Data science for tax administration

Pijnenburg, M.G.F.

5

Are Similar Cases Treated Similarly? A comparison between process workers

For tax administrations similar treatment of similar cases is demanded by law. Nevertheless, in practice it is not always easy to achieve this. Especially in processes involving human professional judgment (e.g., in Knowledge Intensive processes), such as audit selection, debt management and the processing of disputes, it is not even easy to verify if similar cases receive similar treatment. In these processes there is a risk of dissimilar treatment as human process workers may develop their individual experiences and convictions or change their behavior due to changes in workload or season. Awareness of dissimilar treatment of similar cases may prevent disputes, inefficiencies, or non-compliance with regulations that require similar treatment of similar cases. Therefore, in this chapter, we address the research question:

Research Question *Can we develop a data science technique that assists in testing similar treatment of similar (tax) cases?*

In this article two procedures are presented for testing in an objective (statistical) way if different groups of process workers treat similar cases in a similar way. The testing is based on splitting the event log of a process in parts corresponding to the different (groups of) process workers and analyzing the sequences of events in each part. The two procedures are demonstrated on an example using synthetic data and on a real life event log from tax data. The chapter is based on the following article:

- M. Pijnenburg and W. Kowalczyk. Are similar cases treated similarly? a comparison between process workers. In *International Conference on Business Information Systems*, pages 1–15. Springer, 2019

5.1 Introduction

Knowledge Intensive (KI) processes present an expanding topic in the field of Business Process Management (BPM), see for instance the paper of Marin et al. [75]. KI processes differ from classical processes by the presence of human process workers whose domain knowledge has an important influence on the next steps in the process. KI processes occur typically in organizations that complete complex tasks.

The human component in KI processes makes the flow of activities less predictive and raises the question whether similar cases are treated similarly. Differences in treatment of similar cases may result from differences in the expertise, workload, and opinions of the human process workers, especially when the same process is executed at two or more different sites. For example, when similar cases are presented to two process workers, they may process them in different ways.

Awareness of dissimilar treatment of similar cases by different groups of process workers (e.g., at different locations) is important for businesses, not only because a uniform treatment is usually preferred with a view on efficiency, but also since the lack of uniformity may create disputes between customers and the company. For governmental organizations similar treatment of similar cases is even more important, as similar treatment is often demanded by law or policy. As such, the topic is of interest for auditors as well. Auditors, besides the classical task of checking financial statements, are also expected to check compliance with regulations and the law.

Verifying similarity of treatments for two or more (groups of) process workers is complicated by two facts. First, the characteristics ('attributes') of cases presented to process workers may differ from process worker to process worker. For instance, when two process workers are employed in different regions, one should take into account the regional differences of the cases presented to these process workers. Second, human professionals often have the possibility (and are expected) to gather additional information about the cases they treat. Although this (unstructured) information may be stored in a case management system, it is usually not registered in the event log of the process. This 'data incompleteness' introduces an additional stochastic component when modelling the decisions of a process worker. Hence statistical techniques are required to make measure similarity of treatments, see Section 5.3 for more details.

A simple example may illustrate the issue of similar treatment. Consider a Knowledge Intensive service process of a manufacturer of laptops and printers, as shown in Figure 5.1. The service process supports customers whose device breaks down within the warranty period. A defect device that comes in, initially starts two activities that are processed in parallel: registration (V_1) and sending a confirmation (V_2). Part of the registration process is the recording of the *device type* (possible values: 'laptop' and 'printer') and the original purchase price (*price*). After registration and confirm-

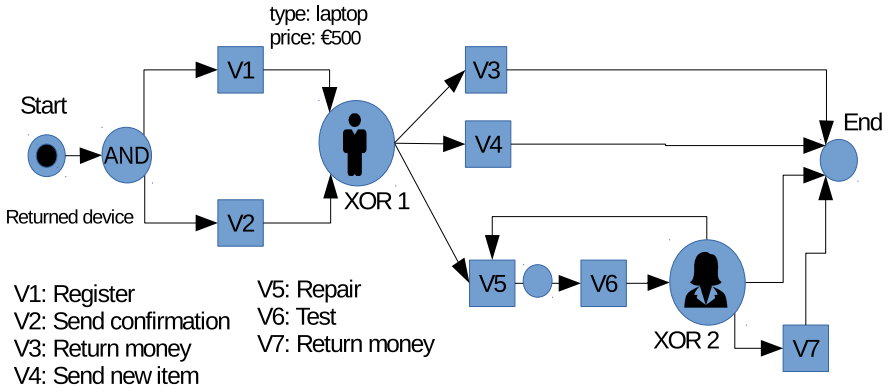

Figure 5.1: Simple process model of a service process of a laptop and printer manufacturer.

ation, a human process worker considers the defect device and has three options: return the original purchase price (V_3), send a new device (V_4), or send the defect device for repair (V_5). The choice of option depends on an assessment of the actual defect, the recorded values, and possibly additional information gathered by contacting the customer. When the repair option is chosen (V_5), the device will be send to an external repair shop and the device will be tested (V_6) afterwards. Depending on these test results, a second process worker will decide on the next activity; a good test result will lead to ship the repaired device back to the customer. Otherwise the process worker may choose to send the device again for repair or to return the money to the customer (V_7).

Traces in the event log of this process may look like this:

- `244 Printer · Start V1 V2 XOR1 V5 V6 XOR2 V5 V6 XOR2 V7 End`
- `69 Printer · Start V2 V1 XOR1 V3 End`
- `224 Laptop · Start V1 V2 XOR1 V4 End`
- `1082 Laptop · Start V1 V2 XOR1 V5 V6 XOR2 V7 End`
- `67 Printer · Start V1 V2 XOR1 V4 End`

where the first number is the purchase price of the device.

The service process is deployed at various regions. The manufacturer is interested in knowing whether similar defects receive the same treatment at each region to deal with a rise of complaints claiming that customers in region 'A' usually receive a

new item quickly, while customers in region 'B' have to wait for a repair. The service process of this manufacturer will serve as an example throughout this paper.

The research into similar treatment is motivated by two real-life examples where knowledge of similar treatment of similar cases is found to be important. One is the debt-collection process of a tax administration where it is expected that debtors in various regions are treated in a uniform way, see Section 5.4.2 for more details. The other is the treatment of patients in two wards of the same hospital according to the same protocol.

The paper is organized as follows. Section 5.2 reviews related work. Section 5.3 describes two testing procedures. Section 5.4 demonstrates these procedures on the 'warranty'-example and the tax data set. We end with conclusions and future research in Section 5.5. The code used in Section 5.4.1 can be found on github: `github.com/PijnenburgMark/Similar_Cases_Treated_Similarly/`.

5.2 Related Work

5.2.1 Process Drift

Detecting different variants of business processes is a topic that is receiving increasing attention [19], [72], [17], [82]. Most works focus on detecting changes in a business process over time, but some papers also mention differences by location (e.g., departments), like the paper of Pauwels and Calders [83], or mention the more general applicability of their method, like the paper of Bolt et al. [17].

One of the earliest papers in the field of process drift is the paper of Bose et al. [19]. The method described in this paper extracts features from an event log and compares the values of these features at two different points in time. The features demonstrated in the paper only allow for finding ' structural changes', i.e., changes in the process model.

The paper of Ostovar et al. [82] differs from the paper of Bose et al. [19] by using a technique for automated discovery of process trees and considering changes in these trees instead of using features extracted from event logs. Moreover, root cause analysis is supported by providing natural language statements to explain the change behind the drift.

In the paper of Pauwels and Calders [83] the concept of process drift is applied to the data set of the BPI challenge 2018 [108]. Besides showing concrete results, the authors add a new model-based approach relying on Dynamical Bayesian Networks. Their approach is strengthened by providing some nice aides for detecting process drift visually.

The paper of Maaradji et al. [72] adds a formal statistical test to the comparison

between different time points and thus adds objectivity. Moreover the detection of process drift is determined by looking at the frequencies of sequences of activities, and is thus capable of detecting structural changes as well as changes in frequencies of process paths.

In our approach, similarly to Maaradji et al. ([72]), we also look at the activities themselves instead of derived features and we also apply a formal statistical test. However, there are several differences. From the contextual viewpoint there are two main differences. First, we consider differences in process execution between groups of users instead of differences in time and second, we focus on the human decision making in the process and thus leaving aside the activities that are triggered automatically by the workflow management software. As a consequence of the latter, we have to consider fewer differences, making the comparison simpler and the statistical testing more powerful. From the technical point of view, there are two differences as well. First we take into account the attributes of the cases that enter the process and second, we consider the frequently occurring situation where some or all decisions in the process are independent of each other. If this assumption holds, the χ^2-test becomes much more powerful and easy to use, because of the mathematical property that the sum of independent χ^2-distributions is again a χ^2-distribution.

5.2.2 Sequence Analysis

Event logs form the basis of process mining and are in essence a number of sequences of activities. For this reason it is not surprising that techniques from sequence analysis (also known as 'sequential pattern analysis' or 'sequential pattern mining') are applicable in process mining. One of the techniques from sequence analysis that is applied frequently in this paper, is Pearson's χ^2-test to test the probabilities of going from one activity to the next against a theoretical model, see the book of Bakeman and Gottman [9]. We apply this test repeatedly under the null hypothesis of no differences in treatment between (groups of) process workers. As described by Bakeman and Gottman [9], Pearson's χ^2-test can be also applied to sequences of events, although the number of possible sequences (and thus the number of observations needed to fulfill the assumptions of the test) grows exponentially in the sequence length. We overcome this problem by making an assumption of independence of decisions (Section 5.3.1) or by focusing on the most frequent traces (Section 5.3.2).

5.3 Testing for similar treatment of similar cases

When testing similar treatment of similar cases for two groups of process workers, a first step is finding the points in the process where human process workers decide on

the next activity. We will call these points 'decision points'. In a process model these are 'XOR-junctions' (eXclusive OR-junctions), i.e., places where the flow of activities can take two or more directions.

We define 'similar treatment of similar cases' as the situation where at each decision point holds that two similar cases (measured by having the same attributes) have the same probability distribution over all possible follow-up activities.

In case only one decision point is present, the situation is relatively simple and the approach is described in Section 5.3.1. When multiple decision points are present, two situations can occur: a) the decisions at the decision points can be considered independent of each other (frequently called a 'Markov assumption' in sequence analysis). In this situation the comparing of two groups of process workers can be done for each decision point individually and the results can later be combined. This case is addresses in Section 5.3.1. b) the decisions cannot be considered independent, in other words the decision at one decision point is influenced by the previous decisions (i.e., it matters for the decision what path a case took through the process model to reach the decision point). The problem that arises in this situation is that the number of possible combinations of decisions increases exponentially in the number of decision points and hence a lot of traces in the event log are needed if a naive approach is taken. In Section 5.3.2 we present a heuristic that requires less data, by assuming that a subset of possible traces accounts for most observed traces in the event log. An assumption which is often reasonable in practice.

5.3.1 Approach 1: Independence of Decisions

5.3.1.1 A test for a single decision point

If we compare two groups of users, say from location L_1 and L_2, and we focus for the moment on *one* decision point, a table like Table 5.1 forms the basis for analysis. In the first column of Table 5.1 'Cluster of case', we see that all cases are grouped based on their attributes into K clusters. This clustering is done to make groups of similar cases. Then in the next column we see the decision that was made at the decision point, i.e., the next activity for the case. Then follow two columns indicating the count data of cases for both groups of users (L_1 and L_2). See Table 5.7 for two examples of filled out tables.

If we consider a subtable of Table 5.1 that belongs to one cluster, we can apply the well known two sample χ^2-test (see for instance the book by Kanji[56]) to test whether the differences between the two user groups can be attributed to chance, or that there is a reason to reject the hypothesis of similar treatment. If we want to apply the χ^2-test to the whole table (i.e., for all clusters simultaneously), we can take the approach as described for instance by Hald [47] (page 746). Note that Table 5.1

		group of process workers	
Cluster of case	output activity	L_1	L_2
x = 1	V^1	$n_1^{1,1}$	$n_2^{1,1}$
	$\vdots$	$\vdots$	$\vdots$
	V^q	$n_1^{1,q}$	$n_2^{1,q}$
x = 2	V^1	$n_1^{2,1}$	$n_2^{2,1}$
	$\vdots$	$\vdots$	$\vdots$
	V^q	$n_1^{2,q}$	$n_2^{2,q}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x = K	V^1	$n_1^{K,1}$	$n_2^{K,1}$
	$\vdots$	$\vdots$	$\vdots$
	V^q	$n_1^{K,q}$	$n_2^{K,q}$

Table 5.1: Three-way contingency table needed to apply the χ^2-test for one decision point for two groups of process workers.

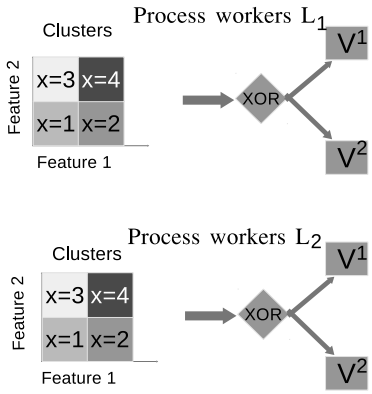

Figure 5.2: Figure illustrating the elements of the table at the left

contains only two locations. This is only for demonstrative purposes as the χ^2-test works equally well for any number of user groups.

Mathematically the test for one decision point comes down to the following. To test the null hypothesis of similar treatment of similar cases, we apply the (two sample) χ^2-test. So we calculate,

$$\chi^2_{XORj} = \sum_{i \in \text{all cells}} \frac{(O_i - E_i)^2}{E_i} = \sum_{x,V,L} \frac{(n_L^{x,V} - E_L^{x,V})^2}{E_L^{x,V}}, \tag{5.1}$$

where $E_L^{x,V}$ denotes the expected number of counts. These expected numbers are estimated assuming no differences between the groups of process workers ($\mathcal{H}_0$), so,

$$E_L^{x,V} = \frac{n_\bullet^{x,V} n_L^{x\bullet}}{n_\bullet^{x\bullet}}, \tag{5.2}$$

where,

$$n_\bullet^{x,V} = \sum_L n_L^{x,V}, \qquad n_L^{x,\bullet} = \sum_V n_L^{x,V}, \qquad n_\bullet^{x\bullet} = \sum_{V,L} n_L^{x,V}. \tag{5.3}$$

For one cluster (e.g., $x = 1$), the degrees of freedom are equal to $(q-1)(t-1)$ [47] (where t is the number of groups of process workers, and q the number of next activities). Since there are K clusters, we have,

$$df_{XORj} = K(q-1)(t-1). \tag{5.4}$$

The statistic (5.1) is χ^2-distributed with degrees of freedom given in (5.4) if the assumptions of the Pearson's χ^2-test are fulfilled.

Experiments have demonstrated that the assumptions of Pearson's χ^2-test are sufficiently fulfilled if *no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater*, see the book by Yates et al. [115]. One way of meeting this condition is to find a balance in the number of clusters K, such that K is small enough for each cluster to contain enough cases, while simultaneously keeping K large enough to make sure that cases that are considered as dissimilar by experts are in different clusters. See Section 5.4 for the values used in the experiments. If after choosing an appropriate number of clusters the assumptions of Pearson's test are still violated, we propose to remove rows from Table 5.1 with infrequent row sums (i.e., activities that are seldom chosen by any groups of process workers) and apply equation (5.1) on the reduced table. Since the frequency of these activities is low for all process workers, the effect of removing will be small in practice. Of course, the degrees of freedom have to be adjusted. If l rows are omitted, then the number of degrees of freedom is given by:

$$df^{adj}_{XORj} = K(q-1)(t-1) - l(t-1). \tag{5.5}$$

By introducing two groups of process workers L_1 and L_2 and comparing the model of a decision point for these two groups, we must be cautious that possible differences are only caused by differences in treatment of the groups of process workers L_1 and L_2 and *not* by differences in the attributes of the cases that are presented to group L_1 and group L_2. Otherwise we would erroneously conclude there is dissimilar treatment of similar cases, while in reality there is dissimilar treatment of *dis*similar cases. In part, these differences in the cases of each group are prevented by the clustering the cases into K clusters based on the known attributes. However, typically not all attributes are recorded. If one suspects the existence of an attribute that has an influence on the decision taken by the process workers and this feature is also related to the group of users a case is assigned to, one must, within each cluster of cases, distribute cases randomly over the groups of process workers. This way one ensures the same probability distribution for the groups and one can safely apply the tests.

5.3.1.2 Multiple decision points that are independent

If there are multiple decision points in the process and we can assume that the decisions by the human process workers are independent from previous decisions in the process, we can easily extend our χ^2-test. When the statistics χ^2_{XORi} have been computed for the individual decision points, a test of similar treatment for the whole process (i.e., all m decision points) can be constructed by adding the individual statistics as well as the degrees of freedom:

$$\chi^2_{overall} = \chi^2_{XOR1} + \ldots + \chi^2_{XORm}, \tag{5.6}$$

with

$$df_{overall} = df^{adj}_{XOR1} + \ldots + df^{adj}_{XORm}. \tag{5.7}$$

The final statistic $\chi^2_{overall}$ is χ^2-distributed with $df_{overall}$ degrees of freedom since the sum of independent χ^2-statistics is again χ^2-distributed with the degrees of freedom equal to the sum of the original ones. The independence of the individual statistics is ensured by the Markov assumption.

The value of the statistic $\chi^2_{overall}$ leads to a p-value indicating the probability that the hypothesis $\mathcal{H}_0$ is due to the randomness in the sample. A value lower than 0.05 is generally accepted as a rejection of $\mathcal{H}_0$ and is thus evident of dissimilar treatment of similar cases.

The complete test procedure in case of independent decision points is summarized in Test Procedure 1.

Test Procedure 1: Procedure to test on similar treatment by two or more groups of process workers when decision points are independent

Input : Event logs
Output: A p-value
1 Find the decision points in the process;
2 **for** *each decision point* **do**
3 Find relevant attributes;
4 Cluster the cases of all event logs into K clusters based on the attributes (perform the clustering on the collective data of all user groups);
5 Construct a contingency table like Table 5.1;
6 Check the assumptions of Pearson's χ^2-test and possibly remove rows with low row sums (adjust degrees of freedom accordingly);
7 Use equation (5.1) to compute the test statistic for this decision point, and equation (5.5) for the degrees of freedom;
8 **end**
9 Apply equation (5.6) and (5.7) and read off the p-value from standard tables of the χ^2-distribution.

5.3.2 Approach 2: Frequent Paths

For some processes the Markov assumption will not hold. A straightforward generalization in this case is to cluster cases and consider for each cluster all possible traces and test if some traces occur significantly more frequently for one group of process workers. Although this approach is in line with the approach taken in the previous section, it has the drawback that a lot of data is needed in order to satisfy the underlying assumptions of Pearson's χ^2-test. Namely, the number of possible traces grows exponentially in the number of decision points. For this reason we will work out an approach that does not take into account *all* traces, but only the most frequent ones (i.e., 'typical paths') for each cluster of similar cases. Then the observed frequencies of these traces for each group of process workers are compared with the expected frequencies (based on the average frequencies over all process workers) using the χ^2-test.

Technically, the frequent paths approach comes down to first clustering cases based on known features, and then finding the most frequent traces for each cluster. If the event log contains many traces, special algorithms can be applied to find the most frequent traces, like the SPADE algorithm described by Zaki [116].

Subsequently, a table like Table 5.2 is created that contains the frequencies of these frequent paths for all groups of process workers. Then equation (5.8) can be

		group of process workers	
Cluster of case	frequent trace	L_1	L_2
x = 1	$V^1V^3V^5$	$n_1^{1,1}$	$n_2^{1,1}$
	⋮	⋮	⋮
	$V^2V^3V^4$	$n_1^{1,r}$	$n_2^{1,r}$
x = 2	⋮	⋮	⋮
⋮	⋮	⋮	⋮
x = K	V^1V^5	$n_1^{K,1}$	$n_2^{K,1}$
	⋮	⋮	⋮
	$V^1V^2V^3V^5$	$n_1^{K,s}$	$n_2^{K,s}$

Table 5.2: Three-way contingency table needed to apply the χ^2-test for all decision points for two groups of process workers when the assumption of independence of decision points does not hold.

applied to each frequent trace s in cluster x,

$$\chi^2 = \sum_{i \in \text{all cells}} \frac{(O_i - E_i)^2}{E_i} = \sum_{x,s,L} \frac{(n_L^{x,s} - E_L^{x,s})^2}{E_L^{x,s}}, \tag{5.8}$$

where $E_L^{x,s}$ is estimated by,

$$E_L^{x,s} = \frac{n_\bullet^{x,s} n_L^{x\bullet}}{n_\bullet^{x\bullet}}, \tag{5.9}$$

and,

$$n_\bullet^{x,s} = \sum_L n_L^{x,s}, \qquad n_L^{x,\bullet} = \sum_s n_L^{x,s}, \qquad n_\bullet^{x\bullet} = \sum_{s,L} n_L^{x,s}. \tag{5.10}$$

The degrees of freedom of the statistic in equation (5.8) equals:

$$df = (N - K)(t - 1), \tag{5.11}$$

where N is the number of rows in the contingency table (Table 5.2), K the number of clusters and t the number of groups of process workers. This can be seen by realizing that for each cluster i the degrees of freedom equals $(s_i - 1)(t - 1)$ [47], where s_i is the number of frequent paths for cases in cluster i. The Test Procedure for the Frequent Path approach is presented in Test Procedure 2.

Test Procedure 2: Procedure to test on similar treatment by two or more groups of process workers when decision points are dependent and most traces belong to a few frequently occurring traces

Input : Event log
Output: A p-value

1 Cluster all cases based on their features into K clusters;
2 Determine for each cluster the most frequent traces in the event log, for instance by applying a frequent sequence algorithm like SPADE. Discard all other traces;
3 Make a contingency table like Table 5.2;
4 Use equation (5.8) to compute the test statistic, and equation (5.11) for the degrees of freedom;
5 Find the p-value from standard tables of the χ^2-distribution.

		XOR-junctions	
		behavior I	behavior II
Features	distribution I	event log 1	event log 2
	distribution II	event log 3	event log 4

Table 5.3: Differences of the four generated event logs

5.4 Experiments

5.4.1 Synthetic data

In this section we will apply the test procedures to the example mentioned in the introduction. The code used for this can be found on github: `github.com/PijnenburgMark/Similar_Cases_Treated_Similarly/`.

We used the process model of Figure 5.1 to generate four event logs. Each event log consists of 500 traces and the value of two meaningful features for each case (price and type). The four event logs differ in the behavior of the two decision points and in the distribution of the two features, see Table 5.3.

Table 5.6 specifies the behavior I and behavior II of the decision points as well as the distributions of the input cases. For instance we see that under behavior I, a printer that is returned with an original purchase price of 200 has the probability of 0.2 for being send for repair, while this probability under behavior II is 0.6. Moreover we see that the behavior of decision point 2 (XOR 2) depends on the number of times

		2^{nd} Event Log			
		event log 1	event log 2	event log 3	event log 4
1^{st} Event Log	event log 1	1.000000	0.000000	0.975989	0.000000
	event log 2	0.000000	1.000000	0.000000	0.057141
	event log 3	0.975989	0.000000	1.000000	0.000000
	event log 4	0.000000	0.057141	0.000000	1.000000

Table 5.4: p-values resulting from applying Test Procedure 1 to all pairs of the four event logs. We used 10 clusters.

the device has been send for repair before (with two options: one time or more than one time) as well as the result of the test that is performed after the repair (activity V_6 in Figure 5.1). Note that if the results of the test are positive, the device is ready and thus there is a probability of one that the process trace will be ended. In our example we set the probability that the test of V_6 gives a positive result to 0.7.

Clearly, event logs 1 and 3 have the same process (i.e., no dissimilar treatment), but differ in the input (event log 1 mainly cheap printers, event log 3 more laptops). In contrast event logs 2 and 4 are generated with other process parameters. In these latter two event logs the process workers have a preference for the repair option instead of sending a new item or returning the money.

We will compare the four event logs pairwise. Large p-values (over 0.05) are expected when comparing event logs that have the same behavior (i.e., event log 1 with event log 3, and event log 2 with event log 4), while low p-value are expected if we compare event logs with different behavior such as event log 1 and event log 2.

The p-values resulting of applying Test Procedure 1 are shown in Table 5.4. The application of this test procedure is justified because the decision points are independent. The results are as expected, i.e., the test procedure is able to clearly distinguish dissimilar treatment from differences in the input. For feature selection we used a standard feature selection method from python's sci-kit learn package based on the ANOVA F-value, and for the clustering we applied a standard Gaussian Mixture clustering algorithm from the same package. The number of clusters was chosen $K = \#$ cases $/100$, where '# cases' are the number of cases that go through the decision point in the event log (for both groups).

We also applied Test Procedure 2, resulting in the p-values of Table 5.5. These values demonstrate that our algorithm properly captures relevant properties of this synthetic data set. In applying Test Procedure 2 we used again Gaussian Mixture clustering. The number of clusters was chosen $K = \#$ traces $/250$, where '# traces' are the number of traces of the combined event logs for both group of process workers.

		2^{nd} Event Log			
		event log 1	event log 2	event log 3	event log 4
1^{st} Event Log	event log 1	1.000000	0.000000	0.982920	0.000000
	event log 2	0.000000	1.000000	0.000000	0.662541
	event log 3	0.982920	0.000000	1.000000	0.000000
	event log 4	0.000000	0.662541	0.000000	1.000000

Table 5.5: p-values resulting from applying Test Procedure 2 to all pairs of the four event logs. We used 4 clusters.

5.4.2 Tax Debt Collection data

For demonstration purposes, we applied Test Procedure 1 to real data of the Netherlands Tax and Customs Administration (NTCA). In the year 2013 a debt collection process was in place, a part of which is sketched in Figure 5.3. As soon as a debt is

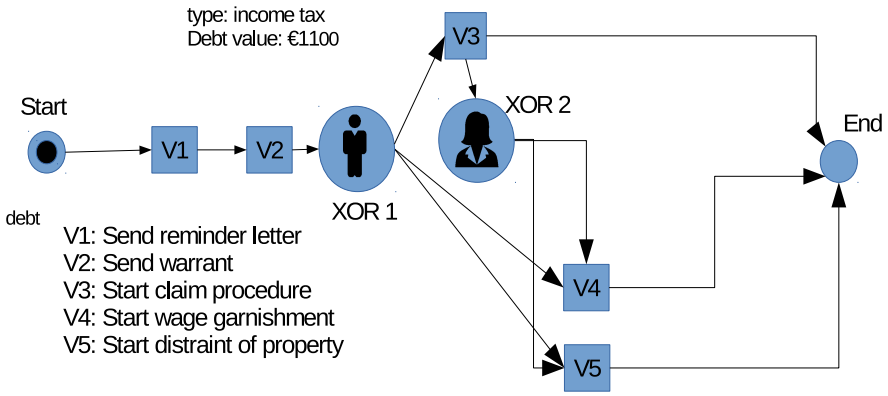

Figure 5.3: Part of the debt collection process of the NTCA in 2013.

over its due date, automatically a reminder letter is sent. When no payment has taken place, a legal notice (warrant) is sent that allows for legal actions afterwards. Usually, there are several legal actions possible and the choice is determined by a human process worker (XOR 1). The legal actions in this part of the process are: a special claim procedure that allows to take money of a savings account (V_3), wage garnishment (V_4), and distraint of property, e.g., a car (V_5). When the special claim procedure V_3 does not lead to payment of the debt, a second human process worker (XOR 2) can decide for actions V_4 or V_5.

We have compared event logs from two locations. We took 1000 debts from each location in the year 2013. We took into account two features of each debt: the debt

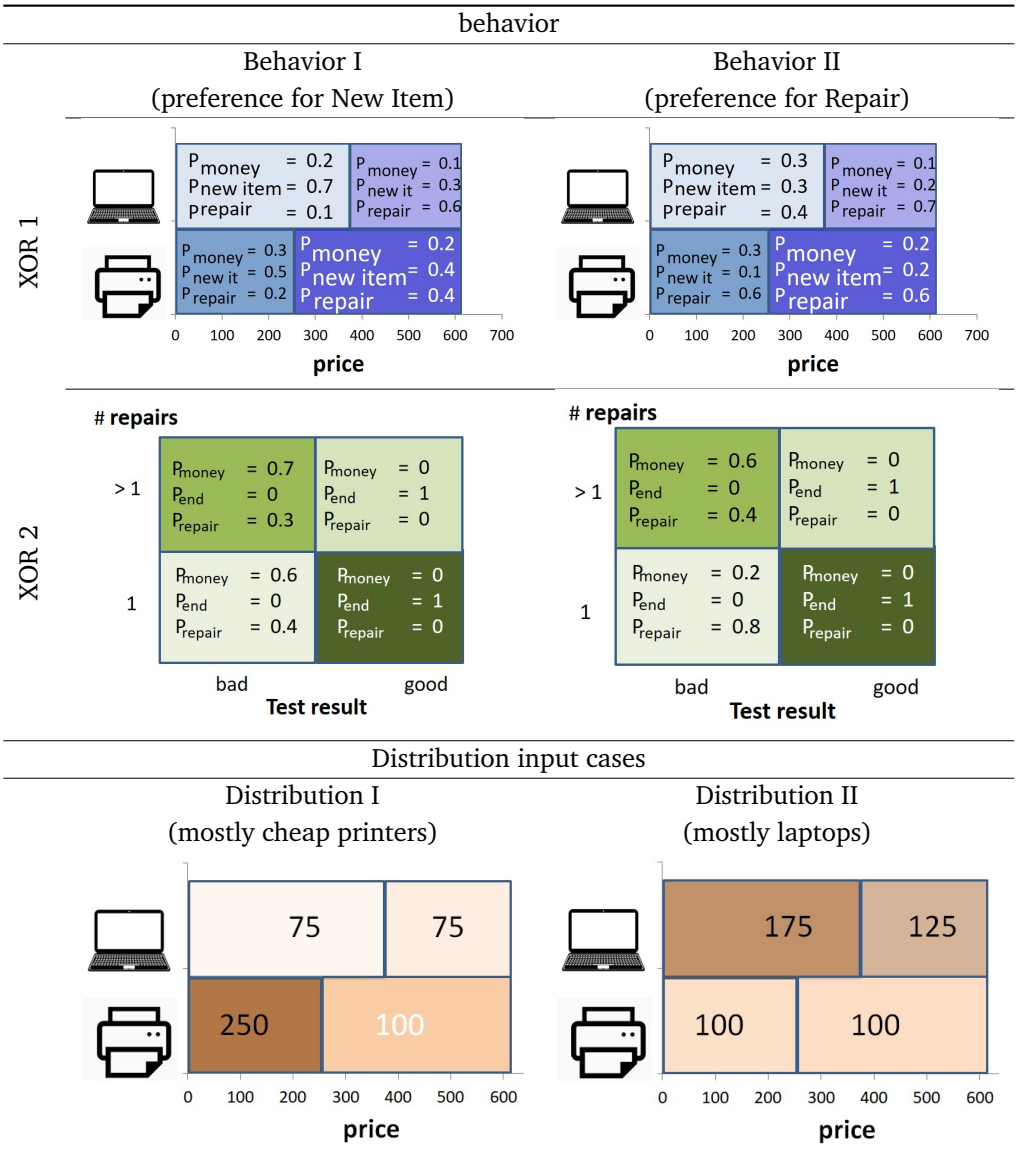

Table 5.6: Worked out example: two types of behavior of the XOR junctions and two distributions of the input cases.

value and the tax type and restricted ourselves to two tax types. The debt value has been used for constructing the clusters for both decision points, while the tax type played a role for the clustering of XOR 1 only. The contingency tables for both decision points are displayed in Table 5.7.

XOR 1		location	
Cluster	Activity	A	B
1	V_3	16	14
1	V_4	42	48
1	V_5	21	17
2	V_3	1	2
2	V_4	16	39
2	V_5	64	35
3	V_3	381	379
3	V_4	25	59
3	V_5	80	90
4	V_3	217	213
4	V_4	8	47
4	V_5	59	50
	Total	930	993

XOR 2	location		
Cluster	Activity	A	B
1	V_4	17	17
1	V_5	64	53
2	V_4	8	9
2	V_5	42	24
	Total	131	103

Table 5.7: Contingency tables belonging to the decision points XOR 1 and XOR 2 of the tax collection data. Note we were not able to analyze 77 of the 2000 cases.

The value of the χ^2-statistic for the first decision point is 59.247 (8 degrees of freedom), while 1.785 (2 degrees of freedom) for the second decision point. For the combination of the two decision point we find, under the assumption of independent decisions, a value of 61.032 (10 degrees of freedom) which corresponds to a p-value of $2.31 \cdot 10^{-9}$. The low p-value indicates that, under this model, we should reject the hypothesis of similar treatment of similar cases. Due to privacy reasons we did not use some important features that could lead to different results.

5.5 Conclusion and Future Research

The paper started by noting that the human factor in Knowledge Intensive processes raises the question of similar treatment by several groups of process workers. The two test procedures that are proposed in Section 5.3 are able to test similarly treatment as is demonstrated in Section 5.4 on an artificial example of a service process of a hardware manufacturer and a debt collection process of a tax administration.

Tax administrations may employ these tests to see in what processes the null hypothesis of similar treatment of similar cases does not hold. The tests allow to pinpoint at what step in the process this occurs and measures may be taken to restore the similar treatment.

The two test procedures presented in this paper are based on rather element-

ary statistical techniques. We choose deliberately to solve the problem with elementary techniques as simple methods display the nature of the problem most clearly. Moreover an elementary approach allows to communicate clearly with business experts. Besides an elementary approach allows for easy extensions to meet more particular needs. For instance most organizations may tolerate a certain small level of dissimilar treatment and the test may be extended to take this into account. However, applying some recently developed algorithms may have some advantages as well. In particular recurrent neural networks can be used successfully for modelling complex sequential data [68]. Unfortunately these networks are 'black-box models' and provide little insight into the nature of detected differences in the behaviour of groups of process workers.

Finally note that other interpretations of 'similar treatment' are possible and might be appropriate in some settings. In this paper similar treatment has been defined in terms of probabilities of going to the next activity in the process model. However similar treatment can also be defined for instance in terms of the amount of time that is spend on each case, or as the level of expertise that is involved in each case. Tests based on these metrics have not been explored yet, but may be the subject of further research.