For example, if an activity is completed and none of its successors have been started within two days (maximum of their PRI 2 thresholds), we can say at this point that a case delay is likely due to PRI 2. A similar approach for calculating the discovery time is used for PRI 6.

Figure 8(a) depicts the discovery times for data set A. Recall that the discovery time is the time at which a true positive or false positive predictions are generated. Figure 8(b) presents the discovery times for data set B5. The horizontal axes in both diagrams represent the number of days since the beginning of a case when the risk of the case delay was discovered. Cases from data set A should be completed within 14 days while the typical case duration for data set B5 is 120 days. The vertical axes depict the cumulative number of delay predictions at a certain point in time. For example, Figure 8(a) shows that more than 1000 correct delay predictions have been generated within the first twelve days. For data set A early predictions (below seven days) are coming mainly from PRI 4 ("Presence of a risky activity") and PRI 8 ("Use of a risky resource"). Early predictions for data set B5 (below 30 days) were generated mainly by PRI 8 ("Use of a risky resource") and PRI 7 ("High resource workload").



(a) (b)

Fig. 8. PRI discovery times for data sets A (a) and B5 (b), "Random" split experiment with 90% as the desired precision level

4.5 Discussion

Some of the limitations of the experiments described above are related to the data available to us. One of the two data sets provided by Suncorp displayed high process variability. Multiple process variants may have influenced the performance of PRIs that rely on the order of activities (PRIs 2 and 6), however the performance of other indicators should not be affected. The other concern

is related to estimating the performance of PRI 7 "High resource workload". This is due to two reasons. The first one is that the event logs available to us contained only completed cases, i.e., traces corresponding to running process instances were filtered out. We also assumed that all resources are involved in one process. Hence, the workload of resources may have been underestimated. In order to more accurately estimate the performance of this PRI complete information about all processes in a company is required. This limitation should not affect the performance of other indicators.

A limitation of the approach is our assumption that a process is in a steady state, i.e. it is not changing over time. To deal with this limitation in this paper we used data from a relatively short period (six months). However, if a process's behaviour is constantly changing, the amount of available up-to-date data may be insufficient for proper configuration of PRIs.

We considered *instance* and *process* contexts, however we did not consider *social* and *external* contexts using the terminology of [17], that may also influence case durations. This is a direction for possible future research. Another direction for future work is to investigate the relation between PRIs and the extent of the expected delay.

5 Conclusions

In this paper, we presented a method for configuration of risk indicators for process delays. The method learns parameters of indicators by analysing event logs and exploiting information about the outcomes of cases completed in the past. Such configuration of indicators takes the specifics of a particular process into account thus improving the accuracy of the predictions. We conducted a number of experiments with different data sets from an Australian insurance company that confirmed that this approach decreases the level of false positive alerts and thus significantly improves the precision of case delay predictions.

The experiments demonstrated the ability to predict case delays with eight selected PRIs. Some of the indicators showed a consistently good performance in all data sets (e.g., PRIs 1, 2 and 6), others are good predictors of delays for some processes but did not predict delays for others (e.g., PRIs 4, 7 and 8). PRIs 3 and 5 produced few predictions for this particular data set due to the fact that it was typically possible to discover these indicators after the deadline was missed. As is often the case in the data retrieval field, there is a trade-off between precision and recall. It is hard to predict more than 50% of case delays with a high degree of precision using our indicators, while many delays can be predicted with a degree of precision of 80%. We expect that our approach can be applied for configuration of indicators for other types of process risks such as cost overruns or low-quality outputs, but this should be explored in future work.

Acknowledgements. This research is funded by the ARC Discovery Project "Risk-aware Business Process Management" (DP110100091). We would like to thank Suncorp for providing the data sets for analysis.

References

1. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops 2011, Part I. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012)
2. Balasubramanian, S., Gupta, M.: Structural metrics for goal based business process design and evaluation. Business Process Management Journal 11(6), 680–694 (2005)
3. Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., Shan, M.C.: Business process intelligence. Computers in Industry 53(3), 321–343 (2004)
4. Grigori, D., Casati, F., Dayal, U., Shan, M.C.: Improving business process quality through exception understanding, prediction, and prevention. In: 27th International Conference on Very Large Databases (VLDB 2001). Morgan Kaufmann Publishers Inc. (2001)
5. Hollands, J.G., Wickens, C.D.: Engineering psychology and human performance. Prentice Hall, New Jersey (1999)
6. International Organization for Standardization. Risk management: vocabulary = Management du risque: vocabulaire (ISO guide 73), Geneva (2009)
7. Jallow, A.K., Majeed, B., Vergidis, K., Tiwari, A., Roy, R.: Operational risk analysis in business processes. BT Technology Journal 25(1), 168–177 (2007)
8. Jans, M., Lybaert, N., Vanhoof, K., van der Werf, J.M.: A business process mining application for internal transaction fraud mitigation. Expert Systems with Applications 38(10), 13351–13359 (2011)
9. Jansen-Vullers, M.H., Reijers, H.A.: Business process redesign in healthcare: Towards a structured approach. Quality Control and Applied Statistics 52(1), 99 (2007)
10. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, vol. 14, pp. 1137–1145. Lawrence Erlbaum Associates Ltd. (1995)
11. Mansar, S.L., Reijers, H.A.: Best practices in business process redesign: use and impact. Business Process Management Journal 13(2), 193–213 (2007)
12. Moeller, R.: COSO enterprise risk management: understanding the new integrated ERM framework. In: Components of COSO ERM. ch. 3, pp. 47–93. John Wiley & Sons, Inc., Hoboken (2007)
13. Nakatumba, J., van der Aalst, W.M.P.: Analyzing Resource Behavior Using Process Mining. In: Rinderle-Ma, S., Sadiq, S., Leymann, F. (eds.) BPM 2009. LNBIP, vol. 43, pp. 69–80. Springer, Heidelberg (2010)
14. Pika, A., van der Aalst, W.M.P., Fidge, C.J., ter Hofstede, A.H.M., Wynn, M.T.: Predicting deadline transgressions using event logs. In: La Rosa, M., Soffer, P. (eds.) BPM Workshops 2012. LNBIP, vol. 132, pp. 211–216. Springer, Heidelberg (2013)
15. Standards Australia and Standards New Zealand. Risk management: principles and guidelines (AS/NZS ISO 31000:2009), 3rd edn., Sydney, NSW, Wellington, NZ (2009)
16. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Berlin (2011)
17. van der Aalst, W.M.P., Dustdar, S.: Process mining put into context. IEEE Internet Computing 16(1), 82–86 (2012)
18. van der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Information Systems 36(2), 450–475 (2011)

19. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering 16(9), 1128–1142 (2004)
20. van Dongen, B., Crooy, R., van der Aalst, W.M.P.: Cycle time prediction: When will this case finally be finished? In: On the Move to Meaningful Internet Systems: OTM 2008, pp. 319–336 (2008)
21. Wickboldt, J.A., Bianchin, L.A., Lunardi, R.C., Granville, L.Z., Gaspary, L.P., Bartolini, C.: A framework for risk assessment based on analysis of historical information of workflow execution in IT systems. Computer Networks 55(13), 2954–2975 (2011)