

## Internal Fraud Risk Reduction: Results of a Data Mining Case Study

Peer-reviewed author version

JANS, Mieke; LYBAERT, Nadine & VANHOOF, Koen (2010) Internal Fraud Risk Reduction: Results of a Data Mining Case Study. In: International Journal of Accounting Information Systems, 11(1). p. 17-41.

DOI: 10.1016/j.accinf.2009.12.004

Handle: <http://hdl.handle.net/1942/8341>

# Internal Fraud Risk Reduction: Results of a Data Mining Case Study

Mieke Jans, Nadine Lybaert, Koen Vanhoof

## Abstract

Corporate fraud these days represents a huge cost to our economy. Academic literature already concentrated on how data mining techniques can be of value in the fight against fraud. All this research focuses on fraud detection, mostly in a context of external fraud. In this paper we discuss the use of a data mining approach to reduce the risk of internal fraud. Reducing fraud risk comprehends both detection and prevention, and therefore we apply descriptive data mining as opposed to the widely used prediction data mining techniques in the literature. The results of using a multivariate latent class clustering algorithm to a case company's procurement data suggest that applying this technique in a descriptive data mining approach is useful in assessing the current risk of internal fraud. The same results could not be obtained by applying a univariate analysis.

## 1 Introduction

Saying that fraud is an important (however not loved) part of business, is nothing new. Fraud is a million dollar business, as several research studies reveal. Among them are important surveys of PriceWaterhouse&Coopers (PwC, 2007) and of the Association of Certified Fraud Examiners (ACFE, 2006). The study conducted in the United States by the ACFE in 2004-2005 and the worldwide study, held by PwC in 2006-2007 yield the following insights. Forty-three percent of companies worldwide have fallen victim to economic crime in the years 2006 and 2007 (PwC, 2007). The average financial damage to companies subjected to the PwC survey was US\$ 2.42 million per company over two years. Participants of the ACFE study estimate a loss of 5% of a company's annual revenues to fraud. Applied to the 2006 United States Gross Domestic Product of US\$ 13,246.6 billion, this would translate to approximately US\$ 662 billion in fraud losses for the United States only. These numbers all address corporate fraud.

Numerous academic studies have used data mining to investigate fraud detection. (Brockett et al., 2002), (Cortes et al., 2002), (Estévez et al., 2006),

(Fanning and Cogger, 1998), (Kim and Kwon, 2006) and (Kirkos et al., 2007) Although prior research has investigated fraud under different domains and using different techniques, the studies have all focus on external fraud<sup>1</sup> and have used predictive data mining. Our study focuses on internal fraud, because this represents mainly these large costs in the PwC and ACFE surveys. While prior studies have only examined fraud detection, our study investigates the combination of fraud detection and prevention. We are convinced that this combination is of priceless value for organizations. If a company only focuses on fraud detection -which is a reactive working method-, it will require insights that may be overtaken by events. On the other hand, when complementing these insights by fraud prevention activities -a proactive working method-, insights will last longer and are of more value to the company. The business may even be ahead of perpetrators' learning. We will use the term fraud risk reduction for encompassing both fraud detection and prevention. The aim of this paper is to provide a framework for both researchers and practitioners to reduce internal fraud risk and to present some first empirical results in this topic. We hope other researchers will complement our work by investigating other techniques, other business processes, suggesting improvements in methodology or any other contribution that will lead to more insights in the topic of internal fraud risk reduction.

We extend prior research on the use of data mining to detect fraud by implementing a fraud prevention aspect. Because of this different aim, we apply however another category of techniques than applied up till now (for fraud detection). In current literature, mainly predictive data mining is used, more precisely classification techniques. The aim of the techniques is to classify whether an observation is fraudulent or not. Since we aim another contribution (risk reduction contrary to detection), we believe descriptive data mining is more suited. Descriptive data mining provides us with insights of the complete data set and not only one aspect of it: fraudulent or not. This characteristic is valuable for assessing the fraud risk in selected business processes.

In the following sections we explain the followed methodology of this study, the data set, the used latent class clustering algorithm, and the results of investigating a business process of the case company. We first apply a univariate analysis to explore the data and thereafter a multivariate analysis. Afterwards we compare the results of both analysis. We end with a conclusion.

---

<sup>1</sup>The dimension internal versus external fraud refers to the relation between the perpetrator and the victim company.

## 2 Methodology

The applied methodology can be summarized by Figure 1. As a first step, an organization should select a business process which it thinks is worthwhile investigating. This selection can be motivated by different aspects: a business process that has a great cash flow, one that is quite unstructured, one that is known for misuses, or one that the business has no feeling with and wants to learn more about. Also the implementation of advanced IT can be a selection characteristic, because according to Lynch and Gomaa (2003) this is a breeding ground for employee fraud. In a second step the stored data will be collected, manipulated and enriched. Manipulation contains organizing the data in the structure and format that is needed for processing. Enrichment is the creation of extra attributes by combining available attributes, for example computing some ratios and averages. These are mainly technical transactions which can be performed by any data analyst. During the third step, the technical data will be translated into behavioral data. This translation builds upon domain knowledge and is not just a technical transformation (like in step two). The core of the methodology, step four, is then to apply descriptive data mining for getting more insights in this behavioral data. In this study, we start with a univariate analysis and turn later to a multivariate clustering. The descriptives should provide the researchers a recognizable pattern of procedures of the selected business process. This pattern should normally be applicable to most of the observations. In addition, some other patterns of minor groups of observations in the data can arise. These are interesting to have a closer look at.

The last step is to audit by domain experts. By auditing observations of a subgroup with a deviating pattern, the domain expert can categorize the observations in four groups: fraudulent cases, cases of circumventing procedures, errors or mistakes, and extreme values. The fraudulent observations are part of fraud detection, while the observations that circumvent procedures or are created by mistake are part of fraud prevention. Fraud prevention is in this methodology primarily based on checking or taking away fraud opportunity. The importance of opportunity is stressed by Cressey's fraud triangle with opportunity being the only element of fraud risk that an employer can influence. The other two elements that raise fraud risk, rationalization and incentive, are personal characteristics. A business has hardly any effect on personal characteristics, so focussing on these elements is not interesting. Opportunity on the other hand is of a company's concern. The fourth category of audited observations, extreme values but very logic when looked into, are of no interest for internal fraud risk reduction. We think for example of a purchase of a main frame which, comparing amounts of purchases, will leap to the eye between purchases of compact discs. The much higher amount of the purchase is however quite natural.

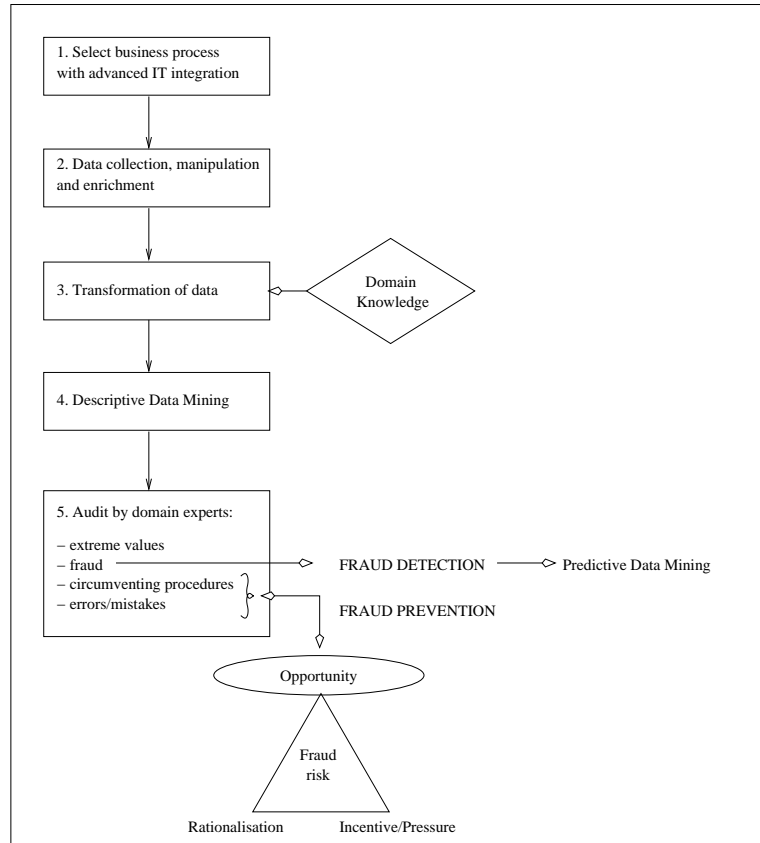


Figure 1: Methodology for internal fraud risk reduction.

### 3 Data Set

The data set was obtained from an international Financial services provider. The corporation is ranked in the top 20 of European financial institutions. The business process selected for internal fraud risk reduction is procurement. Data transactions from the case company's procurement cycle were collected. More specifically, the creation of purchasing orders (PO's) was adopted as process under investigation. This selection is inspired by the lack of fraud files in this business process within the case company<sup>2</sup>, while one assumes this business process is as vulnerable to fraud as every other business process.

The most important attributes to describe a PO and its life cycle are the following: the name of the creator, the supplier, the purchasing group, the type of purchasing document, the number of changes, the number of changes after the last release and the number of price related changes after the last release. 'Changes' are 'events' stored in the log file of the ERP system, so it should not be mistaken for changes in the sense of modifications alone. The creation of a PO, the modifications of a PO, the signs (first approval) and the releases (second approval) are all logged as 'changes'. Concerning the categorical attributes, there are 91 creators in the data set, 3.708 suppliers, 13 purchasing groups and 6 document types. (see Table 1) The histograms of Figure 2, 3, 4 and 5 provide us with some insights on the distribution of these attributes.

Table 1: Categorical attributes.

Categorical	Number in data set
Creator	91
Supplier	3.708
Purchasing Group	13
Document Type	6

As can be seen in Figure 2 the 91 creators do not introduce the same number of PO's in the ERP system. This is caused by the individual characteristics of each purchase. Some creators are responsible for a particular type of purchase which results in entering a lot of PO's, while other creators are responsible for other types of purchase which involves processing only a few PO's. There is one 'creator' responsible for 25% of the PO's in the data set. This is however not a person, but concerns a method of creating a PO, namely by inputting a batch into the ERP system. SAP sees this method as one creator. Also the turnover in terms of personnel has its reflection on the number of PO's per employee. For example an employee that is hired (or retired) halfway 2006, will process fewer PO's than employees registered for a full year.

---

<sup>2</sup>The case company has a department dedicated to handling fraud files, both internal as external frauds

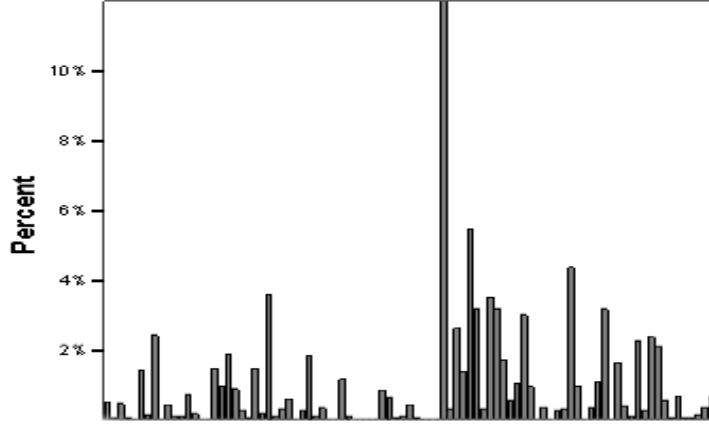


Figure 2: Percentage of PO's in data set per creator.

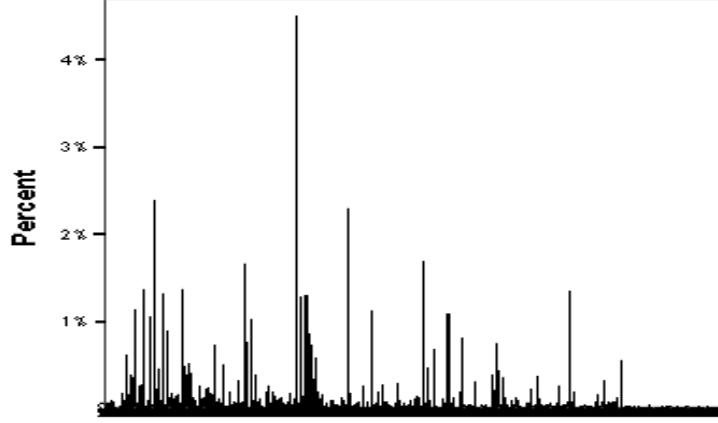


Figure 3: Percentage of PO's in data set per supplier.

Like creators, the frequency of suppliers in the data set is influenced by the specific characteristics of the product or service supplied. For example, there are more PO's concerning monthly leasing contracts for cars than there are for supplying desks. Hence the former supplier will be more frequently present in the data set than the latter. Concerning the 13 purchasing groups, some groups are more present than others in the data set, but this can all be explained by domain knowledge. The same goes for the six different purchasing document types. The run-down of two types (A and C) is clearly visible.

The numerical attributes are described in Table 2. For each attribute, three intervals were created, based on their mean and standard deviation. For the first attribute, the intervals were [2-4], [5-8] and [9-...], for the second attribute [0-0], [1-2] and [3-...] and for the last attribute [0-0], [1-1] and [2-...]. In Table 2 we see that there is a highly skewed distribution for the three attributes, which is to be expected for variables that count these types

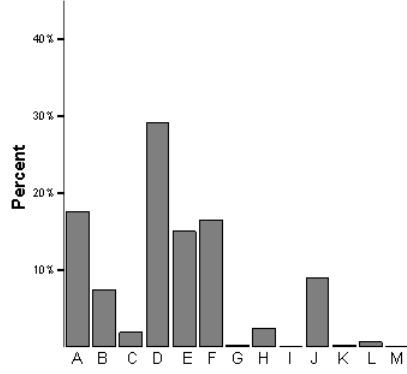


Figure 4: Percentage of PO's in data set per purchasing group.

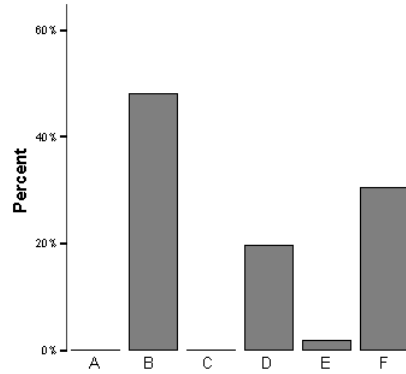


Figure 5: Percentage of PO's in data set per document type.

of changes. The changes are supposed to be small in numbers.

Table 2: Descriptives of numerical attributes.

Attribute	Minimum	Maximum	Mean	Standard deviation	1st interval frequency (%)	2nd interval frequency (%)	3rd interval frequency (%)
Number of changes	1	152	4.37	3.846	71.3	21.5	7.2
Number of changes after last release	0	91	.37	1.343	80.9	11.9	7.2
Price related number of changes after last release	0	46	.15	.882	91.1	6.7	2.2

In terms of collecting this data, the case company provided a txt-dump from their SAP system. All PO's that in 2006 resulted in an invoice are the subject of our investigation, yielding a data set of 36.595 observations. This raw data is then reorganized into appropriate tables to support meaningful analysis. After the creation of these new formats, additional attributes were created as enrichment. Based on domain knowledge and supported by descriptive statistics, a pre-clustering step is made. PO's are split in two groups: old PO's and new PO's. Old PO's are the ones created before July 2005. The fact that they are included in our data is because an invoice of the year 2006 can be linked to a PO created in 2005 or even before 2005. However, if a PO is from before July 2005 (keep in mind that even still in 2006 invoices were linked to this PO), this PO shows a different life cycle than if it were younger. While a 'new' PO will more probably have a life cycle of creation, approval and an attached invoice, an 'old' PO will probably be modified more often in between and have several invoices attached to it.



The subset of old PO's contains 2.781 observations while the subset of new PO's counts 33.814 observations. Both subsets of PO's were subjected to the main part of the methodology (step one through four). Because the latter group is the most prominent in assessing internal fraud risk (most recent and highest value in terms of fraud prevention) and given its magnitude, this paper gives only detailed test results of the new PO's. The other side of the picture is that this large data set poses more problems in the fifth step of our methodology, namely the auditing of interesting observations. We restrict this study to provide recommendations on this matter for the new PO's. For the subset of old PO's however, the audit step is effectively executed and these results will be reported after the discussion of the new PO's. The descriptive statistics provided above, all concern the subgroup of new PO's. In what follows, we provide univariate and multivariate analysis results and recommendations on the fifth step for this subset. Although all these steps are also executed on the subset of old PO's, we restrict this paper to the reflection of only the fifth step for this subset. In the following part, the term 'data set' refers to the subset of new PO's (33.814 observations) unless stated otherwise.

## 4 Latent Class Clustering Algorithm

For a descriptive data mining approach, we have chosen a latent class (LC) clustering algorithm. We prefer LC clustering to the more traditional K-means clustering for several reasons. The most important reason is that this algorithm allows for overlapping clusters. An observation is provided a probability to belong to each cluster, for example .80 for cluster 1, .20 for cluster 2 and .00 for cluster 3. This gives us the extra opportunity to look at outliers where the observation does not really belong to any cluster at all. This is for example the case with probabilities of: .35, .35 and .30. Another reason is that LC clustering algorithm has the ability to handle attributes of mixed scale types and has information criteria statistics to determine the number of clusters. For a more detailed comparison of LC clustering with K-means, see Magidson and Vermunt (2002).

In LC analysis, one starts with the idea that any dependency between the observed or manifest variables can be explained away by some other variable(s). These other variables can be unobserved or unobservable, called latent. We believe internal fraud risk can be represented by such a latent variable and hence fraud risk can be deduced from available information, i.e. manifest variables. We have technical information about a PO (who made it, when, ...) and we have operational information about this PO (how many times is it changed, ...). This operational information describes a behavior. It is this behavior, in combination with technical information, that we be-

lieve leads us to fraud. The main objective is to move from technical data over behavioral data (third step) to behavior (fourth step). Particular this challenge stimulates the use of data mining techniques, in that data mining enables us to recognize patterns we are unaware of. These patterns are used to gain insights in the behavior of people.

In LC clustering, a specific type of LC analysis, objects are assumed to belong to one of a set of  $K$  latent classes, with  $K$  being unknown. Observations in the same class are similar in the probability distributions underneath the manifest variables' scores. It is assumed that a population is a mixture of underlying probability distributions. Parting these different distributions provides us different clusters. The latent variable(s) are believed being capable of doing this.

The basic model for LC clustering has the form

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i|\theta_k)$$

$\mathbf{y}_i$  denotes an observation  $i$ 's scores on the dependent variables.  $\pi_k$  denotes the prior probability of belonging to latent class (or cluster)  $k$ . This model puts the conditional distribution of  $\mathbf{y}_i$  (given the model parameters of  $\theta$ ) as a mixture of class-specific densities,  $f_k(\mathbf{y}_i|\theta_k)$ .

Since there is the assumption of local independence between the manifest variables, this can be rewritten as

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^J f_k(y_{ij}|\theta_{jk})$$

where  $J$  denotes the total number of dependent variables and  $j$  is a particular dependent variable. Instead of specifying the joint distribution of  $\mathbf{y}_i$  given class membership, this is split up into separate univariate distribution functions for each  $y_j$ .

The latent variables are assumed to explain all associations between manifest variables (so that there is local independence between them). The goal is to determine the smallest number of latent classes  $K$  that is sufficient for safeguarding this assumption. The operating procedure is to test Model  $H_0$ , with  $K = 1$ , first. From this model on, latent classes are added.

For assessing the fit of LC models, there are several criteria available. The most widely used approach is the use of the statistic  $L^2$ . The lower this  $L^2$ , the less probability the model fits the data by chance. If  $L^2$  equals 0, this means the variables are perfectly independent from each other and all associations among the manifest variables are explained by the latent variables. For selecting a model, information criteria are also quite popular.

The ones most used are the Akaike, the Bayesian and the consistent Akaike information criteria, or AIC, BIC and CAIC. These criteria are based upon the log-likelihood (LL). Where  $LL$  always gets better (i.e. closer to 0) when  $K$  is raised, the information criteria take the number of parameters (and BIC also the number of degrees of freedom ( $df$ )) into account. The general definitions of BIC, AIC and CAIC are given below:

$$BIC = -2LL + \ln(N)M \quad (1)$$

$$AIC = -2LL + 2M \quad (2)$$

$$CAIC = -2LL + [\ln(N) + 1]M \quad (3)$$

with  $N$  being the sample size and  $M$  the number of parameters. The smaller the criterion, the better the model.

The mode of operation of starting with Model  $H_0$  and building further, is inspired by comparing the information criteria values with these of Model  $H_0$ .

For more and detailed information about LC analysis, we refer to Kaplan (2004) and Hagenaaars and McCutcheon (2002).

## 5 Univariate Clustering

### 5.1 Model Specifications

Before turning to the core of our model of applying a descriptive data mining approach on behavior describing attributes, we apply univariate clustering to explore data. The univariate analysis is applied on obvious attributes. The three most obvious attributes were selected: number of changes (Model A), number of changes after release (Model B) and number of price related changes after release (Model C). For each attribute and its belonging univariate clustering model, we executed the LC clustering algorithm with the number of clusters ( $K$ ) set equal to 1 till 5. This yielded the BIC information criteria values plotted in Figure 6. The AIC and CAIC values showed the same pattern. As you can see, the BIC values drop three times heavily until the 2-cluster model. Beyond the 2-cluster model, the decreases are more modest. Only at Model A, the 4-cluster model is an alternative candidate. The classification statistics of this model are however less satisfactory than those of the 2-cluster model. Based on these values, we decide to use three times the 2-cluster model for further exploration.

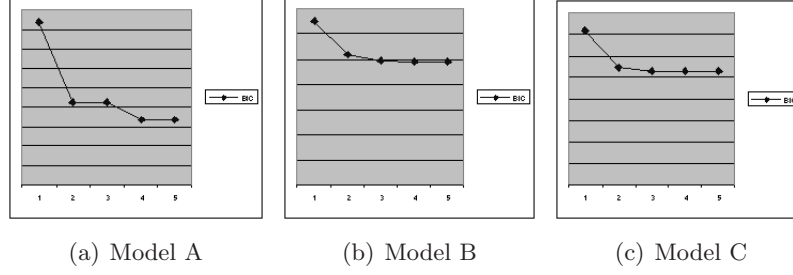


Figure 6: BIC values univariate clustering.

## 5.2 Results

In Table 3 we find the clustering results for Model A, B and C. For each of the three univariate models, one large and one small cluster is given as output of the 2-cluster models. The small clusters contain 1,428, 1,085 and 344 PO's (out of the 33,814) respectively for Model A, B and C. The 344 cases of the small cluster of Model C are fully incorporated in the 1,085 PO's of Model B. This is the only classification consistency between the three models. Except for the cluster size, the mean value of the attribute in each cluster is given. For each model, cases with a small value (of the count attribute) are classified in the large cluster while the small cluster is characterized by a higher mean value of that attribute.

Table 3: Results of univariate clustering.

	Cluster 1	Cluster 2
Model A		
Cluster size (%)	0.96	0.04
Number of changes	3.84	16.79
Model B		
Cluster size (%)	0.95	0.05
Number of changes after last release	0.17	3.82
Model C		
Cluster size (%)	0.99	0.01
Price related number of changes after last release	0.09	5.01

The clustering in each model is based on the value of one attribute. By looking at the following four attributes, we get insight in which kind of PO's are classified in these small clusters, aside from the high score on the clustering attribute: document type, purchasing group, creator and supplier.

In Figure 7 we see the distribution of the document types in the small clusters of Model A, B and C. Model A and C both highlight document type D, while at Model B both B and D hold a prominent place.

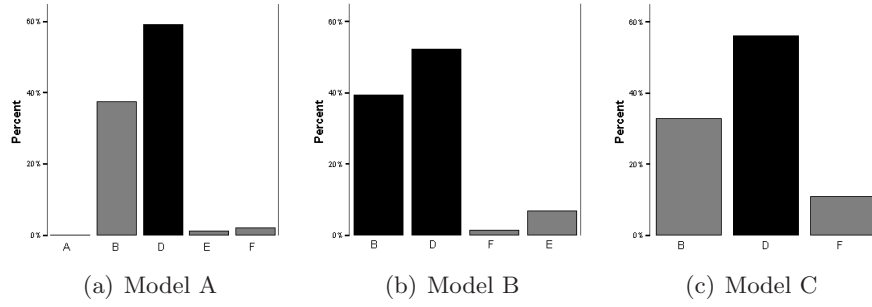


Figure 7: Distribution of document types in small univariate clusters.

In Figure 8 we see the distribution of the purchasing groups in the small clusters of Model A, B and C. Model A highlights three purchasing groups (A, E and F), while Model B and C only put E in the spotlights.

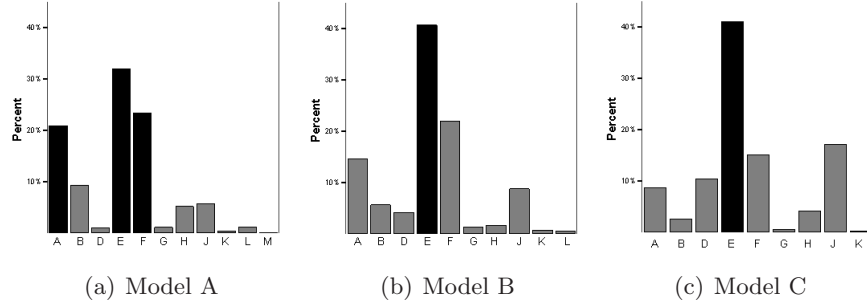


Figure 8: Distribution of purchasing groups in small univariate clusters.

Looking at outliers concerning creators and suppliers, we can see that there is again some overlap, but no complete consistency among the three models. In Figure 9 we find a top three of creators, named C1, C2 and C3. Building further on Model A, these would be the creators interesting to have a closer look at. If we use another attribute to cluster on, for example the clustering attribute of Model B, we would get another composition of creators in the small cluster. This distribution is presented in Figure 10. One creator out of Model A's top three returns with a frequency of 13.4%, namely C2. (This is beyond the scale of this graph, which is set equal to other graphs concerning the frequency of creators.) In spite of this consistency between Model A and B, Model B highlights another top three. The two new creators are named C4 and C5. Turning to Model C, C2 is again represented and now in an even bigger percentage of 25.6%. Also C4 of Model B returns, but again three new creators stand in the spotlight, C6, C7 and C8. So depending on which attribute we choose to perform the univariate clustering, other creators would get our attention.

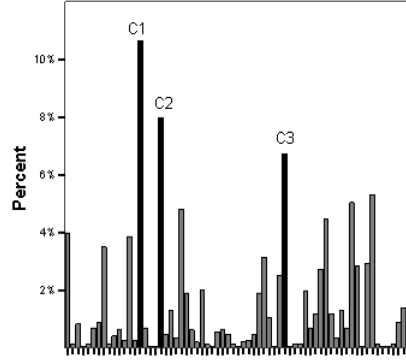


Figure 9: Distribution of creators in small cluster of Model A.

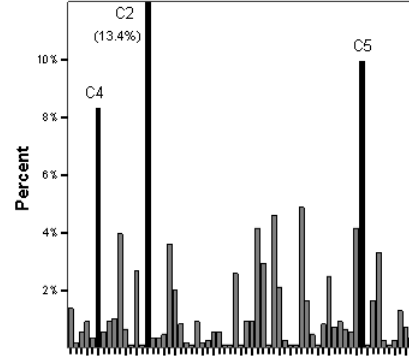


Figure 10: Distribution of creators in small cluster of Model B.

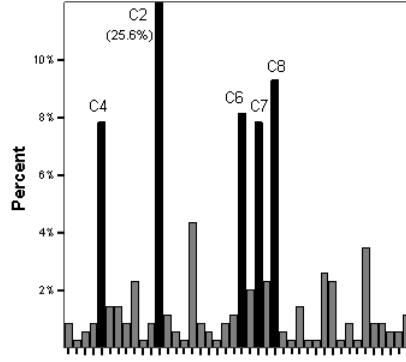


Figure 11: Distribution of creators in small cluster of Model C.

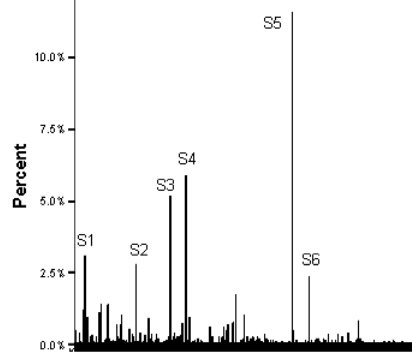


Figure 12: Distribution of suppliers in small cluster of Model A.

The same analysis can be made of the suppliers. The small cluster of Model A highlights a top six, S1 till S6. In the small cluster of Model B, two of those six suppliers, S5 and S6, are again represented in the top three along with the new supplier S7. In yet another composition of the small cluster, based on the clustering attribute of Model C, two more new suppliers would get attention, and some other would not. A top three arises, with S8 and S9 being new suppliers, together with S3 from the top six out of Model A.

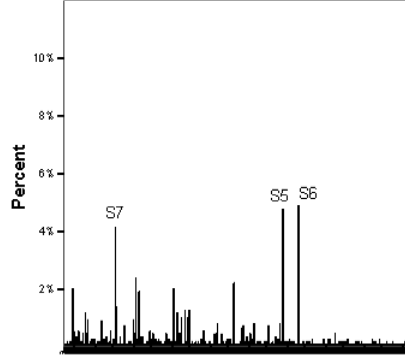


Figure 13: Distribution of suppliers in small cluster of Model B.

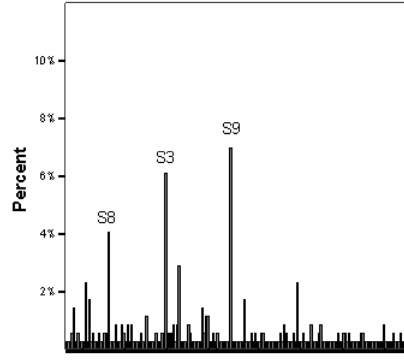


Figure 14: Distribution of suppliers in small cluster of Model C.

## 6 Multivariate Clustering

### 6.1 Model Specifications

The univariate clustering yielded contradictory information, depending on which attribute was taken to cluster on. A multivariate analysis takes several attributes at the same time into account. Before we can apply this analysis, we have to execute the third step of our methodology, namely to translate technical data into attributes that describe behavior. For performing this step, we take into account the particular type of fraud risk we wish to reduce. The fraud risk linked with entering PO's into the ERP system is connected with the number of changes one makes to this PO, and more specifically, the changes made after the last release. There is namely a built-in flexibility in the ERP system to modify released PO's without triggering a new release procedure. For assessing the related risk, we selected four attributes to mine the data. A first attribute is the number of changes a PO is subjected to in total. A second attribute presents the number of changes that is executed on a PO after it was released for the last time. The third attribute we created is the percentage of this last count that is price related. So what percentage of changes made after the last release is related to price issues? This is our third attribute. The last attribute concerns the magnitude of these price changes. Considering the price related changes, we calculate the mean of all price changes per PO and its standard deviation. On itself, no added value was believed to be in it. Every purchaser has its own field of purchases, so cross sectional analysis is not really an option. However, we combine the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) to create a theoretical upper limit per PO of  $\mu+2\sigma$ . Next, we count for each PO how often this theoretical limit was

exceeded. This new attribute is also taken into account in our data mining approach. In this core model, no categorical attributes were added. As a robustness check however, attributes like document type and purchasing group were included in the model. The results did not significantly change by these inclusions.

After the selection of attributes, we need information to set the value of K. We therefor execute the LC clustering algorithm with K set equal to 1 till 5. This yields the BIC values plotted in Figure 15. The BIC values drop heavily until the 3-cluster model. Beyond the 3-cluster model, the decreases are more modest. Based on these values, we decide to use this 3-cluster model.

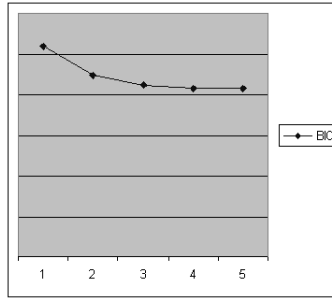


Figure 15: BIC values multivariate clustering.

## 6.2 Results

The profile of the 3-cluster model is presented in Table 4. It gives the mean value of each attribute in each cluster. To compare with the data set as a whole, the mean values of the population are also provided.

Table 4: Profile of data set and 3-cluster model.

	Population	Cluster 1	Cluster 2	Cluster 3
Cluster size	100	0.7663	0.2212	0.0125
Number of changes	4.37	3.3378	6.7608	25.459
Changes after release	0.37	0.0193	1.2376	6.1257
Percentage price related	0.0756	0	0.3185	0.4094
Count over limit	0.01	0.0072	0.0194	0.2725

Looking at the profile of the 3-cluster model, there is an interesting cluster to notice, the third cluster, if it was even only for its size. Cluster 1 comprehends 76.6% of the total data set, cluster 2 22.1% and cluster 3 only 1.25%. Why is there 1.25% of all PO's behaving differently than the remaining PO's? Regarding the mean attribute values of this small cluster, this cluster is, besides from its size, also interesting in terms of fraud risk. The



mean number of changes per PO in this cluster, is 25, as opposed to a mean number of changes of 4 in the data set. Why are these PO's modified so often? Not only are these PO's changed so much in their entire life cycle, they are also modified significantly more after they were last released (6 times) in comparison with the mean PO in the data set (0.37 times). These are odd characteristics. The mean percentage in cluster 3 of changes after the last release that is price related is also the highest percentage of the three clusters (40.9%). All together this means that the average PO in cluster 3 is changed 25 times in total, of which 6 changes occur after the last release and 2.4 of those 6 changes are price related. Concerning the magnitude of the price related changes, we can conclude that these changes of PO's in cluster 3 are more often much larger than the average price change in that PO if we compare this with price related changes of PO's in the other clusters. In cluster 3, there are on average 0.2725 price related changes larger than  $\mu + 2\sigma$  per PO, in comparison with 0.0072 and 0.00194 per PO in cluster 1 and 2 and 0.01 changes in the entire data set.

Taking these numerical characteristics into account, one can conclude that cluster 3 has a profile with a higher fraud risk than the other two clusters.

Numerical attributes tell us that cluster 3 carries a fraud risky profile, but also categorical attributes behave in a different fashion than they behave in the data set as a whole. So there are the creators of the PO. One person for example created 39 out of the 408 PO's from cluster 3 (hereby representing 9.56% of cluster 3), while the same person only created 131 out of the 33.814 PO's, which counts only for 0.39% of the entire data set.

For calculating the probability of taking this person (called xxx) by chance 39 times of 408, given the prior distribution, we use the hypergeometric distribution. This looks as follows.

$$h_m = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}}$$

The hypergeometric distribution is a discrete probability distribution that describes the number of successes  $m$  in a sequence of  $n$  draws without replacement, given a finite population  $N$  with  $M$  successes. In our situation concerning person xxx this leads to:

$$h_{39} = \frac{\binom{131}{39} \binom{33.814-131}{408-39}}{\binom{33.814}{408}} < 1^{-15}$$

So if we select 408 cases at random out of the population of 33.814 observations, there is a probability less than  $1^{-15}$  that we pick 39 cases with user-id xxx, given the prior distribution of 131 successes in the population. This event is very unlikely to happen by coincidence.

Not only creators made such significant increases in representation, but also some suppliers are significantly more represented in cluster 3 than they are in the full data set. We screened all creators and suppliers on significant increases in representation between the data set and cluster 3 with a significance level of  $h < 1^{-5}$ . 14 suppliers and 12 creators met this criterium. Not all of them are however equally important since an increase of 0.03% representation to 0.98% is not as impressive as an increase of 1.47% to 7.6%. Table 5 gives us more insights into the importance of the 14 suppliers and 12 creators.

Table 5: Descriptives of creators and suppliers with a significant higher representation in cluster 3.

Representation (r) in cluster 3	Number of suppliers	Number of creators
$r < 1\%$	4	
$1\% < r < 2\%$	4	
$2.2\% < r < 4.5\%$	3	
$6\% < r < 7.5\%$	3	
Total	14	
$r < 2\%$		3
$2.9\% < r < 3.5\%$		3
$5\% < r < 10\%$		6
Total		12

Not only the creators and suppliers, but looking at Figure 16 and 17 we also find the distributions of purchasing groups and purchasing document types in cluster 3 differing from the distributions in the population. Purchasing group E stands out with its 39.2% in cluster 3, while it was purchasing group D that was highly represented in the total data set (29.1%). Concerning the document types, type B was found most prevalent in the total data set (48.1%), while type D was most prevalent in cluster 3 (64.0%).

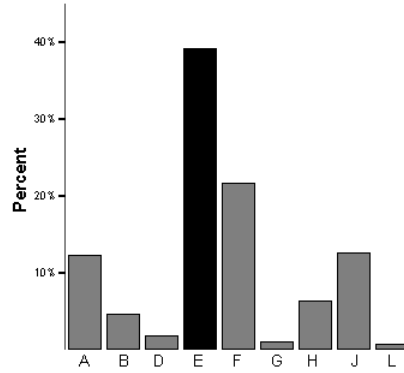


Figure 16: Distribution of purchasing groups in cluster 3.

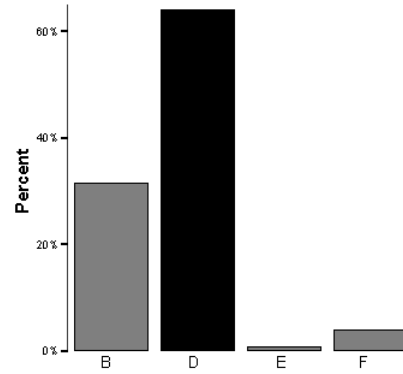


Figure 17: Distribution of document types in cluster 3.

To summarize these results, we find a small cluster with a high fraud risky

profile, due to the large values for the behavior describing profiles. When looking at this small cluster in terms of creators, suppliers, purchasing groups and document type, we find other patterns than in the total data set. It would be interesting to have a closer look at these PO's to find answers on why they behave so differently from the total data set. This would be the fifth step of our methodology.

## 7 Audit by Domain Experts

Since it is more than likely that auditing all 408 PO's of cluster 3 is too time consuming, it would be interesting to take a sample of PO's that are made by one of the creators described above or involve one of those suppliers (or both). The smallest sample to extract from this cluster is to take only those PO's of the six creators and three suppliers that are most represented in the cluster. This yields a sample of 38 PO's. Why is it that they merely induce PO's in this small cluster than in the other two clusters? What makes these purchases this risky? Also the recurrence of a particular purchasing group (E) and purchasing document type (D) can shed an interesting light on deciding which PO's to audit. Auditing this kind of PO's can learn the company a lot about the opportunities that exist to commit fraud, in view of the fraud risky profile that the behavioral attributes describe. However, as already mentioned, the audit step is not (yet) executed for the subset of new PO's (cluster 3), given its size of 408 PO's. At least it is not performed on the data of the small cluster. The possibility LC clustering provides to select observations that do not belong to any cluster is explored. 42 PO's were identified and audited in-depth (the fifth step). There was no uniform profile for these cases. The audit resulted in a few questions with regard to the use of the ERP-system. Nothing however showed misuse of procedures or any other fraud risk.

The entire methodology, provided in Figure 1, is however also applied on the subset of old PO's. The results of the descriptive data mining step are similar to the discussed results. The small interesting cluster (in perspective of a fraud risky profile) of old PO's only contained 10 observations, with nine of them stemming from the same purchasing group and six of them created by the same employee. These 10 observations were audited by domain experts. The results of their investigation are summarized in Table 6.

Table 6: Summary of investigation by domain experts.

Category	Number of cases
Extreme values	0
Fraud	0
Circumventing procedures	9
Errors/Mistakes	1

These are very good results in the light of internal fraud risk reduction since all investigated PO's are contributing to risk assessment by laying bare opportunities. Nine PO's, the ones in the particular purchasing group, are created and modified all over and over again. This is against procedures and makes investigating these PO's very difficult. By creating such complex histories of a PO, the opportunity of committing fraud increases. Only insiders can unravel what really happened with these PO's, since they are such a mess. This off course increases the opportunity and risk of internal fraud. Also, the investigation of this practice has put things in another perspective concerning the separation of functionalities. A follow-up investigation by the audit and investigations department of the case company for this matter is approved.

In the tenth PO a mistake is made. As explained before, a mistake that stays unnoticed creates a window of opportunity for internal fraud. The employee that first makes a mistake by accident, can afterwards consider how to turn this opportunity to one's advantage.

By investigating the 10 selected observations, additional odd practices came to light, which also induced extra investigations. On top of this, the case company gave priority on auditing the procurement cycle in depth.

## 8 Multivariate versus Univariate Analysis

The results of using a multivariate descriptive data mining approach based on behavior describing attributes, provides us with interesting results. In the smaller subset of old PO's we encounter PO's that are changed over and over again. Also in the larger subset, changing the PO a lot of times is a primal characteristic of the selected observations. However, one could wonder if this outcome was not much easier to obtain, simply by applying univariate clustering instead of multivariate clustering. We do not go into the discussion about one method being generally better than another. What we can and want to say however, is that in our case, we did not find the same results by using a univariate LC clustering algorithm as we found by applying the multivariate LC clustering algorithm. Firstly, if we have a look at the profiles of all small clusters, both from the univariate as from the multivariate analysis (Table 7)<sup>3</sup>, the profiles of the univariate small clusters are not as marked as the small cluster 3 we discussed. The marked profile is of high importance in order to make a narrow selection of cases for further auditing in the light of internal fraud risk. The small cluster of Model C is only for 43% incorporated in cluster 3, so this is not a selection of some core

---

<sup>3</sup>The values between brackets are of the univariate clustering attribute and should not be taken into account when assessing the resulting profile.

of cluster 3. We can conclude that the multivariate aspect of our analysis was indispensable to come to the presented profile.

Table 7: Profiles of small clusters from Model A, B, C and the multivariate model.

	Model A	Model B	Model C	Multivariate
Cluster size (PO's)	1,428	1,085	344	408
Number of changes	(16.87)	12.13	17.88	25.46
Changes after release	2.51	(5.18)	7.73	6.13
Percentage price related	0.27	0.40	0.84	0.41
Count over limit	0.09	0.06	0.17	0.27

When we look at the four attributes purchasing group, document type, creator and supplier, the univariate samples do not present the same results either. Concerning the purchasing group, with E most prevalent in cluster 3, only Model B showed a comparable distribution of purchasing groups in its small cluster, although Model C also highlighted purchasing group E, but in another distribution profile. The same comparison can be made for the document type. This time Model A shows a comparable distribution of document types in its small cluster with the distribution in cluster 3, but Model B and C do not. Model B even put both document types B and D in the spotlights, instead of only document type D. Regarding creators and suppliers, the distributions are presented in Figure 18 and 19.

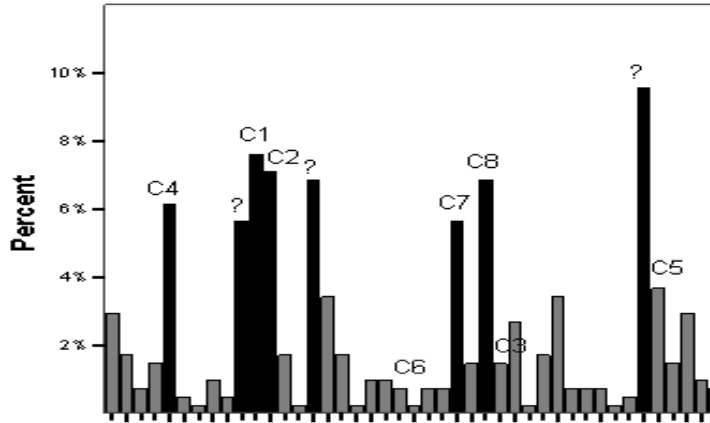


Figure 18: Distribution of creators in cluster 3.

The distribution of creators in cluster 3 shows a top eight. Only five of those eight came to light in the univariate models, taken all of them together. Not one model would have given the same results. Further, the univariate models brought creators forward that in cluster 3 are not that important, like C6, C3 and C5. Concerning the suppliers, a top five is presented in cluster 3, consisting of S1, S3, S9, S5 and S6. Other outliers from the univariate (S8,

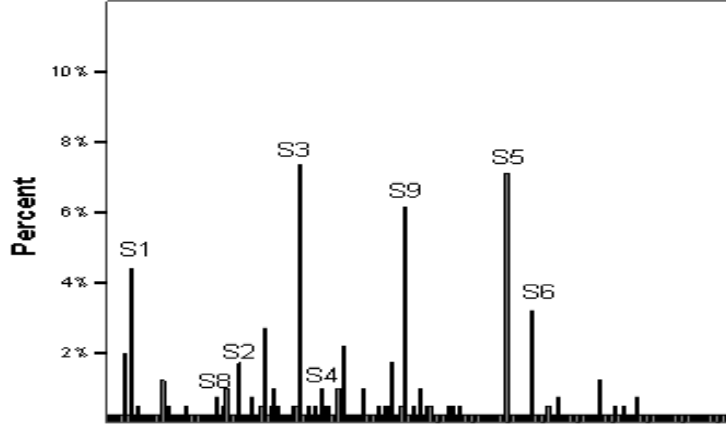


Figure 19: Distribution of suppliers in cluster 3.

S2 and S4) models are pushed back. Again, not one model alone could have presented the same top five suppliers.

To have a view on where the top eight of creators of cluster 3 is situated in the three univariate models, we refer to Figure 20, 21 and 21. The top eight is marked by the letters A through H. In the small cluster of Model A, we find a top three (as already discussed), with two of the creators similar to the top eight of cluster 3 and one new outlier. The rest of the top eight is situated a bit lower in frequencies. But even in that range of frequencies, new creators are put forward in this model. The same situation is found for the outliers of Model B. Two creators of the top three are the same as in the top eight of cluster 3, except that it does not concern the same two creators. Here A and D are in the top three, while in Model A this was C and D. Again, in the lower frequency range the remaining six creators of the top eight were found, along with some new creators. In the small cluster of Model C, a top five creators presents itself, with four out of five coming from cluster 3's top eight. One has even an extreme frequency of 25.6% (again creator D). One of the top eight (C) has fallen very low in frequency, while new creators rise. Regarding the outliers in terms of creators, no univariate model shows the same results as the multivariate model.

After situating the top eight of creators of cluster 3 in the univariate models, we do the same with the top four of suppliers. In Figure 23, 24 and 25 we mark this top four with letters W through Z. Looking at the distribution of suppliers in the small cluster of Model A, we had distinguished a top six, here marked in bold. Three of the top six are also to be found in the top four of cluster 3, and three new outliers are found. One of the top four of cluster 3 (Y) has fallen low. In Model B, we have a top three, with only one supplier similar to the top four of cluster 3, namely supplier Z. Not only are new suppliers put forward in the top three, but also in the frequency range where suppliers X and Y are to be found. Supplier W, a part of the top

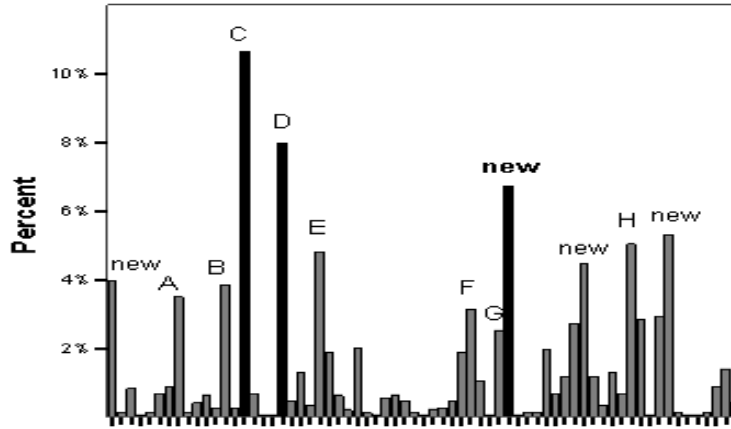


Figure 20: Location of top 8 creators cluster 3 (A-H) in Model A.

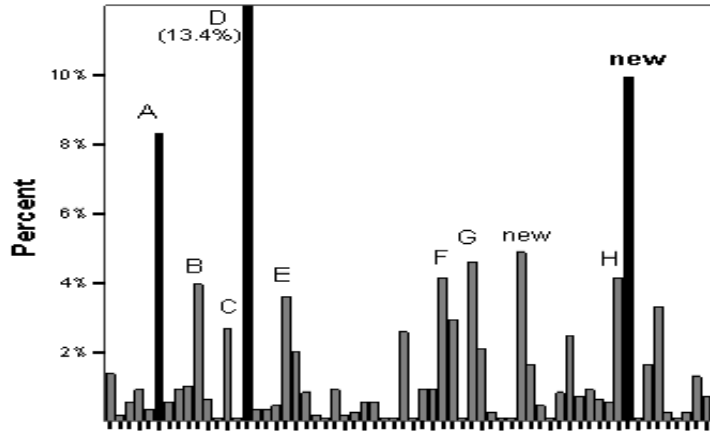


Figure 21: Location of top 8 creators cluster 3 (A-H) in Model B.

four in cluster 3, is in this model not even a peak on the graph. It has a frequency of .6% in this small cluster. The opposite is true for one of the two new outliers in this top three, which is not represented at all in cluster 3. At last, if we look at the suppliers that draw attention in the small cluster of Model C, we find a top three, with two of them also being part of the top four of cluster 3. However, at least five other suppliers outnumber supplier W and Z if it comes to frequency. Like the situation with the creators, also the suppliers could not be represented by a univariate model in the same way as by the multivariate model.

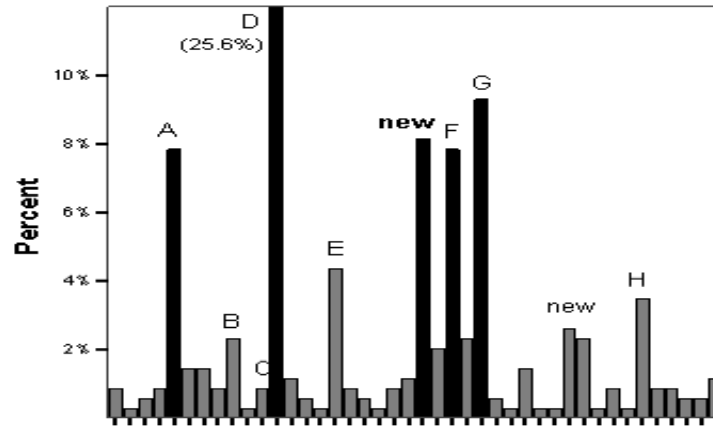


Figure 22: Location of top 8 creators cluster 3 (A-H) in Model C.

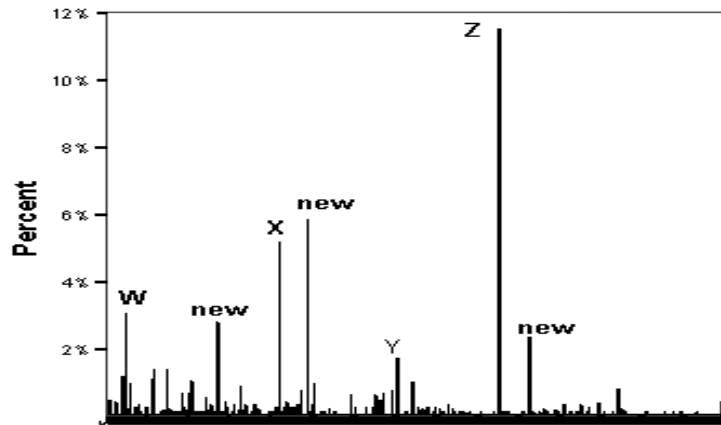


Figure 23: Location of top 4 suppliers cluster 3 (W-Z) in Model A.



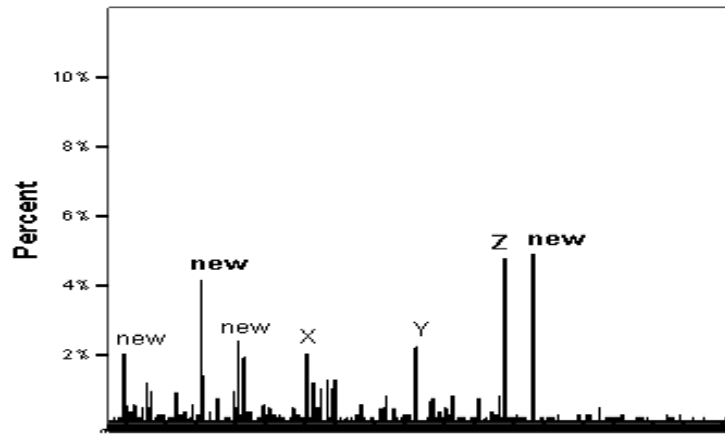


Figure 24: Location of top 4 suppliers cluster 3 (W-Z) in Model B.

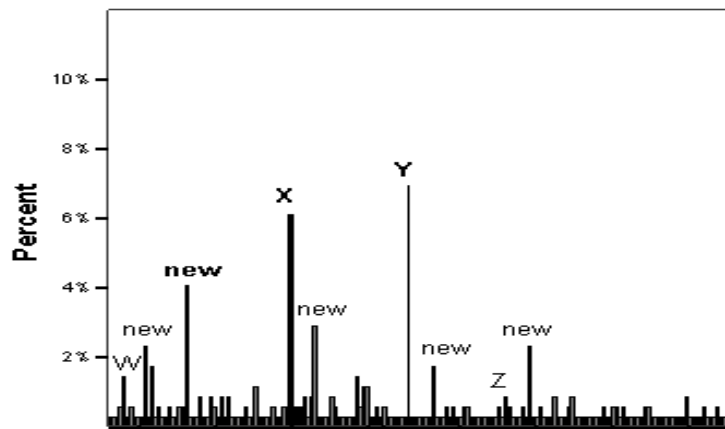


Figure 25: Location of top 4 suppliers cluster 3 (W-Z) in Model C.

## 9 Conclusion

In this paper, a methodology for reducing internal fraud risk is presented. This is a contribution to the literature in that it concerns internal fraud whereas the literature focusses on external fraud. Further we broaden our scope from fraud detection to fraud risk reduction, which encompasses both fraud detection as prevention. We were able to apply our suggested methodology in a top 20 ranked European financial institution. The results of the case study suggest that the use of a descriptive data mining approach and the multivariate latent class clustering technique, can be of additional value to reduce the risk of internal fraud in a company. Using univariate latent class clustering did not yield the same results. The application of the suggested methodology at the case company produced a tone of more concern about the topic of internal fraud along with concern about the opportunity of committing this crime.

## References

- ACFE (2006). 2006 ACFE Report to the nation on occupational fraud and abuse. Technical report, Association of Certified Fraud Examiners.
- Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert (2002). Fraud classification using principal component analysis of RIDITs. *The Journal of Risk and Insurance* 69(3), 341–371.
- Cortes, C., D. Pregibon, and C. Volinsky (2002). Communities of interest. *Intelligent Data Analysis* 6, 211–219.
- Estévez, P., C. Held, and C. Perez (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications* 31, 337–344.
- Fanning, K. and K. Cogger (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management* 7, 21–41.
- Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge University Press.
- Kaplan, D. (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks: Sage Publications.
- Kim, H. and W. J. Kwon (2006). A multi-line insurance fraud recognition system: a government-led approach in Korea. *Risk Management and Insurance Review* 9(2), 131–147.

- Kirkos, E., C. Spathis, and Y. Manolopoulos (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications* 32, 995–1003.
- Lynch, A. and M. Gomaa (2003). Understanding the potential impact of information technology on the susceptibility of organizations to fraudulent employee behaviour. *International Journal of Accounting Information Systems* 4, 295–308.
- Magidson, J. and J. K. Vermunt (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*.
- PwC (2007). Economic crime: people, culture and controls. the 4th biennial global economic crime survey. Technical report, PriceWaterhouse&Coopers.