Ekonomi och samhälle
Economics and Society

Skrifter utgivna vid Svenska handelshögskolan
Publications of the Hanken School of Economics

Nr 255

Argyris Argyrou

# Developing Quantitative Models for Auditing Journal Entries

Helsinki 2013

Developing Quantitative Models for Auditing Journal Entries

Key words: auditing, journal entries, financial statement fraud, self-organizing map, extreme value theory



Argyris Argyrou
Hanken School of Economics
Department of Accounting and Commercial Law
P.O.Box 479, 00101 Helsinki, Finland


Hanken School of Economics
ISBN 978-952-232-194-7 (printed)
ISBN 978-952-232-195-4 (PDF)
ISSN-L 0424-7256
ISSN 0424-7256 (printed)
ISSN 2242-699X (PDF)

Edita Prima Ltd, Helsinki 2013

# Acknowledgements

In completing my PhD, I have incurred a great deal of intellectual debt. I am deeply indebted to Dr. Andriy Andreev, Prof. Barbro Back, and Mr. Anders Tallberg; I could have not completed my PhD without their support and tutelage. I am also grateful to Prof. Pontus Troberg for his advice and guidance as well as to Prof. Jaakko Hollmén for his insightful comments and suggestions. I owe a great debt of gratitude to Mr. Demetris Malais for his unparalleled assistance that has made this thesis feasible.

I also appreciate greatly the generous financial support I have received from the Hanken Foundation and the Graduate School of Accounting.

# Dedication

***I dedicate this Thesis to my nephews, Andreas and Polydoros,***
***in the hope that it will motivate and inspire them.***

# List of publications

1. Argyris Argyrou. Clustering hierarchical data using Self-Organizing Map: A graph-theoretical approach. In Jose C. Principe and Risto Miikkulainen, editors, Advances in Self-Organizing Maps, Volume 5629 of LNCS, pp. 19 - 27. Springer-Verlag, June 2009.

2. Argyris Argyrou and Andriy Andreev. A semi-supervised tool for clustering accounting databases with applications to internal controls. Expert Systems with Applications, 38(9):11176 - 11181, September 2011.

3. Andriy Andreev and Argyris Argyrou. Using Self-Organizing Map for data mining: A synthesis with accounting applications. In Dawn E. Holmes and Lakhmi C. Jain, editors, Data Mining: Foundations and Intelligent Paradigms, Volume 2, chapter 14, pp. 321-342, Springer, September 2011.

4. Argyris Argyrou. Auditing journal entries using Self-Organizing Map. In Proceedings of the 18th Americas Conference on Information Systems (Track: Accounting Information Systems), Seattle, USA, August 2012.

5. Argyris Argyrou. Auditing journal entries using extreme value theory. In Proceedings of the 21st European Conference on Information Systems (Track: Accounting Information Systems), Utrecht, The Netherlands, June 2013.

# A brief note to the readers

In this brief note, I discuss four issues pertaining to the Thesis in order to guide the readers through it, and so that both the Thesis and my stance on it could be better understood.

The Thesis carries out inter-disciplinary research that intersects the domains of Accounting, and Information Systems (I.S.); this intersection is broadly referred to as Accounting Information Systems (A.I.S). Readers in the Accounting discipline could criticise the Thesis on the grounds that it contains more material on I.S. than on accounting. Similarly, readers in the I.S. discipline could make the reverse point. Who is right? Both sides could be right if they only interpret the Thesis within the confines of their respective disciplines. However, I invite the readers to take into account the inter-disciplinary nature of the Thesis and to interpret the findings within the boundaries of A.I.S.

The Thesis does not follow the paradigm of positive accounting research: it does not include any theory that leads to testable and falsifiable hypotheses; neither does it represent theoretical constructs as observable variables; nor does it test whether the variables are statistically significant. However, this is not to say that the Thesis fails to conduct research according to accepted paradigms. The Thesis opts for design science as the apposite paradigm in order to conduct research.

The Thesis uses data that are not available in the public domain. Readers could call into question the authenticity and provenance of the data. Although I can not publish the data for confidentiality reasons, I welcome the opportunity to present the data to the readers, should they consider such presentation to be conducive for their interpreting the Thesis.

The Thesis follows the structure of composite theses, rather than that of doctoral monographs, that is, it consists of five papers and a summary section. Readers could argue that the five papers are not as closely connected with one another as they should be. I accept that there is an element of "disconnectedness" that may reduce the coherence of the Thesis. Part of the "disconnectedness" is attributable to the structure of composite theses; and, I hope, the remaining part can be atoned for by the novelty and quality of the Thesis.

# CONTENTS

TABLES



FIGURES

PART II: PUBLICATIONS

# 1 INTRODUCTION

The Thesis examines how the auditing of journal entries can detect and prevent financial statement fraud; financial statement fraud occurs when an intentional act causes financial statements to be materially misstated (AICPA, 2002). Although it is not a new phenomenon, financial statement fraud has attracted much attention and publicity in the aftermath of numerous cases of financial malfeasance (e.g. ENRON, WorldCom). These cases have undermined investors' confidence to capital markets, resulted in unprecedented economic costs, and led to the conviction of senior executives.

In response to these cases, Sarbanes and Oxley Act (U.S. Congress, 2002) was enacted in order to protect investors by increasing the accuracy and reliability of corporate disclosures. In essence, Section 302 of the Act requires senior management to establish and maintain internal controls, and to certify that financial statements present the results of a company fairly. Section 404 of the Act requires senior management to evaluate, and report on, the effectiveness of internal controls over financial reporting (i.e. ICFR). Further, Section 404 requires auditors to express two opinions in an annual audit report: first, their own opinion on whether a company maintains effective ICFR; and second, an opinion on whether the management's assessment of ICFR is fairly stated. The latter requirement has been relaxed in Auditing Standard No. 5 (Public Company Accounting Oversight Board (PCAOB), 2007); A.S. No. 5 requires auditors to express only their own opinion on the effectiveness of a company's ICFR.

Existing literature has pursued two main lines of inquiry. The first line has sought to identify what conditions are present in financial statement fraud; the conditions can be classified into incentives, opportunities, and attitudes. The second line of inquiry has proposed decision aids in order to assist auditors to estimate fraud risk. The decision aids are based on questionnaires, statistical theory (e.g. logistic regression), and machine learning (e.g. artificial neural networks).

Notwithstanding the breadth of existing literature, little is known on how the auditing of journal entries can detect and prevent material misstatements to financial statements. The lack of knowledge becomes more pronounced given that journal entries are deemed to pose a high risk of material misstatements due to fraud (PCAOB, 2004). It is further exacerbated when considering that auditors are required to test the appropriateness of journal entries recorded in a general ledger (SAS 99; AICPA, 2002). It also contrasts sharply with the principle that controls over the recording and processing of journal entries underpin the completeness, accuracy, and timeliness of financial reporting (Canadian Institute of Chartered Accountants, 2004). Indeed, pertinent literature acknowledges the lack of knowledge (Hogan et al., 2008; Grabski, 2010). A noteworthy exception is a study by Debreceny and Gray (2010), who used digit analysis, or

Benford's Law, to detect fraudulent journal entries.

Journal entries are prone to be manipulated by unscrupulous management who are disposed to perpetrate fraud, as the case of WorldCom has amply demonstrated (Beresford et al., 2003, pp. 17 and 56). Fraud results in considerable costs to a number of parties, for example: perpetrators may be liable to fine or imprisonment; auditors may be exposed to litigation; and, investors may experience negative stock returns. These considerations warrant a better understanding of how the auditing of journal entries can detect and prevent financial statement fraud.

In order to further this line of research, the Thesis adopts the paradigm of design-science research in order to build and evaluate three quantitative models for auditing journal entries. The Thesis first employs self-organizing map (SOM; Kohonen 1982, 1997) and extreme value theory (Coles, 2001) in order to design the models as constructs. Second, it codes the constructs in MATLAB (MATLAB Release R2012a, 2012) to build functioning instantiations; and finally, it evaluates the instantiations by conducting a series of experiments on an accounting dataset. The dataset has been made available by an international shipping company, and it contains the complete set of the company's journal entries for fiscal years 2006 and 2007.

The first model aims at clustering journal entries as a means of supplementing internal control procedures. To this end, it combines SOM, the graph-theoretical approach, which is introduced in the first paper, and the International Accounting Standards Board XBRL Taxonomy (IASCF, 2009). The Taxonomy specifies the semantical relationships between accounting concepts as $child \prec parent$ links. The relationships are quantified by the graph-theoretical approach in order to pre-process hierarchical variables into a numerical representation for SOM-based processing. SOM is employed for its dual ability to perform vector quantization and vector projection in order to identify and visualise meaningful clusters that may exist in an accounting database.

The second model exploits the ability of SOM to perform vector quantization in order to derive a reference model that can describe the behaviour of legitimate journal entries. Subsequently, it calculates the quantization error, a distance metric, between the SOM-based reference model and novel journal entries containing seeded errors. Finally, it considers a journal entry whose quantization error exceeds an optimum threshold as "suspicious".

The third model conjectures that "suspicious" journal entries have a very low probability of occurring and also a monetary amount large enough to materially misstate financial statements. It uses the method of peaks-over-threshold, a subset of extreme value theory, to estimate an optimum threshold that can differentiate the distribution of legitimate from that of "suspicious" journal entries.

The Thesis makes a number of contributions that are relevant to the domain of accounting information systems. It proposes a graph-theoretical approach that can extend SOM to categorical-hierarchical data; the approach can yield SOM grids having higher resolution and cluster validity than alternative approaches can. The Thesis also presents a synthesis of an expansive and fragmented literature pertinent to SOM, discusses the main financial applications of SOM, and demonstrates how SOM can identify meaningful and interesting clusters that may exist in accounting databases.

The Thesis suggests a model that can cluster journal entries in well-separated and homogeneous clusters that can be interpreted within an accounting context. The model can provide a holistic picture of an accounting database at a given point in time and assist managers to assess the accuracy of financial statements. Further, the Thesis proposes two models that can detect those journal entries that may cause financial statements to be materially misstated. The Thesis has a novelty value in that it investigates financial statement fraud from the hitherto unexplored perspective of journal entries. Finally, the findings of the Thesis are likely to have practical applications in accounting, as the models can be implemented as a Computer Assisted Audit Technique (i.e. CAAT). This potentiality is the subject of ongoing research that is being carried out by the author.

In the next section, the Thesis presents the accounting context by reviewing the regulatory environment and the existing literature pertinent to financial statement fraud. Section 3.1 presents the main tenets of design-science research, and Section 3.2 describes the data. Section 3.3 explains how the graph-theoretical approach can be employed to transform categorical-hierarchical data into a numerical representation for SOM-based processing. Sections 3.4 and 3.5 describe the fundamentals of graph theory and SOM, respectively, and, Section 3.6 elaborates on non-parametric bootstrap. Section 4 discusses the main contributions and limitations of the Thesis as well as suggests potential directions for further research. Finally, Section 5 summarises the salient features and contributions of the five papers that are included in the Thesis.

# 2 CONTEXTUAL BACKGROUND

To present the contextual background, the Thesis first discusses the statutory and auditing pronouncements that pertain to financial statement fraud; then, it outlines the economic and legal repercussions of financial statement fraud. Subsequently, the Thesis reviews the two main currents in existing literature and examines the current state of knowledge on how the auditing of journal entries can prevent financial statement fraud.

## 2.1 Regulatory environment

The review of the regulatory environment centres on the Sarbanes and Oxley Act (U.S. Congress, 2002), the Statement on Auditing Standards 99 (AICPA, 2002), and the Auditing Standard No. 5 (PCAOB, 2007). In addition, it summarises the economic costs and legal repercussions that result from financial statement fraud.

The Sarbanes and Oxley Act was enacted to "protect investors by improving the accuracy and reliability of corporate disclosures" (U.S. Congress, 2002, p. 1). In particular, Section 302 of the Act requires the management (i.e. CEO, CFO) of companies that are subject to the Act to establish and maintain internal controls, evaluate the effectiveness of these controls, report the conclusions concerning their evaluations, and disclose significant deficiencies and material weaknesses in the design and operation of internal controls. It also requires management to certify that financial statements and other financial information, included in annual or quarter reports, present fairly the financial conditions and results of operations in all material respects. Further, Section 404(a) of the Act requires management to evaluate, and report on, the effectiveness of internal controls over financial reporting (ICFR) at the end of a fiscal year. Section 404(b) requires external auditors to evaluate, and opine on, the effectiveness of a company's ICFR as well as to attest to, and report on, management's assessment regarding ICFR.

To make the implementation of Section 404 more efficient and effective, PCAOB (2007) replaced Auditing Standard No. 2. with Auditing Standard No. 5 " An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements". In contrast to the former auditing standard, A.S. No. 5 requires auditors to express only their own opinion on the effectiveness of a company's ICFR, as effective ICFR provide assurance concerning the reliability of financial reporting. The audit of ICFR should be integrated with that of financial statements; and, although the two audits have different objectives, auditors should plan and perform audit procedures in order to achieve both sets of objectives.

Further, A.S. No. 5 suggests auditors should evaluate those internal controls that address the risk of material misstatements to financial statements, for example: controls over

significant and unusual journal entries; controls over related party transactions and significant estimates made by management; and, controls that mitigate incentives, opportunities, and attitudes for management to falsify financial results.

The Statement On Auditing Standards 99: Consideration of Fraud in a Financial Statement Audit[1] (AICPA, 2002), which is the interim standard AU Section 316 of PCAOB, defines financial statement fraud as "...an intentional act that results in a material misstatement in financial statements that are the subject of an audit." (AU Section 316 .05)[2]. The presence of "intent" distinguishes "error" from "fraud"; an unintentional misstatement in financial statements is considered to be an instance of error, rather than that of fraud. The auditors are not required to determine "intent"; instead, they are responsible for planning and performing an audit to obtain reasonable assurance about whether financial statements are free of material misstatements, whether caused by error or fraud (AU Section 316 .01).

Misstatements in financial statements, including omissions, are deemed to be material if they could be reasonably expected to influence the economic decisions informed users, considered as a group, take on the basis of financial statements. Further, material misstatements are classified into two categories (AU Section 316 .06): (i) misstatements arising from fraudulent financial reporting, and (ii) misstatements arising from misappropriation of assets. A company can perpetrate the former category by manipulating accounting records that are used in preparing financial statements, omitting significant information from financial statements, and misusing accounting principles. The latter category involves the theft of a company's assets, when such an event causes financial statements not to be presented according to the reigning generally accepted accounting practices (i.e. GAAP).

In addition, AU Section 316 .58 requires auditors, among other things, to test the appropriateness of journal entries recorded in a general ledger and other adjustments made in preparing financial statements. The rationale behind this requirement is that management can misstate financial statements by recording fictitious and inappropriate journal entries as well as adjusting recorded amounts via "top-down" or "paper" journal entries (AU Section 316 .08 and .58). Further, the Standing Advisory Group of PCAOB considers journal entries to be an area that poses a high risk of material misstatement in

[1]SAS 99 supersedes SAS 82, which superseded SAS 53. In brief, SAS 53 distinguishes between "irregularities" and "errors" depending on whether an act that causes financial statements to be materially misstated is intentional or not; "irregularities" is classified into management fraud and defalcation of assets. SAS 82 uses the same terminology as SAS 99 does.

[2]In contrast to AU Section 316, the International Standard on Auditing 240 (International Federation of Accountants (IFAC), 2009) defines fraud as "An intentional act ... to obtain an unjust or illegal advantage." However, the AICPA's Auditing Standards Board states that the difference in the definition does not create significant differences between the application of ISA 240 and AU Section 316 (redrafted) (AICPA, 2012). It also states that the AU Section 316 (redrafted) does not change nor expand extant AU Section 316 in any significant respect (AICPA, 2013).

financial statements due to fraud (PCAOB, 2004)[3].

COSO (2010) estimates that the prevalence of fraudulent financial reporting is about six times that of misappropriation of assets, and their combined monetary magnitude to be about US$ 120 billion. Companies may resort to fraudulent financial reporting in order to meet external or internal earnings expectations, conceal a deteriorating financial condition, increase share price, and secure equity or debt financing on favourable terms. To perpetrate fraudulent financial reporting, companies can employ two main techniques. First, improper revenue recognition involves creating fictitious revenue transactions and recording revenue prematurely; second, overstating existing assets or capitalising items that should have been otherwise expensed. COSO (2010) emphasises that fraudulent financial reporting entails considerable legal and economic costs for companies, directors, and shareholders as well as undermines the public confidence in capital markets. The story of Arthur Andersen serves as a poignant reminder.

Palmrose (1987) examined 472 legal cases against the fifteen largest U.S audit firms in the period from 1960 to 1985. The study found that management fraud accounted for 43% (201 cases out of 472) of the legal cases, whereas business failures accounted for 18%, and the rest 39% was classified as other. Palmrose (1991) investigated the extent to which legal cases against auditors become public knowledge in order to explain how this knowledge may impose indirect costs (e.g. loss of reputation) to the litigants. The study investigated 761 legal cases in the period from 1960 to 1990; it found that about 20% of the cases were disclosed to the public, and that high-profile cases involved management fraud and financial distress.

## 2.2 Pertinent literature

Existing literature has followed two main lines of research. First, it has sought to identify the conditions that are present in, albeit not necessarily causative of, financial statement fraud. The conditions can be classified into three categories forming a “fraud triangle” (AICPA, 2002): incentives and pressure to commit fraud; opportunities for committing fraud; and attitudes towards fraud. The second line of research has striven to propose decision aids (e.g. models based on logistic regression) that may enhance the ability of auditors to detect and prevent financial statement fraud. These two lines of research serve as the organizing principles for the review that follows.

To investigate the link between equity-based remuneration and occurrences of financial statement fraud, Erickson et al. (2006) used accounting and auditing enforcement releases (i.e. AAERs)[4] as a proxy to select fifty fraudulent companies in the period from

[3]For completeness, the high-risk areas also include: revenue recognition, significant or unusual accruals, estimates of fair value, related party transactions, and quarterly financial statements.

[4]AAERs are issued by the U.S. SEC: http://www.sec.gov/divisions/enforce/friactions.shtml

1996 to 2003. The authors selected a paired-matched sample of 100 non-fraudulent companies as well as a second unmatched sample. In either sample, the authors found no evidence to support the claim that executive equity incentives could increase the likelihood of financial statement fraud. On the other hand, Efendi et al. (2007) found that CEOs having stock options that were "in-the-money" were more likely to issue financial statements containing an accounting irregularity; the term "accounting irregularity" included both intentional (i.e. fraud) and unintentional (i.e. errors) acts. Burns and Kedia (2006) documented that CEOs whose option portfolio was sensitive to stock price were more likely to engage in accounting practices that could cause financial statements to be subsequently restated.

In a similar line of inquiry, Armstrong et al. (2010) selected a sample consisting of three types of accounting irregularities that took place from 2001 to 2005: restatement of financial statements as a result of accounting manipulations; firms accused of accounting manipulation in a lawsuit; and, firms accused of accounting manipulation in an AAER. The authors adopted a matched-pairs design, based on propensity score, that was robust to a potential misspecification of the functional relationship between control variables and the outcome variable. The study found no evidence to support a positive association between CEO equity-based incentives and accounting irregularities.

Summers and Sweeney (1998) examined whether insider trading could act as a risk-factor auditors could use to estimate the likelihood of fraudulent financial reporting. The results suggested that insiders at fraudulent companies reduced their net position by selling shares. The study acknowledged certain limitations, such as: it had a selection bias towards the news-worthiness of fraud; the fraud sample did not include undetected fraud nor fraud discovered during an audit; and no hold-out sample was used to validate the results.

Conditions that may provide executives with an opportunity to commit financial statement fraud include: related party transactions; ineffective monitoring; lack of internal controls; the ability of management to override internal controls; and, complex organisational structures (AICPA, 2002). Overwhelming evidence has indicated that the companies most inclined to commit financial statement fraud are characterised by poor corporate governance, weak internal controls, and management's ability to override internal controls (Loebbecke et al., 1989; Dechow et al., 1996; Farber, 2005). For example, the corporate governance of WorldCom was found to have contributed to the fraud that was perpetrated at the company by discouraging dissent, restricting access to financial information, and lacking ethical business practices (Beresford et al., 2003, pp. 18-19).

Existing literature has also examined whether the use of decision aids could enhance

the ability of auditors to detect fraud. Decisions aids include questionnaires, checklists, linguistic analyses, and analytical procedures.

Pincus (1989) examined whether the use of red-flag questionnaires could aid auditors to assess the risk of material misstatement. The results suggested that both groups of auditors - users and non-users of questionnaires - assessed the risk to be higher for the fraud cases than for the non-fraud cases. Although the two groups performed similarly for the non-fraud cases, the non-users outperformed the users in assessing the risk for the fraud cases. The study offered two reasons to explain the counter-intuitive result: first, non-users considered relevant information users did not, because the information was not included in the questionnaire; and second, users did not weigh the red-flags properly.

The findings of (Pincus, 1989) have been extended and corroborated by Asare and Wright (2004) who found that auditors using a standard risk checklist, organised according to SAS 82, made a lower risk assessment than the auditors without a checklist did.

Loebbecke and Willingham (1988), cited in (Loebbecke et al., 1989), compared the risk factors, contained in SAS 53, against a number of AAERs in order to assess which factors could best identify material management fraud. The study organised the risk factors into three categories: (i) conditions that may lead to fraud, (ii) motivation to commit fraud, and (iii) attitude that can accommodate dishonest behaviour. Further, it proposed a model for assessing the likelihood of material management fraud. To examine whether the model could be applied as an engagement tool by auditors, Loebbecke et al. (1989) conducted a survey of audit partners of KPMG. However, the authors could assess neither the discriminatory power of individual factors nor that of their model in its entirety, because the survey covered only cases of fraud (Bell and Carcello, 2000).

To estimate the likelihood of fraudulent financial reporting, Bell and Carcello (2000) developed a model based on logistic regression and seven risk-factors that were found to be significantly correlated with the occurrence of fraud. The model could supplement auditors' unaided risk assessment; further, it could provide empirical evidence on how effective individual risk factors were in discriminating between fraudulent and non-fraudulent audit engagements. However, the model's discriminatory power may have been overstated as a result of possible hindsight bias introduced by the auditors, who were involved in fraudulent engagements.

Apostolou et al. (2001) used an Analytic Hierarchy Process to assess how much importance auditors placed on the risk-factors, specified in SAS 82, when they assessed the risk of material misstatements to financial statements due to management fraud. The study found that auditors valued the following three categories of risk-factors, in an

ascending order of importance: (i) management characteristics and influence over the control environment, (ii) operating and financial stability characteristics, and (iii) industry conditions characteristics.

Goel et al. (2010) examined the verbal content and presentation style of annual reports in order to identify the linguistic features that may distinguish fraudulent from non-fraudulent annual reports. The sample consisted of 126 fraudulent and 622 non-fraudulent companies matched by industry, year, and size. The study found that fraudulent annual reports contained more passive voice, higher lexical variety, and more uncertainty markers than non-fraudulent annual reports did.

Humpherys et al. (2011) analysed the text contained in the Management's Discussion and Analysis (MDA) section of 10-K filings. The study used the AAERs issued from 1995 to 2004 as a proxy for fraudulent companies; the sample consisted of 101 fraudulent companies and an equal number of non-fraudulent companies matched by industry and year. The study found that fraudulent MDA contained more active language, had lower lexical diversity, and used more complex words than non-fraudulent MDA did.

A considerable body of literature has investigated whether analytical review procedures (e.g. ratio analysis, machine learning algorithms) could enhance the ability of auditors to assess fraud risk. Auditors can use analytical review procedures (i.e. ARP) to form expectations about recorded balances and then judge whether the differences between recorded and expected balances are reasonable.

Kaminski et al. (2004) investigated whether ratio analysis could discriminate fraudulent from non-fraudulent financial reporting. The study selected 79 fraudulent companies by using AAERs, issued between 1982 and 1999, as a proxy and an equal number of non-fraudulent companies matched by size, industry, and year. The study concluded that ratio analysis could not detect fraudulent financial reporting. However, it qualified its conclusions by acknowledging a number of limitations that resulted from the small sample size, the choice of ratios in the absence of a theoretical foundation, and the use of AEERs as a proxy for fraudulent financial reporting.

Coakley (1995) proposed a model, based on artificial neural networks (i.e. ANN), for analysing patterns that may be present in the fluctuation of ratios. The ANN-based model could detect material errors in account balances more reliably than traditional ARP could. In the same line of enquiry, Koskivaara (2000) put forward an ANN-based model that could forecast monthly account balances as well as detect seeded errors. In addition, Koskivaara and Back (2007) proposed an ANN-based prototype, coded in a Linux environment, that could model the dynamics of monthly account balances and be applied in the domain of continuous auditing.

To detect financial statement fraud, Green and Choi (1997) developed a model based on

ANN and five ratios. The study used AAERs, issued from 1982 to 1990, as a proxy in order to select 79 fraudulent companies, and a matched-pairs design to select an equal number of non-fraudulent companies. The study showed that the proposed model had lower error rates (i.e. Type I and II) than a random classifier had. However, the comparison may be neither valid nor realistic, as experienced auditors are able to exercise professional judgement, and thus perform better than a random classifier does.

In the end, analytical review procedures (i.e. ARP) are not likely to detect financial statement fraud, because management can manipulate accounting records to conjure expected, and eliminate unexpected, variations. A typical example is Arthur Andersen, who was criticised for relying excessively on ARP in order to detect accounting irregularities at WorldCom (Beresford et al., 2003, p. 236).

Existing literature has acknowledged certain limitations. Using AAERs as a proxy for financial statement fraud may induce selection bias for mainly two reasons. First, AAERs do not report instances of fraud that may occur in companies that are not listed nor do they report fraud that was detected and rectified in the process of an audit. Second, SEC may issue AAERs only for the most egregious cases. Further, a matched-pairs design relies heavily on assumptions about the functional relationship between explanatory and outcome variables (Armstrong et al., 2010). The presence of heterogeneous data and the multi-faceted tenor of financial statement fraud may hinder researchers in their efforts to construct models that can detect fraud. Two additional issues, seldom addressed in the literature, are the asymmetrical misclassification costs of Type I and Type II errors as well as the prevalence, or prior probability, of fraud in the population. The importance of these issues should not be underestimated, as they affect the classification accuracy of models.

## 2.3 Financial statement fraud and journal entries

The fraud perpetrated at WorldCom exemplifies how journal entries can be manipulated to achieve, albeit artificially, expected revenue growth. In brief, journal entries were recorded to reduce operating line costs by capitalising these costs and improperly releasing accruals. It is estimated that about US$9.25 billion - US$ 7.3 billion of which refer to operating line costs - were recorded in false or unsupported journal entries from 1999 to 2002 (Beresford et al., 2003, pp. 17 and 56).

A synthesis of the literature pertaining to financial statement fraud acknowledges that " We were unable to find any research that directly addresses … the use of unusual or top-level journal entries and financial statement fraud"; and, it suggests that additional research is warranted "… in understanding the process of journal entry review to detect and prevent fraud" (Hogan et al., 2008, p.244).

The lack of published research may be due to the difficulties researchers encounter in gaining access to journal entries, as databases are proprietary. Further, in order to conceal fraudulent journal entries, management may intentionally forbid third parties to access a database. While this reasoning may appear implausible, it illustrates why WorldCom did not allow Arthur Andersen to access their computerised general ledger (Beresford et al., 2003, p.247). Had Arthur Andersen been given access, they could have detected the fraud (Beresford et al., 2003, p.236); the fraud was eventually discovered by the Internal Audit Function, because they had access to the computerised general ledger (Beresford et al., 2003, p.247).

A noteworthy study is that by Debreceny and Gray (2010) who used digit analysis, or Benford's Law, to detect fraudulent journal entries. In essence, the authors estimated the difference between the observed distribution of the first digit of US$ amounts and that expected by Benford's Law. If the difference was statistically significant under a chi-square test, then the US$ amount was deemed to be suspicious. The results suggested that the differences were significant for all entities in the sample. One possible explanation is that the results may be an artefact of the chi-square test, as a large number of observations can induce statistically significant results (Grabski, 2010). A further explanation may be that either fraudulent journal entries were the norm in the sample or that Benford's Law is not applicable to journal entries (Grabski, 2010).

In addition, Debreceny and Gray (2010) employed Hartigan's dip test in order to investigate whether the last digits of US$ amounts followed a uniform distribution, which was expected. The authors suggested (ibid. Table 8, p. 174) that the distributions of last digits were not uniform for about 2/3 of the entities in the sample. These entities had a statistically significant dip test ($p<0.01$); the result implied the distributions were at least bimodal.

# 3 RESEARCH DESIGN AND METHODOLOGY

The Thesis adopts the paradigm of design-science research, Section 3.1, in order to build and evaluate three quantitative models for auditing journal entries. To build the models, it employs SOM, Section 3.5, a graph-theoretical approach, Sections 3.3 and 3.4, and extreme value theory. It evaluates the models by conducting a series of experiments on a dataset containing journal entries, Section 3.2. To conduct the experiments, it employs non-parametric bootstrap, Section 3.6, in order to estimate parameters of interest and their associated confidence intervals.

## 3.1 Conceptual framework

The foundations of design-science research have been laid down in "*The Sciences of the Artificial*", first published in 1969, by H.A. Simon, who draws a distinction between natural science and science of design. The former addresses "how things are", whereas the latter deals with "how things ought to be" (Simon, 1996, p.114). This distinction reflects the contrast between descriptive and normative approaches to research (Simon, 1996, p.5).

These foundations paved the way for March and Smith (1995)[5] to propose a design-science research framework in the domain of Information Systems (I.S). The framework encompasses the research activities of natural science and design science as well as the outputs produced by design science. Natural science aims at understanding reality by proposing scientific claims and then justifying their validity. By contrast, design science aims at constructing or building tools for a purpose and then evaluating the performance of the tools against suitable criteria.

Design science can produce four research outputs or artefacts. First, constructs describe a problem within its domain. Second, models capture the semantical relationships between constructs and provide a representation, or an abstraction of reality, that is useful for an artefact to achieve its intended purpose. Third, methods consist of a set of steps (e.g. algorithms) to perform a task; and fourth, instantiations constitute the realisation of an artefact in its intended environment.

The foregoing framework was extended by Hevner et al. (2004), who proposed seven guidelines on how researchers should conduct and evaluate design-science research. First, design-science research must produce a purposeful artefact defined in terms of constructs, models, methods, or instantiations. Second, it must solve an important and real problem. Third, to evaluate the performance of an artefact, design-science research must define appropriate criteria, collect and analyse data, and use methodologies that

[5]This paper was first presented at the Second Annual Workshop on Information Technology and Systems (WITS 1992), Dallas, USA, December 1992.

exist in the underlying knowledge base. Fourth, design-science research can contribute to research in three ways: (i) applying existing knowledge in new and innovative ways, (ii) extending and improving existing foundations in the knowledge base (e.g. algorithms), and (iii) developing performance criteria and evaluative methods. Fifth, the scientific rigour depends on selecting and applying suitable techniques to construct, and specifying appropriate criteria to evaluate, artefacts. Sixth, design-science research must utilise available means in order to solve a problem. Finally, it must be communicated in sufficient detail to allow an audience to understand how an artefact was constructed and evaluated as well as to decide whether the artefact could be applied in a specific context.

To extend the I.S research framework in (Hevner et al., 2004, p.80), Hevner (2007) identified three inter-connected cycles of research activities in order to provide researchers with additional insights on how to perform design-science research. First, the relevance cycle connects the two design-science processes (i.e. build and evaluate) to the environment that provides the research problems to be addressed as well as specifies the criteria for evaluating the research output. The relevance cycle aims at establishing how an artefact can be tested and evaluated in the environment in which the artefact operates. Second, the rigour cycle allows the design-science processes to be grounded in existing methods and theories, which are drawn from the knowledge base of the research domain. It also clarifies the contribution design-science research makes to the knowledge base (e.g. new methods, peer-reviewed literature). Finally, the design cycle connects the two aforesaid cycles, and iterates between designing and evaluating an artefact until it can accomplish a satisfactory design that meets the requirements of relevance and rigour.

In order to clarify the paradigm of design-science research, the *MIS Quarterly* published a special issue (Vol. 32 No. 4 December 2008) containing five papers that satisfied three criteria (March and Storey, 2008, p. 727): (i) address a distinct and important problem, (ii) develop a novel solution, and (iii) evaluate the proposed solution. The editors of the issue suggested that the authors should conduct and evaluate design-science research according to the framework presented in (Hevner et al., 2004).

The “evaluate” activity of design science calls for the development of suitable criteria for evaluating how well artefacts can perform in the environments in which they operate. A set of criteria was proposed by (March and Smith, 1995), as shown in Table 1. However, these criteria have two main shortcomings in that they lack a theoretical justification and fail to provide a holistic evaluation of artefacts (Aier and Fischer, 2011). To overcome these shortcomings, Aier and Fischer (2011) built on Kuhn’s five criteria for scientific progress in order to propose a set of six criteria for evaluating the progress of I.S design theories. Further, the authors compared this set of criteria against that proposed by (March

and Smith, 1995), as shown in Table 1.

Table 1: Criteria for evaluating outputs of design science and I.S design theories

| | Research outputs of design science | | | | I.S Design theories |
|---|---|---|---|---|---|
| March and Smith (1995) | Construct | Model | Method | Instantiation | Aier and Fischer (2011) |
| Completeness | X | X | | | Internal Consistency |
| Ease of use | X | | X | | Utility, Simplicity |
| Effectiveness | | | | X | Utility |
| Efficiency | | | X | X | Utility |
| Elegance | X | | | | Simplicity |
| Fidelity with real world phenomena | | X | | | External consistency |
| Generality | | | X | | Broad purpose and scope |
| Impact on the environment and on the artefacts' users | | | | X | Utility |
| Internal consistency | | X | | | Internal Consistency |
| Level of detail | | X | | | Internal Consistency |
| Operationality | | | X | | Utility |
| Robustness | | X | | | Broad purpose and scope |
| Simplicity | X | | | | Simplicity |
| Understandability | X | | | | Simplicity |
| | | | | | Fruitfulness of further research |
| Järvinen (2001, p.111) | | | | | |
| Communication | X | | | | |
| Form and content | | X | | | |
| Richness of knowledge representation | | X | | | |
| Experiences of users | | X | | | |
| Application domain | | | X | | |
| Positive and negative unanticipated outcomes | | | | X | |
| Impact on social, political, and historical contexts | | | | X | |
| Cost-benefit | | | | X | |

This table combines tables 5 and 10 (Aier and Fischer, 2011, p.143 and p.163), and table 5.2 (Järvinen, 2001, p.111)

In order to explain how design-science research can solve important and relevant problems in I.S, Gregor and Hevner (2011) elaborated on three inter-dependent issues. First, the authors defined the problem space according to (Hevner et al., 2004, p.79) as "... people, organizations and their existing or planned technologies". They also proposed that the definition of artefacts should be broadened to include any designed solution to a problem. Second, the authors identified the four main tenets of design-science research: identify a problem (i.e. research question); build an artefact to solve the problem; evaluate the artefact; and demonstrate research contribution. Third, the authors argued that a research contribution can be made to either, or to all, of the following three levels: (i) specific instantiations (e.g. products, processes), (ii) design principles (e.g. constructs, models, methods), and (iii) emergent design theories.

A recurring and pervasive theme in the literature focuses on establishing suitable criteria for differentiating between activities that constitute design-science research and those activities that involve only development and consultancy. To this end, David et al. (2002) suggested three criteria. The first criterion considers whether the proposed research is novel given the current status of the domain; for example, using new tools (e.g. programming language) to implement an existing artefact is considered to be

development, rather than research. The second criterion states that studying difficult rather than simple problems can yield more knowledge. Finally, when research aims at improving an existing artefact, it must demonstrate that the new artefact performs better than the existing artefact does.

In discussing the methodology of design science, Iivari (2010, p.53) has argued that given the two research activities - build and evaluate - the former is not as well understood as the latter is. Although the two predominant frameworks, (March and Smith, 1995; Hevner et al., 2004), suggest methods for evaluating artefacts, they offer little guidance for building artefacts. Iivari (2010) proposed that the application of transparent and rigorous constructive research[6] methods in building artefacts can differentiate design-science research from inventions built by practitioners.

To address this objective, Iivari (2010) put forward four guidelines in order to make the construction process as transparent as possible. The first guideline considers the practical relevance and potentiality of a research problem; the second addresses the role of existing artefacts in evaluating the incremental contribution of improved artefacts. The third guideline proposes the use of analogies and metaphors for stimulating creativity (e.g. neural networks, genetic algorithms); and, the final guideline calls for kernel theories to support and inform design science.

Kasanen et al. (1993) introduced constructive research as a valid methodology in the context of management accounting. In this context, constructive research involves solving explicit managerial problems either in novel or improved ways, demonstrating the novelty and working of the solution, showing how the solution connects to a theoretical framework, and explaining how it enhances knowledge. Two well-known "constructions" of solving managerial problems are the Activity Based Costing and the Balance Scorecard. In either case, the authors have articulated the "constructions" in accounting theory, demonstrated the functioning of the "constructions" in practice, and paved the way for additional research.

## 3.2 Data description

To select journal entries, the Thesis first obtained access to the database of an international shipping company. It then exported the complete set of journal entries for fiscal years 2006 and 2007 to a text file; an extract from the text file is shown in Appendix A. The text file consists of 61,908 lines and 19 columns representing

[6]Iivari (1991) introduced constructive research methodology in order to extend the "nomothetic" and "idiographic" methodologies identified by (Burrell and Morgan, 1979). In essence, "nomothetic" research methods (e.g. experiments, mathematical analyses) are often employed by natural sciences to test the veracity of hypotheses according to the hypothetico-deductive method; "idiographic" methods (e.g. case studies, action research) focus on exploring the background of the subjects being investigated. By contrast, constructive research methodology "constructs" a new reality in terms of conceptual models, frameworks, and technical artefacts (e.g. softwares).

individual accounting transactions and variables, respectively; the variables are defined in Appendix B. The individual accounting transactions (i.e. lines within journal entries) become the unit of analysis for the proposed models. In addition, the Thesis exported the Chart of Accounts, as depicted in Appendix C. Individual accounts were excluded, because they contain confidential information (e.g. names of shareholders).

The Thesis combines the two sources of information, journal entries and Chart of Accounts, in order to select eight variables, as described in Table 2. It aggregates the transactions according to "Account Class" and calculates descriptive statistics for this variable, as shown in Table 4 on page 32 and Table 5 on page 33.

The Thesis is not able to describe how the case company implements the system of double-entry bookkeeping in its computerised accounting system. The reason is the case company uses SAP, which is administered by I.T. professionals; it would be a very difficult task for the Thesis to describe such a complicated Enterprise Resource Planning system.

Table 2: Description of variables

| Name | Type | Unique values |
|---|---|---|
| Account Number | Alphanumeric | 360 |
| Account Description | Text | 360 |
| Posting Date | Date | 24 |
| Debit-Credit Indicator | Binary | 2 |
| USD Amount | Numerical | |
| Transaction Details | Text | |
| Account Class | Categorical-hierarchical | 30 |
| Account Category | Categorical-hierarchical | 7 |

The Thesis can not cluster the accounting transactions directly, because the categorical hierarchical variables (e.g. "Account Class") lack quantitative information for SOM to calculate a similarity metric (e.g. Euclidean distance). For this reason, the Thesis must first pre-process these variables into a numerical representation that can be amenable to SOM-based processing.

## 3.3 Data pre-processing: A graph-theoretical approach

SOM has been extended to non-metric spaces, for example: WEBSOM for text documents (Kohonen et al., 2000) and (Lagus et al., 2004); SOM for symbol strings (Kohonen and Somervuo, 1998, 2002); and SOM for categorical data (Hsu, 2006). However, these extensions can not capture the semantical relationships that exist between hierarchical-categorical data, and hence they are not appropriate for the purposes of the Thesis.

Consequently, the Thesis proposes a graph-theoretical approach in order to pre-process

the hierarchical-categorical data into a numerical representation that takes the form of a distance matrix. The rows of the distance matrix become the input vectors to SOM.

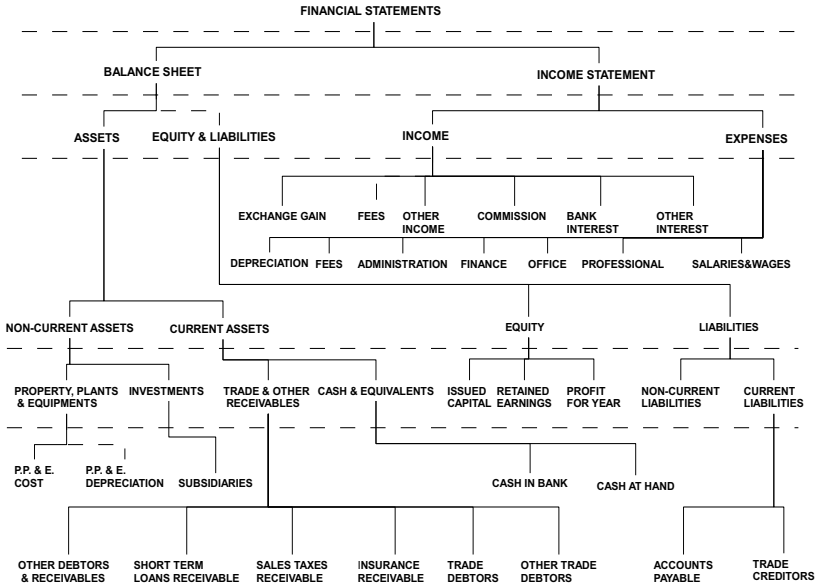



Figure 1: The Accounting Tree, a partial graphical instantiation of the IFRS XBRL Taxonomy

The Thesis first encodes "Account Class" as a directed acyclic graph (i.e. DAG), depicted in Fig. 1. The "Accounting Tree" does not aim at describing accounting concepts in a pictorial format. Instead, it aims at capturing and maintaining the semantical relationships that exist between accounting concepts; the relationships are specified in terms of *child* $\prec$ *parent* links in the IASB XBRL Taxonomy[7] (IASCF, 2009). The root vertex, "Financial Statements", represents the complete set of accounting concepts, and all other vertices are ordered in such a way that each vertex represents a sub-set of its parent vertex. For example, "Bank Account" $\prec$ "Cash in Bank" $\prec$ "Cash and Cash Equivalents" $\prec$ "Current Assets" constitute *child* $\prec$ *parent* relationships specified in the IASB XBRL Taxonomy.

In the second step, the Thesis quantifies the *child* $\prec$ *parent* relationships by using

[7]The Thesis uses the Taxonomy only as a component for the graph-theoretical approach. The approach is introduced in the first paper, and applied in the second and third papers; the fourth and fifth papers do not use this approach.

Dijkstra's algorithm[8] (Dijkstra, 1959). This operation yields a distance matrix[9] that satisfies the conditions of a metric space, as specified in Section 3.4, and thus it can become the input dataset to SOM.

## 3.4 Graph-theoretical approach

The Thesis proposes a data pre-processing approach, based on graph theory, that can capture, maintain, and quantify the semantical relationships that exist between hierarchical data. In doing so, the graph-theoretical approach can transform hierarchical data into a numerical representation, and thus render them amenable for SOM-based processing. The graph-theoretical approach is predicated on two assumptions: first, the hierarchical data conform to a taxonomy that describes the semantical relationships between the data, (e.g. IASB XBRL Taxonomy); and second, these relationships are static.

Table 3: Graph theory: notations and definitions

| | |
|---|---|
| $G = (V, E, w)$ | A graph |
| $V = \{v_1, v_2, \dots v_n\}$ | Set of vertices |
| $E = \{e_1, e_2, \dots e_m\}$ | Set of edges |
| $w : E \rightarrow \mathbb{R}^+$ | Function assigning a positive real number to an edge |
| $\mid V \mid$ | Degree of graph, cardinality of $V$ |
| $\mid E \mid$ | Order of graph, cardinality of $E$ |
| $e = \left\{v_i, v_j\right\}$ | Edge connecting vertices $v_i$ and $v_j$ |
| $d_{ij} = w(e)$ | Distance between $v_i$ and $v_j$ |
| $D = (d_{ij})_{nn}$ | Distance matrix |

For the purposes of the Thesis, a tree is defined as a type of graph, $G = (V, E, w)$, that satisfies at least two of the following three necessary and sufficient properties: (i) $G$ is acyclic, (ii) $G$ is connected, and (iii) $\mid E \mid = \mid V \mid -1$; any two of these properties imply the third (Jungnickel, 2002, p. 8). Let $T = (V, E, w)$ be a tree that is: (i) rooted, with $v_0$ the root vertex, and (ii) ordered, which means that there is a linear ordering of its vertices such that for each edge $e = \left\{v_i, v_j\right\}$ then $v_i < v_j$. It can be deduced that in $T$: (i) all vertices excluding $v_0$ have at most one "parent" vertex, (ii) at least one vertex has no "child" vertices, and (iii) there is a unique path between any two vertices. A level-order traversal of a tree starts from the root vertex, $v_0$, and proceeds from left-to-right to visit each vertex at distance $d$ from $v_0$ before it visits any vertex at distance $d + 1$.

[8] If the conditions set out in Section 3.4 are satisfied, then the Thesis can quantify the relationships by using a level-order traversal of the "Accounting Tree". This approach yields the same results as, and enjoys a lower algorithmic complexity than, Dijkstra's algorithm does.

[9] It is worth noting that the distance matrix is symmetric and all the values of its main diagonal are nil; consequently, distances can be calculated by using either the upper or lower triangular part of the matrix. This operation can improve the algorithmic complexity, because it reduces the number of calculations from $n^2$ to $\frac{n(n-1)}{2}$.

The graph-theoretical approach operates in three phases. First, it encodes a set of hierarchical data as a rooted and ordered tree. The root vertex represents the complete set of hierarchical data, and all other vertices are ordered in such a way that each vertex represents a subset of its "parent" vertex. The edges indicate the covering relations between the vertices. For example, consider a finite order set $P$; $x, y \in P$; $T = (V, E, w)$; and $v_x, v_y \in V$ correspond to $x$ and $y$ respectively. If $x$ is covered by $y$ (i.e. $x \prec y$), then $v_x$ is a "child" vertex of $v_y$. Each edge is assigned a weight, which can be any positive real number (i.e. $w : E \rightarrow \mathbb{R}^+$).

Second, the graph-theoretical approach traverses the tree in a level-order manner in order to calculate the distances between the root vertex and all other vertices. The distance between the root vertex $v_o$ and a vertex $v_i$ is the sum of the weighed-edges that exist in the unique path between $v_o$ and $v_i$. To calculate the distances for all pairs of vertices, the graph-theoretical approach designates each vertex as the root vertex and repeats the level-order traversal. This process yields distance matrix $D = (d_{ij})_{nn}$, where $d_{ij}$ denotes the distance between vertex $v_i$ and vertex $v_j$; $d_{ij}$ satisfies the conditions of a metric space, as follows (Jungnickel, 2002, p.65): (i) $d_{ij} > 0$ for all $i \neq j$ , (ii) $d_{ij} = 0$ if and only if $i = j$, (iii) $d_{ij} = d_{ji}$, and (iv) $d_{iz} \leq d_{ij} + d_{jz}$. Consequently, distance matrix $D$ constitutes the numerical representation of the hierarchical data, and its rows become input vectors to SOM.

## 3.5 Self-Organizing Map

### 3.5.1 Introduction

Self-organizing map performs two operations: (i) vector quantization, and (ii) vector projection. In the first operation, SOM regresses a set of codevectors into input data in a non-parametric and non-linear manner; this operation reduces input data to a smaller, albeit representative, set of codevectors. Second, SOM projects the codevectors onto a regular two-dimensional grid of neurons; this grid can be either hexagonal or rectangular depending on whether a neuron has either six or four neighbours. In either case, the neurons are spatially-ordered, and thus the grid can preserve the neighbourhood relations, or topology, between input data as faithfully as possible.

Given an input dataset, $X = (\vec{x}_{ij})_{nd}$, SOM initialises a set of codevectors, $M = (\vec{m}_{kj})_{Kd}$, where $\vec{x}_i$, $\vec{m}_k \in \mathbb{R}^{\mathsf{d}}$ are row vectors of $X$ and $M$ respectively, $n$ represents the number of input vectors, $d$ denotes the dimensionality of input space, and $K$ stands for the number of neurons on the SOM grid; $K$ can be estimated by using the formula: $K = 5\sqrt{n}$ (Vesanto et al., 2000, p.30). A neuron $k$ is described by the tuple $(\vec{m}_k, \vec{p}_k)$, where $\vec{m}_k \in \mathbb{R}^{\mathsf{d}}$ is a codevector, and $\vec{p}_k \in \mathbb{R}^2$ is a location vector on the SOM grid. The codevectors are used for vector quantization and their corresponding locations on the SOM grid for vector

projection.

### 3.5.2 Formation

The formation of SOM involves three iterative processes: (i) competition, (ii) co-operation, and (iii) adaptation (Haykin, 1999, p.447). First, in the competition process, each input vector, $\vec{x}_i \in \mathbb{R}^{\mathsf{d}}$, is compared with all codevectors, $\vec{m}_k \in \mathbb{R}^{\mathsf{d}}$, and the best match in terms of the smallest Euclidean distance, $\| \vec{x}_i - \vec{m}_k \|$, is mapped onto neuron $k$ that is termed the best-matching unit (i.e. BMU) and denoted by the subscript $c$ (Kohonen, 1997, p.86):

$$\| \vec{x}_i - \vec{m}_c \| = \min_k \{ \| \vec{x}_i - \vec{m}_k \| \} \quad , \tag{1}$$

or equivalently: $c = \underset{k}{\operatorname{argmin}} \{ \| \vec{x}_i - \vec{m}_k \| \}$ .

Second, in the co-operation process, the BMU locates the centre of a neighbourhood kernel, $h_{ck}(t)$, which is usually a Gaussian function defined as:

$$h_{ck}(t) = exp\left[ -\frac{\| \vec{p}_c - \vec{p}_k \|^2}{2\sigma^2(t)} \right] \quad , \tag{2}$$

where $\vec{p}_c$, $\vec{p}_k \in \mathbb{R}^2$ are the location vectors of BMU and neuron $k$ respectively, $t$ denotes discrete time, and $\sigma(t)$ is a monotonically decreasing function of time that defines the width of the kernel (Kohonen, 1997, p.87).

The neighbourhood kernel enables the SOM grid to preserve the topology or neighbourhood relations between input data by allowing the codevectors to be updated according to their respective proximity to the BMU. The closer to the BMU a codevector is, the greater the extent of its updating is, whereas codevectors lying outside the neighbourhood of BMU are not updated at all. As a result, the neurons on the SOM grid become spatially-ordered in the sense that neighbouring neurons have similar codevectors, and thus they represent similar areas in the input space.

Third, in the adaptive process, the sequence-training SOM updates recursively codevector $\vec{m}_k$ as follows:

$$\vec{m}_k(t+1) = \vec{m}_k(t) + a(t) h_{ck}(t) \left[ \vec{x}_i(t) - \vec{m}_k(t) \right] \quad , \tag{3}$$

where $0 < a(t) \leqslant 1$ is a learning rate at discrete time $t$; and, $a(t)$ is a non-increasing function of time, for example: $a(t) = a_0 \left( 1 - \frac{t}{T} \right)$ or $a(t) = a_0 \left( \frac{0.005}{a_0} \right)^{\frac{t}{T}}$, where $a_0$ is the initial learning rate, and $T$ is the training length.

Batch-SOM is a variant of the updating rule, described in Eq. 3. It estimates the BMU according to Eq. 1, but updates the codevectors only at the end of each epoch, which is a complete presentation of input data, rather than recursively. (Vesanto et al., 2000,

p.9):

$$\vec{m}_k(t+1) = \frac{\sum_{i=1}^{n} h_{ck}(t)\vec{x}_i}{\sum_{i=1}^{n} h_{ck}(t)} \quad . \tag{4}$$

Batch-SOM can be expressed in terms of Voronoi cells: $V_k = \{\vec{x}_i \mid \| \vec{x}_i - \vec{m}_k \| < \| \vec{x}_i - \vec{m}_j \| \; \forall k \neq j\}$, as follows (Vesanto et al., 2000, p.11):

$$\vec{m}_k(t+1) = \frac{\sum_{k=1}^{K} h_{ck}(t)\vec{s}_k(t)}{\sum_{k=1}^{K} h_{ck}(t)N_k} \quad , \tag{5}$$

where $K$ is the number of codevectors, and hence that of Voronoi cells, $N_k$ and $\vec{s}_k(t) = \sum_{\vec{x} \in V_k} \vec{x}$ denote the number and sum of input vectors that belong in Voronoi cell $V_k$, respectively.

SOM converges to a stationary state when the codevectors do not get updated any further. This entails that $E\{\vec{m}_k(t+1)\}$ must be equal to $E\{\vec{m}_k(t)\}$ for $t \to \infty$, at a stationary state (Kohonen, 1998). The foregoing leads to the convergence criterion: $E\{h_{ck}(\vec{x}_i - \lim_{t\to\infty} \vec{m}_k(t))\} = 0$, where $E\{.\}$ denotes the expectation function (Kohonen, 1997, p.113).

### 3.5.3 Clustering and performance metrics

The Unified-distance matrix (i.e. U-matrix) enables SOM to identify and visualise clusters that may be present in the input data (Ultsch and Siemon, 1990). The U-matrix calculates the average distance between a neuron's codevector and that of its immediate neighbouring neurons:

$$h(k) = \frac{1}{\mid N_k \mid} \sum_{i \in N_k} d(\vec{m}_k, \vec{m}_i) \quad . \tag{6}$$

Where $N_k$ stands for the neighbourhood of neuron $k$, and $\mid N_k \mid$ represents the number of neurons in that neighbourhood. The distances, $h(k)$, are depicted on a SOM grid as shades of grey; dark and light areas indicate large and small distances respectively, and denote cluster boundaries and clusters in this order.

A SOM grid can be evaluated in terms of its topology preservation and resolution. A SOM grid preserves the topology, or neighbouring relations, between input data if input vectors that are near to one another are also mapped onto adjacent neurons on the SOM grid. Topographic error (i.e. T.E.) quantifies the topology preservation of a SOM grid as follows: $T.E. = \frac{1}{n} \sum_{i=1}^{n} \varepsilon(\vec{x}_i)$; if the first and second BMUs of $\vec{x}_i$ are adjacent, then $\varepsilon(\vec{x}_i) = 0$, otherwise $\varepsilon(\vec{x}_i) = 1$ (Kiviluoto, 1996). Further, quantization error (i.e. Q.E.) measures resolution as the average distance between input vectors and their respective BMU: $Q.E. = \frac{1}{n} \sum_{i=1}^{n} \| \vec{x}_i - \vec{m}_c \|$; and, the lower a quantization error is, the higher the resolution of a SOM grid is.

The internal validity of clustering can be evaluated via the Davies-Bouldin index (Davies and Bouldin, 1979) as follows:

$$DBI = \frac{1}{C}\sum_{i=1}^{C}\max_{i\neq j}\left\{\frac{\Delta(C_i)+\Delta(C_j)}{\delta(C_i,C_j)}\right\} \quad . \tag{7}$$

Where $C$ is the number of clusters produced by SOM, $\delta(C_i, C_j)$ denotes inter-cluster distances, and $\Delta(C_i)$ and $\Delta(C_j)$ represent intra-cluster distances. A small value indicates highly-compact clusters whose centroids are well-separated.

If the input data are organized into categories or classes, then the external validity of clustering can be assessed by using entropy (Shannon, 1948):

$$H(Z) = -\sum_{j=1}^{C}\frac{m_j}{m}\sum_{i=1}^{K}\frac{m_{ij}}{m_j}log_2\frac{m_{ij}}{m_j} \quad . \tag{8}$$

Where $H(Z)$ denotes the entropy of clustering, C is the number of clusters produced by SOM, K is the number classes, $m_{ij}$ represents the number of observations in class $i$ that are clustered by SOM in cluster $j$, $m_j$ stands for the size of cluster $j$, and $m$ is the size of all clusters.

## 3.6 Non-parametric bootstrap

To conduct the experiments, the Thesis uses non-parametric bootstrap (Efron and Tibshirani, 1993) in order to mitigate the inferential restrictions arising from the small dataset. Because the sample size is small, the Thesis can not rely on the asymptotic assumption (i.e. $n \to \infty$) to estimate parameters of interest (e.g. mean, standard deviation) in the population nor the levels of accuracy (i.e. confidence intervals) associated with the estimates.

If the asymptotic assumption is taken to be valid, then confidence intervals can be estimated as follows: $\hat{\vartheta} \pm z^{\alpha}\hat{\sigma}$. Where $\hat{\vartheta}$ denotes the point estimate of a parameter of interest, $\hat{\sigma}$ is an estimate of $\hat{\vartheta}$'s standard deviation, and $z^{\alpha}$ is the $100.\alpha$th percentile of the standard normal distribution. The confidence intervals are, by definition, symmetrical about point estimate $\hat{\vartheta}$ and have a coverage probability of $1-2\alpha$. However, when the sample size is small, the foregoing formulation may yield biased estimators and incorrect confidence intervals.

To clarify the use of non-parametric bootstrap, the Thesis presents an example, drawn from the second paper, that involves estimating the topographic error. Let $\hat{\vartheta}$ stand for the point estimate of topographic error, $\vartheta$; and, let $X = (\vec{x}_{ij})_{nd}$ denote the dataset for fiscal year 2006, where $n = 25,440$ reflects the number of transactions. The transactions are assumed to be independent and identically distributed (i.e. iid) observations drawn from

an unknown probability distribution $F$, $x_1, x_2, \ldots, x_n \sim iid\, F$. The empirical distribution of the transactions, $\hat{F}$, is the discrete distribution that assigns probability $\frac{1}{n}$ to each transaction, where $n = 25,440$ as noted above.

In the first step, the Thesis draws with replacement one hundred random samples of size $n = 25,440$ from the empirical distribution of the transactions, $\hat{F}$. This operation yields one hundred bootstrap replications of the original dataset ($X = (\vec{x}_{ij})_{nd}$), that is $X_1^*, X_2^*, \ldots, X_{100}^*$. Second, for each bootstrap replication the Thesis trains a SOM and estimates the topographic error, $\hat{\vartheta}^*$. This procedure produces one hundred non-parametric bootstrap replications of the topographic error, $\hat{\vartheta}_1^*, \hat{\vartheta}_2^*, \ldots, \hat{\vartheta}_{100}^*$.

Third, in a procedure analogous to the first step, the Thesis draws with replacement 1,000 random samples of size $n = 100$ from the empirical distribution of $\hat{\vartheta}_1^*, \hat{\vartheta}_2^*, \ldots, \hat{\vartheta}_{100}^*$. This operation yields 1,000 bootstrap samples each containing one hundred estimates of topographic error. Finally, the Thesis applies $BC_a$ (DiCiccio and Efron, 1996) to the output of the preceding step in order to estimate the non-parametric two-sided 95% confidence intervals of the mean and standard deviation of topographic error.

# 4 DISCUSSION

## 4.1 Research objectives and motivation

The Thesis develops three quantitative models for auditing journal entries as a means of detecting and preventing material misstatements to financial statements. Material misstatements are those misstatements and omissions that are likely to influence the economic decisions informed users take on the basis of financial statements. Material misstatements that are caused intentionally constitute financial statement fraud.

The motivation behind the Thesis is fourfold. First, existing literature has hitherto provided tenuous empirical evidence on the link between auditing journal entries and financial statement fraud. Second, auditing standards require auditors to test the appropriateness of journal entries recorded in a general ledger. Third, journal entries are deemed to pose a high risk of financial statement fraud, as the case of WorldCom has demonstrated.

Finally, financial statement fraud results in considerable costs to a number of parties. For example, auditors may lose reputation capital and be exposed to litigation; capital markets may suffer from reduced liquidity; investors may experience negative stock returns; and regulators may be criticised for inept supervision of listed companies.

## 4.2 Proposed models

These considerations have led the Thesis to build and evaluate three quantitative models for auditing journal entries. The Thesis first employs SOM and extreme value theory to design the models as constructs; then, it codes the constructs to build functioning instantiations; and finally, it evaluates the instantiations by conducting a series of experiments on an accounting dataset, described in Section 3.2.

The first model can cluster journal entries in homogeneous and well-separated clusters that can be interpreted within an accounting perspective. A closer investigation reveals that the model can provide all the necessary and relevant financial information to function as financial statements. Results suggest the model corresponds accurately to the financial statements prepared by the case company; this correspondence lends much credence to the model. The model can provide a holistic picture of an accounting database at a given point in time as well as assist managers to supervise the processing of accounting transactions and assess the accuracy of financial statements.

The second model aims at detecting “suspicious” journal entries that may cause financial statements to be materially misstated. Empirical results indicate that the model can detect both “suspicious” and legitimate journal entries accurately. Additional analyses indicate that the performance of the model is robust to varying prior probabilities of

"suspicious" journal entries as well as to asymmetrical misclassification costs of Type I and Type II errors. The model may have a practical value in that it can assist auditors to detect "suspicious" journal entries. This potentiality may yield additional research; for example, a further study could code and implement the model as a Computer Assisted Audit Technique (i.e. CAAT).

The third model recognises that "suspicious" journal entries have a very low probability of occurring and a monetary amount large enough to materially misstate financial statements. The model employs the method of peaks-over-threshold, a subset of extreme value theory, to estimate an optimum threshold that can differentiate the distribution of legitimate from that of "suspicious" journal entries. The threshold is expressed as a function of three accounting variables: (i) Monetary amount; (ii) "Debit-Credit"; and (iii) "Account Class". As a result, the threshold can be interpreted from an accounting perspective; further, it can perform more efficiently than a uniform threshold can.

## 4.3 Critique

The Thesis is not impervious to criticism. Thus, criticism can be levelled at the lack of a more extensive dataset and at the construction of the models. The small size of the dataset may restrict the possible inferences that can be drawn as well as limit the generalisation of findings and conclusions to a broader set of companies. In order to mitigate the inferential restrictions arising from the small sample size, the Thesis has employed non-parametric bootstrap in conducting experiments, and in estimating parameters of interest and their associated confidence intervals. Even though the small size of the dataset has limited the scope of the Thesis, the dataset has yielded valuable experiments that have led to instructive findings.

Two factors may support the reasonable assumption that the dataset is representative of that of similar companies. First, regardless of differences in national GAAPs, the principles of double-entry bookkeeping are uniform; in this regard, not much has changed since Luca Pacioli codified the principles in 1494. Second, the ubiquitous use of relational databases to record and process journal entries makes it plausible that the structure of the dataset is similar to that of other companies.

The proposed models can be criticised on the ground that they do not account for time. This shortcoming manifests itself in two ways. First, the models are retrospective in that all relevant information (i.e. journal entries) ought to be available at the end of a fiscal period for auditors to test what has already occurred in that fiscal period. Although this approach typifies most audit engagements, it does little to facilitate continuous auditing, for which a real-time response is needed. Nonetheless, this limitation is not irredeemable; instead, future research could extend the Thesis by employing suitable algorithms. For example, Lee and Roberts (2008) suggest a real-time approach for novelty detection that

is based on Kalman filter and extreme value theory; and, Adams and MacKay (2007) propose an algorithm for Bayesian online change-point detection.

Second, the models are static (i.e. time-invariant); they assume that the statistical properties of journal entries do not vary over time. For example, the model, proposed in the fifth paper, estimates a constant threshold. However, it may not be realistic to use a constant threshold throughout a fiscal year, because there is much more scope for manipulating journal entries at year-end than there is during the rest of the year.

There are two reasons why the Thesis develops models that are static. First, it is more difficult to design a dynamic rather than a static model, other things being equal. Second, past attempts to incorporate time had failed because of lack of data. In particular, the distributions of monthly and quarterly journal entries do not vary enough over time for the models to be able to incorporate time.

## 4.4 Directions for future research

The Thesis has raised a number of questions that may support additional investigations. Expanding the proposed models could be a promising line of future research. Ongoing research aims at incorporating a temporal covariate in the model, proposed in the fifth paper, in order to capture information about cycles and trends that may exist in the distribution of journal entries. In doing so, the expanded model could estimate a threshold that takes into account the distribution of journal entries over time. An additional study could include in the foregoing model a covariate that captures auditors' assessment of audit risk; consequently, the higher the risk an audit engagement poses, the lower the estimated threshold would be. A further study could implement the proposed models as a Computer Assisted Audit Technique (i.e. CAAT). This potentiality can produce concrete practical applications in accounting, and thus forms the focal point of continuing research that is being carried out by the author.

Further studies could compare the results obtained by the models against decisions taken by auditors. It is possible that the models can outperform auditors in detecting "suspicious" journal entries. Additional research could investigate the models from the perspective of potential users, and evaluate them within a framework of end-user satisfaction.

# 5 SUMMARY AND CONTRIBUTION OF PAPERS

The Thesis includes five peer-reviewed publications, and investigates how the auditing of journal entries can detect and prevent financial statement fraud. The investigation bifurcates into two sub-currents of research that have so far received insufficient attention: first, clustering of accounting databases as a method for supplementing internal control procedures; and second, detecting "suspicious" journal entries that may materially misstate financial statements. The former sub-current is addressed by the first three papers, and the latter by the fourth and fifth papers. The two sub-currents of research coalesce to provide empirical evidence on, and to deepen the repository of knowledge about, the link between reviewing journal entries and detecting financial statement fraud.

The Thesis takes place during a period when companies, due to the ongoing financial crisis, could be tempted to misstate their financial statements in order to achieve, albeit artificially, expected revenue growth or present a favourable, and misleading, debt position. The Thesis could benefit a diverse group of parties, for example: assist auditors to fulfil their duties as laid down in statute and auditing standards; enhance the ability of regulatory authorities to supervise listed companies; and, improve investors' confidence to capital markets. Further, the Thesis may have broader implications for the efficient functioning of capital markets.

## 5.1 Clustering hierarchical data using Self-Organizing Map: A graph-theoretical approach

Clustering accounting databases is problematic, as it contains hierarchical variables (e.g. "Account Class"). These variables lack quantitative information for SOM to be able to calculate a similarity measure (e.g. Euclidean, Mahalanobis) that satisfies the conditions of a metric space. Current practices involve encoding hierarchical variables into binary terms, {1,0}, that are then treated as numerical values as well as using a modified formulation of SOM, Generalizing SOM (Hsu, 2006).

The paper argues that these techniques suffer from a number of limitations. Consequently, it proposes a graph-theoretical approach, described in Sections 3.3 and 3.4, that can pre-process hierarchical data into a numerical representation, and thus render them amenable for SOM-based processing.

Experimental results suggest that the graph-theoretical approach can yield SOM grids that exhibit significantly higher cluster validity and resolution than binary encoding does. Further analyses indicate that the graph-theoretical approach enjoys a lower algorithmic complexity than Generalizing SOM does. Finally, the graph-theoretical approach is decoupled from SOM, and hence it can be used by other algorithms to

pre-process hierarchical data into a numerical representation.

However, the graph-theoretical approach assumes that the semantical relationships between hierarchical data are static. Further, it increases the dimensionality of the input space in direct proportion to the number of values of a hierarchical variable. These limitations can be the focal points of future research.

## 5.2 A semi-supervised tool for clustering accounting databases with applications to internal controls

This paper designs and validates a semi-supervised tool for clustering accounting databases as a means of supplementing internal control procedures. To design the tool, the paper combines SOM, graph-theoretical approach, and IFRS XBRL Taxonomy. The paper assesses the internal consistency of the tool via a series of experiments on an accounting dataset, described in Section 3.2. Further, it benchmarks the tool against the financial statements of the case company in order to evaluate how accurately the tool corresponds to the company's economic reality.

Empirical results suggest that the tool can preserve the topology of the input dataset faithfully as well as exhibit high resolution. Both the mean and standard deviation of the topographic error have small values and narrow 95% confidence intervals: $CI_{0.95} = 0.269 - 0.336$, and $CI_{0.95} = 0.150 - 0.192$, respectively; the corresponding values for the quantization error are nil to the fifth decimal place. The tool can group accounting transactions into homogeneous and well-separated clusters; the mean and standard deviation of Davies-Bouldin index have small values and narrow 95% confidence intervals: $CI_{0.95} = 1.037 - 1.107$, and $CI_{0.95} = 0.153 - 0.207$, in this order.

More importantly, the clusters can be interpreted from an accounting perspective. The tool can provide all the necessary and relevant financial information to function as financial statements. Indeed, the results from the benchmarking exercise indicate that the tool corresponds accurately to the financial statements prepared by the case company; this correspondence lends much credence to the proposed tool. A closer investigation reveals that the tool can compress a number of similar transactions to a single neuron as well as preserve the semantical relationships between one account class and another, and, by extension, between accounting transactions.

The tool can supplement the practice of accounting by providing a holistic picture of an accounting database at a given point in time, assisting a manager to supervise the processing of accounting transactions, and enabling a manager to assess the accuracy of financial statements. In addition to the practical applications, the paper can pave the way for further research, for example: validating the results of the paper against those of

auditors, evaluating how useful the tool is within an end-user satisfaction framework, and investigating the potential of the tool to produce financial statements. Finally, ongoing research aims at coding the tool as a Matlab-toolbox to be released in the public domain.

## 5.3 Using Self-Organizing Map for data mining: A synthesis with accounting applications

This paper concludes the trilogy on the clustering of accounting databases; it includes unpublished material as well as insights that have been obtained in preparing the previous two papers. It provides a research synthesis of the literature pertinent to SOM as well as presents a case study that illustrates how SOM can identify and visualise clusters that may exist in accounting databases.

The research synthesis centres on three organising principles: first, the formation of SOM; second, those modifications that extend SOM to a broader class of data-mining tasks (e.g. clustering of symbol strings); and third, the main financial applications of SOM, such as: financial benchmarking, bankruptcy prediction, and analysis of self-declared styles of hedge funds. The case study uses the accounting dataset for fiscal year 2007, described in Section 3.2, and the research methodology explained in the first and second papers.

The paper extends current literature in that it synthesises a fragmented and expansive literature relevant to SOM, focuses on the main financial applications of SOM, and demonstrates the application of SOM to accounting by means of a case study.

## 5.4 Auditing journal entries using Self-Organizing Map

This paper aims at detecting "suspicious" journal entries that may cause financial statements to be materially misstated. It employs SOM for its ability to perform vector quantization in order to derive a reference model that can describe the behaviour of legitimate journal entries. Subsequently, it calculates the quantization error, a distance metric, between the reference model and novel journal entries containing seeded errors. Finally, it considers the journal entries whose quantization error exceeds an optimum threshold as "suspicious".

The paper conducts a series of experiments on an accounting dataset, described in Section 3.2. It examines twelve scenarios by combining four cost ratios of Type I to Type II errors with three prior probabilities of "suspicious" journal entries occurring in the population. For each scenario, the paper addresses four categories of potential errors that may affect a journal entry, and performs one hundred experiments by using non-parametric bootstrap.

Empirical results suggest that the proposed model can detect both "suspicious" and

legitimate journal entries accurately. For example, in the scenario whereby "suspicious" journal entries occur with a 5% probability and a Type II error is ten times more costly than Type I error, the model has a true negative rate of 91% ($C.I_{0.95} = 0.861 - 0.938$) and a true positive rate of 88.3% ($C.I_{0.95} = 0.854 - 0.907$); the respective 95% confidence intervals are shown in parentheses. Additional analyses indicate that the performance of the model is robust to varying prior probabilities of "suspicious" journal entries as well as to asymmetrical misclassification costs of Type I and Type II errors.

The model may have a practical value in that it can assist auditors to detect "suspicious" journal entries. This potentiality may yield additional research. For example, a study could compare the performance of the model against that of unaided auditors, who may detect "suspicious" journal entries only on the basis of their judgement and past experience. A further study could code and implement the model as a Computer Assisted Audit Technique (i.e. CAAT).

However, the model has three shortcomings that may reduce its ability to detect "suspicious" journal entries. First, it is predicated upon the assumption that a dataset containing only legitimate journal entries exists in advance. Second, the model does not examine the scenario in which the prior probability of "suspicious" journal entries is less than 1%. Finally, the model estimates a threshold in terms of the Euclidean distance; consequently, the threshold can not be interpreted from an accounting perspective.

Motivated by these shortcomings, the next paper adopts a different paradigm to detect "suspicious" journal entries. First, it presumes that a dataset contains both legitimate and "suspicious" journal entries; second, it considers "suspicious" journal entries to be rare events, i.e. events occurring with a very low probability (e.g. less than 1%); and finally, it estimates a threshold that can be interpreted within an accounting context.

## 5.5 Auditing journal entries using extreme value theory

This paper proposes a bipartite model for detecting "suspicious" journal entries in order to assist auditors to assess the risk of material misstatements to financial statements. The paper conjectures that "suspicious" journal entries are rare, i.e. have a very low probability of occurring, and also have a monetary amount large enough to likely cause financial statements to be materially misstated.

The first component of the model employs the peaks-over-threshold method, a subset of extreme value theory, in order to estimate an optimum threshold that can differentiate the distribution of legitimate from that of "suspicious" journal entries. The second component models the number of monthly "suspicious" journal entries in terms of a univariate Poisson distribution, and draws inferences by using Bayesian analysis.

The results suggest that the proposed model can estimate a threshold that, at least in

principle, can be interpreted within the context of extreme value theory; by contrast, a heuristic threshold lacks any probabilistic interpretation. In this respect, the model can mitigate the subjectivity and bias that may occur when auditors select a threshold only on the basis of their past experience and knowledge. Further, the model can estimate a threshold as a function of three accounting variables: (i) "Monetary Amount", (ii) "Debit-Credit", and (iii) "Account Category". As a result, the threshold enjoys two main advantages: first, it can be interpreted from an accounting perspective; and second, it can perform more efficiently than selecting the largest x% (e.g. 5%) of the monetary amounts, as the latter alternative is a function of only one variable.

Ongoing research aims at incorporating a temporal covariate in order to capture information about cycles and trends that may be present in the distribution of journal entries. An extension to the model could include a variable for auditors' assessment of audit risk; the higher the audit risk an audit engagement poses, the lower the threshold for the engagement would be.

Table 4: Descriptive statistics for the Balance Sheet items

| | Debit | | | | | | Credit | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Percentile | | | | | | Percentile | |
| | N | Mode | Mean | Median | 0.75 | 0.25 | N | Mode | Mean | Median | 0.75 | 0.25 |
| NON-CURRENT ASSETS | | | | | | | | | | | | |
| Property, Plants and Equipment_Costs | 80 | 109 | 1,887 | 793 | 2,208 | 367 | 8 | -3,073 | -1,275 | -1,076 | -552 | -1,880 |
| Property, Plants and Equipment_Depr. | 6 | 40 | 548 | 294 | 609 | 50 | 136 | -2,677 | -1,427 | -624 | -151 | -2,010 |
| Investments in Subsidiaries | 4 | 100 | 18,438 | 11,825 | 31,875 | 5,000 | | | | | | |
| CASH AND CASH EQUIVALENTS | | | | | | | | | | | | |
| Cash in Bank | 4,324 | 1,274 | 45,202 | 10,085 | 21,872 | 2,029 | 2,776 | -2 | -72,105 | -871 | -116 | -14,551 |
| Cash at Hand | 182 | 0 | 549 | 313 | 700 | 107 | 490 | -3 | -181 | -49 | -16 | -144 |
| TRADE AND OTHER RECEIVABLES | | | | | | | | | | | | |
| Sales Taxes Receivable | 1,310 | 4 | 164 | 11 | 49 | 4 | 102 | -225 | -2,321 | -225 | -102 | -1,372 |
| Trade Debtors | 4,182 | 1,252 | 12,021 | 3,195 | 13,327 | 714 | 3,933 | -1,274 | -13,286 | -8,535 | -1,789 | -16,780 |
| Other Trade Debtors | 1,550 | 536 | 15,065 | 1,350 | 4,703 | 297 | 1,365 | -198 | -19,070 | -1,427 | -315 | -4,712 |
| Accounts Receivable | 366 | 0 | 349 | 125 | 400 | 34 | 734 | -10 | -174 | -36 | -13 | -134 |
| Short-term Loans Receivable | 130 | 8 | 185,759 | 165 | 3,557 | 66 | 56 | -1,889 | -181,259 | -1,385 | -165 | -3,257 |
| Insurance Receivable | 5,033 | 1,052 | 8,748 | 2,609 | 8,000 | 836 | 5,820 | -1,052 | -7,638 | -1,925 | -666 | -7,500 |
| Other Debtors and Receivables | 162 | 207 | 7,536 | 1,131 | 2,785 | 213 | 242 | -103 | -5,292 | -491 | -118 | -2,500 |
| CURRENT LIABILITIES | | | | | | | | | | | | |
| Accounts Payable | 3,967 | 203 | 2,298 | 275 | 685 | 125 | 8,534 | -150 | -967 | -219 | -101 | -620 |
| Trade Creditors | 1,104 | 0 | 40,243 | 731 | 7,880 | 91 | 2,734 | -135 | -15,541 | -1,765 | -354 | -7,200 |
| EQUITY | | | | | | | | | | | | |
| Issued Capital | | | | | | | | | | | | |
| Retained Earnings | 10 | 234,679 | 854,951 | 370,574 | 801,059 | 234,679 | 12 | -469,358 | -1,292,217 | -352,019 | -14,585 | -1,361,695 |
| Profit for the year | 10 | 167 | 1,377,255 | 352,019 | 1,012,936 | 5,565 | 8 | -234,679 | -889,793 | -234,679 | -234,679 | -1,545,891 |

Table 5: Descriptive statistics for the Profit and Loss items

| | Debit | | | | | | Credit | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Percentile | | | | | | Percentile | |
| | N | Mode | Mean | Median | 0.75 | 0.25 | N | Mode | Mean | Median | 0.75 | 0.25 |
| INCOME | | | | | | | | | | | 0 | |
| Other Income Received | 5 | 224 | 502,951 | 6,419 | 1,089,236 | 952 | 44 | -331 | -70,058 | -331 | -147 | -949 |
| Fees Received | 2 | 1,500 | 15,000 | 15,000 | 28,500 | 1,500 | 66 | -28,500 | -15,000 | -15,000 | -1,500 | -28,500 |
| Insurance Commission Received | 1,069 | 667 | 6,136 | 984 | 2,516 | 360 | 4,100 | -132 | -2,520 | -582 | -143 | -1,628 |
| Bank Interest Received | 27 | 1 | 30,852 | 502 | 3,754 | 9 | 638 | -157 | -2,572 | -800 | -265 | -2,224 |
| Other Interest Received | | | | | | | 16 | -153,653 | -175,009 | -152,818 | -137,658 | -210,311 |
| Exchange Difference Gain | 803 | 0 | 983 | 5 | 86 | 0 | 671 | -18 | -3,072 | -121 | -18 | -522 |
| EXPENSES | | | | | | | | | | | | |
| Administration Expenses | 2,027 | 28 | 215 | 54 | 201 | 20 | 452 | -11 | -334 | -93 | -48 | -474 |
| Office Expenses | 689 | 68 | 801 | 169 | 767 | 68 | 187 | -17 | -1,159 | -372 | -89 | -927 |
| Salaries and Wages | 275 | 197 | 12,329 | 4,570 | 8,777 | 1,335 | 118 | -197 | -9,293 | -3,238 | -338 | -7,092 |
| Fees and Commissions | 239 | 59 | 1,218 | 248 | 1,053 | 34 | 43 | -1,146 | -1,653 | -1,000 | -255 | -1,589 |
| Professional Expenses | 52 | 4,500 | 6,246 | 1,760 | 4,500 | 447 | 16 | -53,081 | -10,458 | -1,708 | -647 | -18,466 |
| Finance Expenses | 808 | 2 | 124 | 23 | 51 | 13 | 56 | -56 | -106 | -30 | -12 | -81 |
| Depreciation Expenses | 131 | 2,677 | 1,065 | 624 | 1,746 | 136 | 4 | -609 | -197 | -69 | -45 | -348 |

## Appendix A: Extract from the text file containing journal entries

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 - 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AMI ADM | Jan-07 | 220719 | RON PAPE | 05/01/2007 | C | GBP | 0.4875 | 44 | 86 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | RON PAPE | 05/01/2007 | D | GBP | 0.4875 | 44 | 86 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 220719 | RON PAPE | 05/01/2007 | C | GBP | 0.4875 | 13 | 25 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | RON PAPE | 05/01/2007 | D | GBP | 0.4875 | 13 | 25 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 220719 | RON PAPE | 07/01/2007 | C | GBP | 0.4875 | 506 | 985 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | RON PAPE | 07/01/2007 | D | GBP | 0.4875 | 506 | 985 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 220719 | RON PAPE | 07/01/2007 | C | GBP | 0.4875 | 5 | 10 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | RON PAPE | 07/01/2007 | D | GBP | 0.4875 | 5 | 10 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 220719 | RON PAPE | 07/01/2007 | C | GBP | 0.4875 | 17 | 34 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | RON PAPE | 07/01/2007 | D | GBP | 0.4875 | 17 | 34 | | | Z | 0 | 0 |
| ********** | | | | | | | | | | | | | | |
| AMI ADM | Jan-07 | 220715 | AM/10/06/H | 01/01/2007 | C | USD | 1 | 136 | 136 | | | | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | USD | 1 | 8 | 8 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | USD | 1 | 11 | 11 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | EUR | 0.6896 | 25 | 32 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | EUR | 0.6896 | 25 | 32 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | EUR | 0.6896 | 35 | 44 | | | Z | 0 | 0 |
| AMI ADM | Jan-07 | 850013 | AM/10/06/H | 01/01/2007 | D | USD | 1 | 9 | 9 | | | Z | 0 | 0 |
| ********** | | | | | | | | | | | | | | |
| AMI INPAY | Oct-07 | 212001 | HSH USD | 10/10/2007 | D | USD | 1 | 7,289 | 7,289 | | | | 0 | 0 |
| AMI INPAY | Oct-07 | 220004 | HSH USD | 10/10/2007 | C | USD | 1 | 7,289 | 7,289 | | | | 0 | 0 |
| AMI INPAY | Oct-07 | 345504 | AM007.1204 | 10/10/2007 | D | USD | 1 | 400 | 400 | | | | 0 | 0 |
| AMI INPAY | Oct-07 | 345506 | AM007.1204 | 10/10/2007 | C | USD | 1 | 259 | 259 | | | | 0 | 0 |
| AMI INPAY | Oct-07 | 345506 | AM007.1204 | 10/10/2007 | C | USD | 1 | 140 | 140 | | | | 0 | 0 |
| ********** | | | | | | | | | | | | | | |
| AMI INPAY | Dec-07 | 212001 | HSH USD | 21/12/2007 | D | USD | 1 | 13,303 | 13,303 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 220169 | AM007.1560 | 21/12/2007 | C | USD | 1 | 4,079 | 4,079 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 345504 | AM007.1560 | 21/12/2007 | D | USD | 1 | 651 | 651 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 345506 | AM007.1560 | 21/12/2007 | C | USD | 1 | 651 | 651 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 220169 | AM007.1589 | 21/12/2007 | C | USD | 1 | 9,224 | 9,224 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 345504 | AM007.1589 | 21/12/2007 | D | USD | 1 | 1,473 | 1,473 | | | | 0 | 0 |
| AMI INPAY | Dec-07 | 345506 | AM007.1589 | 21/12/2007 | C | USD | 1 | 1,473 | 1,473 | | | | 0 | 0 |

# Appendix B: Definition of variables

| Column | Variable | Description | |
|---|---|---|---|
| 1 | Journal Code | | |
| | | ADM | ADMINISTRATION EXPENSES |
| | | CASH | PETTY CASH |
| | | EXP | EXPENDITURE |
| | | GLB | GLB INCOME |
| | | INC | INCOME |
| | | INPAY | INWARD PAYMENTS |
| | | JV | JOURNAL VOUCHER |
| | | OUTP | OUTWARD PAYMENT |
| | | RED | REALISED EXCHANGE DIFFERENCE |
| | | REXP | RETURN EXPENSES - C/N |
| | | RINC | RETURN INCOME - C/N |
| 2 | Posting Period | The month to which a transaction is related | |
| 3 | Account Code | | |
| 4 | Reference | Alphanumerical code identifying a transaction | |
| 5 | Transaction Date | The date on which a transaction is entered | |
| 6 | Debit/Credit | D = Debit; C = Credit | |
| 7 | Currency Code | | |
| | | CYP | Cyprus pounds |
| | | EUR | Euro |
| | | GBR | British pounds |
| | | USD | USA dollars |
| 8 | Currency Rate | | |
| | | US$/CYP | |
| | | US$/EUR | |
| | | US$/GBR | |
| 9 | Transaction Amount | | |
| 10 | Reporting Amount (US$) | | |
| 11 | Details | Text describing a transaction, left blank intentionally | |
| 12 - 16 | For internal purposes only | | |
| 17, 18 | Vat Code | Vat Rate | |
| | A | 5% | |
| | B | 10% | |
| | D | 18% | |
| | E | 0% | |
| | F | 8% | |
| | G | 7% | |
| | H | 8% | |
| | I | 0% | |
| | N | 0% | |
| | O | 0% | |
| | R | 15% | |
| | S | 15% | |
| | Z | 0% | |
| | Blanks | 0% | |
| 19 | Vat Amount | Var. 10 multiplied by the corresponding VAT rate. | |

## Appendix C: Chart of Accounts

NON-CURRENT ASSETS
- 1000 Property, Plants and Equipment_Costs
- 1100 Property, Plants and Equipment_Depr.
- 1200 Investments in Subsidiaries

CASH AND CASH EQUIVALENTS
- 2110 Cash in Bank
- 2140 Cash at Hand

TRADE AND OTHER RECEIVABLES
- 2170 Sales Taxes Receivable
- 2200 Trade Debtors
- 2204 Other Trade Debtors
- 2207 Accounts Receivable
- 2300 Short-term Loans Receivable
- 2400 Insurance Receivable
- 2600 Other Debtors and Receivables

CURRENT LIABILITIES
- 3455 Accounts Payable
- 3480 Trade Creditors

EQUITY
- 4100 Issued Capital
- 4500 Retained Earnings
- 9990 Profit for the year

INCOME
- 6000 Other Income Received
- 6100 Fees Received
- 6260 Insurance Commission Received
- 6300 Bank Interest Received
- 6305 Other Interest Received
- 6440 Exchange Difference Gain

EXPENSES
- 8500 Administration Expenses
- 8501 Office Expenses
- 8550 Salaries and Wages
- 8600 Fees and Commissions
- 8610 Professional Expenses
- 8700 Finance Expenses
- 8800 Depreciation Expenses

## REFERENCES


Adams, R. P. and MacKay, D. J. (2007), 'Bayesian Online Changepoint Detection', *Technical Report, University of Cambridge, Cambridge U.K Available at: http://hips. seas. harvard. edu/files/adams-changepoint-tr-2007. pdf* .

AICPA (2002), 'Statement On Auditing Standards 99 (SAS 99): Consideration of Fraud in a Financial Statement Audit', *American Institute of Certified Public Accountants* .

AICPA (2012), 'Substantive Differences Between the International Standards on Auditing and Generally Accepted Auditing Standards', *American Institute of Certified Public Accountants* . Available at: www.aicpa.org/interestareas/frc/auditattest/downloadabledocuments/clarity/substantive_differences_isa_gass.pdf.

AICPA (2013), 'Summary of Differences Between Clarified SASs and Existing SASs', *American Institute of Certified Public Accountants* . Available at: www.aicpa.org/interestareas/frc/auditattest/downloadabledocuments/clarity/clarity_sas_summary_of_differences.pdf.

Aier, S. and Fischer, C. (2011), 'Criteria of progress for information systems design theories', *Information Systems and e-Business Management* **9**(1), 133–172.

Apostolou, B. A., Hassell, J. M., Webber, S. A. and Sumners, G. E. (2001), 'The relative importance of management fraud risk factors', *Behavioral Research in Accounting* **13**, 1–24.

Armstrong, C. S., Jagolinzer, A. D. and Larcker, D. F. (2010), 'Chief executive officer equity incentives and accounting irregularities.', *Journal of Accounting Research* **48**(2), 225 – 271.

Asare, S. K. and Wright, A. M. (2004), 'The effectiveness of alternative risk assessment and program planning tools in a fraud setting.', *Contemporary Accounting Research* **21**(2), 325–352.

Bell, T. B. and Carcello, J. V. (2000), 'A decision aid for assessing the likelihood of fraudulent financial reporting', *Auditing: A Journal of Practice and Theory* **19**(1), 169–184.

Beresford, D. R., Katzenbach, N. d. and Rogers, C. B. J. (2003), *Report of Investigation by the Special Investigative Committee of the Board of Directors of WorldCom, INC.*, Available at: http://news.findlaw.com/wsj/docs/worldcom/bdspcomm60903rpt.pdf.

Burns, N. and Kedia, S. (2006), 'The impact of performance-based compensation on misreporting', *Journal of Financial Economics* **79**(1), 35–67.

Burrell, G. and Morgan, G. (1979), *Sociological Paradigms and Organizational Analysis*, Heinemann Educational Books Ltd.

Canadian Institute of Chartered Accountants (2004), *IT Control Assessments in the context of CEO/CFO Certification*, Toronto, Canada.

Coakley, J. R. (1995), 'Using pattern analysis methods to supplement attention directing analytical procedures', *Expert Systems with Applications* **9**(4), 513–528.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, Springer-Verlag, London, UK.

COSO (2010), *Fraudulent Financial Reporting: 1998-2007. An Analysis of U.S. Public Companies*, The Committee of Sponsoring Organizations of the Treadway Commission (COSO). Available at: http://www.coso.org/documents/COSOFRAUDSTUDY2010_001.PDF.

David, J. S., Gerard, G. J. and McCarthy, W. E. (2002), Design science: An REA perspective on the future of AIS, *in* V. Arnold and S. G. Sutton, eds, 'Researching Accounting as an Information Systems Discipline', American Accounting Association, Information Systems Section, Sarosota, Florida.

Davies, D. and Bouldin, D. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2), 224–227.

Debreceny, R. S. and Gray, G. L. (2010), 'Data mining journal entries for fraud detection: An exploratory study', *International Journal of Accounting Information Systems* **11**(3), 157–181.

Dechow, P. M., Sloan, R. G. and Sweeney, A. P. (1996), 'Causes and consequences of earnings manipulations: An analysis of firms subject to enforcement actions by the SEC', *Contemporary Accounting Research* **13**(1), 1–36.

DiCiccio, T. J. and Efron, B. (1996), 'Bootstrap confidence intervals', *Statistical Science* **11**(3), 189–228.

Dijkstra, E. W. (1959), 'A note on two problems in connexion with graphs', *Numerische Mathematik* **1**, 269–271.

Efendi, J., Srivastava, A. and Swanson, E. P. (2007), 'Why do corporate managers misstate financial statements? the role of option compensation and other factors', *Journal of Financial Economics* **85**(3), 667–708.

Efron, B. and Tibshirani, R. (1993), *An introduction to the bootstrap*, Chapman & Hall, New York, USA.

Erickson, M., Hanlon, M. and Maydew, E. L. (2006), 'Is there a link between executive equity incentives and accounting fraud?', *Journal of Accounting Research* **44**(1), 113–143.

Farber, D. B. (2005), 'Restoring trust after fraud: Does corporate governance matter?', *The Accounting Review* **80**(2), 539–561.

Goel, S., Gangolly, J., Faerman, S. R. and Uzuner, O. (2010), 'Can linguistic predictors detect fraudulent financial filings?', *Journal of Emerging Technologies in Accounting* **7**, 25–46.

Grabski, S. (2010), 'Discussion of "Data mining journal entries for fraud detection: An exploratory study"', *International Journal of Accounting Information Systems* **11**(3), 182–185.

Green, B. P. and Choi, J. H. (1997), 'Assessing the risk of management fraud through neural network technology.', *Auditing: A Journal of Practice and Theory* **16**, 14 – 28.

Gregor, S. and Hevner, A. R. (2011), 'Introduction to the special issue on design science', *Information Systems and eBusiness Management* **9**(1), 1–9.

Haykin, S. (1999), *Neural Networks. A Comprehensive Foundation*, second edn, Prentice Hall International, Upper Saddle River, New Jersey, USA.

Hevner, A. R. (2007), 'The three cycle view of design science research', *Scandinavian Journal of Information Systems* **19**(2), 87–92.

Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004), 'Design science in information systems research', *MIS Quarterly* **28**(1), 75–105.

Hogan, C. E., Rezaee, Z., Riley, R. A. and Velury, U. K. (2008), 'Financial statement fraud: Insights from the academic literature', *Auditing: A Journal of Practice and Theory* **27**(2), 231–252.

Hsu, C. (2006), 'Generalizing self-organizing map for categorical data', *IEEE Transactions on Neural Networks* **17**(2), 294–304.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K. and Felix, W. F. (2011), 'Identification of fraudulent financial statements using linguistic credibility analysis', *Decision Support Systems* **50**(3), 585–594.

IASCF (2009), *IFRS Taxonomy Guide 2009 (XBRL)*, International Accounting Standards Committee Foundation (IASCF), London, United Kingdom.

Iivari, J. (1991), 'A paradigmatic analysis of contemporary schools of IS development', *European Journal of Information Systems* **1**(4), 249–272.

Iivari, J. (2010), Twelve theses on design science research in information systems, *in* A. R. Hevner and S. Chatterjee, eds, 'Design Research in Information Systems: Theory and Practice', Vol. 22 of *Integrated Series in Information Systems*, Springer, chapter 5.

International Federation of Accountants (IFAC) (2009), 'International Standard on Auditing 240 (ISA 240): The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements'. Available at: http://www.ifac.org/sites/default/files/downloads/a012-2010-iaasb-handbook-isa-240.pdf.

Järvinen, P. (2001), *On Research Methods*, english edn, Opinpaja Oy, Tampere, Finland.

Jungnickel, D. (2002), *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Volume 5, English edn, Springer, Berlin, Germany.

Kaminski, K. A., Wetzel, S. T. and Guan, L. (2004), 'Can financial ratios detect fraudulent financial reporting?', *Managerial Auditing Journal* **19**(1), 15–28.

Kasanen, E., Lukka, K. and Siitonen, A. (1993), 'The constructive approach in management accounting research', *Journal of Management Accounting Research* **5**, 243–264.

Kiviluoto, K. (1996), Topology preservation in Self-Organizing maps, *in* 'Proceeding of International Conference on Neural Networks (ICNN'96)', pp. 294–299.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**(1), 59–69.

Kohonen, T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences, Volume 30, second edn, Springer-Verlag, Heidelberg, Germany.

Kohonen, T. (1998), 'The self-organizing map', *Neurocomputing* **21**(1-3), 1–6.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000), 'Self organization of a massive document collection', *IEEE Transactions on Neural Networks* **11**(3), 574–585.

Kohonen, T. and Somervuo, P. (1998), 'Self-organizing maps of symbol strings', *Neurocomputing* **21**(1-3), 19–30.

Kohonen, T. and Somervuo, P. (2002), 'How to make large self-organizing maps for nonvectorial data', *Neural Networks* **15**(8-9), 945–952.

Koskivaara, E. (2000), 'Artificial neural network models for predicting patterns in auditing monthly balances', *The Journal of the Operational Research Society* **51**(9), 1060–1069.

Koskivaara, E. and Back, B. (2007), 'Artificial neural network assistant (ANNA) for continuous auditing and monitoring of financial data.', *Journal of Emerging Technologies in Accounting* **4**, 29–45.

Lagus, K., Kaski, S. and Kohonen, T. (2004), 'Mining massive document collections by the WEBSOM method', *Information Sciences* **163**(1-3), 135–156.

Lee, H. and Roberts, S. J. (2008), On-line Novelty Detection Using the Kalman Filter and Extreme Value Theory, *in* '19th International Conference on Pattern Recognition, 2008.(ICPR 2008)', pp. 1–4.

Loebbecke, J. K., Eining, M. M. and Willingham, J. J. (1989), 'Auditors' experience with material irregularities: Frequency, nature, and detectability.', *Auditing: A Journal of Practice and Theory* **9**(1), 1–28.

Loebbecke, James K. Willingham, J. J. (1988), Review of SEC Accounting and Auditing Enforcement Releases. Unpublished Working Paper.

March, S. T. and Smith, G. F. (1995), 'Design and natural science research on information technology', *Decision Support Systems* **15**(4), 251–266.

March, S. T. and Storey, V. C. (2008), 'Design science in the information systems discipline: an introduction to the special issue on design science research.', *MIS Quarterly* **32**(4), 725–730.

MATLAB Release R2012a (2012), The MathWorks Inc., Natick, Massachusetts, USA.

Palmrose, Z. (1987), 'Litigation and independent auditors - the role of business failures and management fraud', *Auditing: A Journal of Practice and Theory* **6**(2), 90–103.

Palmrose, Z. (1991), 'An analysis of auditor litigation disclosures.', *Auditing: A Journal of Practice and Theory* **10**(Supplement), 54–71.

Pincus, K. V. (1989), 'The efficacy of a red flags questionnaire for assessing the possibility of fraud', *Accounting, Organizations and Society* **14**(1-2), 153–163.

Public Company Accounting Oversight Board (PCAOB) (2004), 'Standing Advisory Group Meeting: Financial Fraud', *Available at: http://pcaobus.org/News/Events/Documents/09082004_SAGMeeting/Fraud.pdf* .

Public Company Accounting Oversight Board (PCAOB) (2007), 'Auditing Standard No. 5: An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements'. Available at: pcaobus.org/Rules/Rulemaking/Docket 021/2007-06-12_Release_No_2007-005A.pdf.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**, 379–423 & 623–656.

Simon, H. A. (1996), *The Sciences of the Artificial*, 3rd edn, MIT Press, Cambridge (MA), USA.

Summers, S. L. and Sweeney, J. T. (1998), 'Fraudulently misstated financial statements and insider trading: An empirical analysis', *The Accounting Review* **73**(1), 131–146.

Ultsch, A. and Siemon, H. (1990), Kohonen's self organizing feature maps for exploratory data analysis, *in* 'Proceedings International Neural Network Conference', Kluwer Academic Press, Dordrecht, Netherlands, pp. 305–308.

U.S. Congress (2002), 'Sarbanes-Oxley Act of 2002, H.R.3763'.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), SOM Toolbox for Matlab 5, Technical Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland. Available at: http://www.cis.hut.fi/somtoolbox/.

# PART II: PUBLICATIONS

**Clustering hierarchical data using Self-Organizing Map:**
**A graph-theoretical approach**

**Argyris Argyrou**

**Abstract**
The application of Self-Organizing Map (SOM) to hierarchical data remains an open issue, because such data lack inherent quantitative information. Past studies have suggested binary encoding and Generalizing SOM as techniques that transform hierarchical data into numerical attributes. Based on graph theory, this paper puts forward a novel approach that processes hierarchical data into a numerical representation for SOM-based clustering. The paper validates the proposed graph-theoretical approach via complexity theory and experiments on real-life data. The results suggest that the graph-theoretical approach has lower algorithmic complexity than Generalizing SOM, and can yield SOM having significantly higher cluster validity than binary encoding does. Thus, the graph-theoretical approach can form a data-preprocessing step that extends SOM to the domain of hierarchical data.

**Keywords:** Clustering, hierarchical data, Self-Organizing Map, graph theory

# 1 Introduction

The Self-Organizing Map (SOM) (Kohonen, 1997) represents a type of artificial neural network that is based on unsupervised learning. It has been applied extensively in the areas of dimensionality reduction, data visualization, and clustering (Vesanto, 2002). The original formulation of SOM uses the Euclidean distance as a similarity metric (Kohonen et al., 1996, p.4), and hence its domain of application is restricted to metric spaces (Kohonen and Somervuo, 1998). SOM has been extended to non-metric spaces by using generalized means and medians as the distance measures and the batch variant of SOM (Kohonen and Somervuo, 1998); for example, speech recognition (Kohonen and Somervuo, 1997), and clustering of protein sequences (Kohonen and Somervuo, 2002). An online algorithm for SOM of symbol strings was provided by Somervuo (2004). However, neither a metric distance nor a string metric (e.g. Levenshtein distance) can yield meaningful results in the domain of hierarchical data, and thus the application of SOM in this domain remains an open issue. For example, consider clustering the data: {cat, rat, mouse}. A string metric would find that {cat} and {rat} are more closely related to each other than {rat} and {mouse} are, while a metric distance would produce meaningless results.

To address this issue, prior studies have suggested two main techniques that transform hierarchical attributes into numerical attributes. First, the most prevalent technique encodes a categorical attribute in binary terms {1,0}, where 1 and 0 denote the presence and absence of an attribute respectively. The binary encoding is then treated as a numerical attribute in the range {1,0}. Second, Hsu (2006) introduced Generalizing SOM (GSOM), whereby a domain expert describes a set of categorical data by means of

a concept hierarchy, and then extends it to a distance hierarchy in order to represent and calculate distances between the categorical data. However, both techniques suffer from theoretical and practical limitations.

Motivated by this open issue, the paper puts forward a graph-theoretical approach that processes hierarchical data into a numerical representation, and thus renders them amenable for clustering using SOM. To elaborate, based on graph theory, the paper encodes a set of hierarchical data in the form of a rooted and ordered tree. The root vertex represents the complete set of the hierarchical data, and each vertex represents a sub-set of its "parent" vertex. An edge between a pair of vertices is assigned a weight, which can be any positive real number, representing the distance between the two vertices. Thus, the distance between a pair of vertices, $v_i$ and $v_j$, is the sum of the weighed-edges that exist in the path from $v_i$ to $v_j$. The paper uses a level-order traversal algorithm to calculate the distances between each vertex and all other vertices. This process yields a symmetric distance matrix $D = \left(d_{ij}\right)_{nn}$, where $n$ is the number of vertices, and $d_{ij}$ the distance between $v_i$ and $v_j$.

In the present case, the paper encodes the animals that are contained in the zoo-dataset (Asuncion and Newman, 2007) in the form of a rooted and ordered tree, and calculates the distances between all pairs of animals by using a level-order traversal of the tree, as shown in Fig. 1. The symmetric distance matrix $D = \left(d_{ij}\right)_{nn}$ thus derived forms the numerical representation of the zoo-dataset, where $n = 98$ reflecting the number of animals, and $d_{ij}$ denotes the distance between a pair of animals. The distance metric $d_{ij}$ satisfies the conditions of a metric space, as follows (Jungnickel, English edition 2002, p.65): (i) $d_{ij} \geq 0$, (ii) $d_{ij} = 0$ if and only if $i = j$, (iii) $d_{ij} = d_{ji}$, and (iv) $d_{iz} \leq d_{ij} + d_{jz}$. Each row in $D$ represents an animal, and becomes an input vector – $\vec{x}_j \in \mathbb{R}^{98}$, $j = 1, 2, \ldots 98$ – to SOM[1].

The paper trains two SOMs, batch and sequence, for each of the two representations of the zoo-dataset, original binary encoding and paper's graph-theoretical approach. For each of the four combinations, the paper selects one hundred samples by using bootstrap; and for each of the 400 bootstrapped samples, it trains a SOM with a Gaussian neighborhood and an 8 x 5 hexagonal lattice. The paper evaluates the quality of each SOM in terms of: (i) the entropy of clustering, (ii) quantization error, (iii) topographic error, and (iv) the Davies-Bouldin index. Based on these quality measures, the paper uses the Wilcoxon rank-sum test at the one-tailed 5% significance level to assess whether the graph-theoretical approach can yield significantly better SOM than binary encoding does. Further, the paper compares the algorithmic complexity of the graph-theoretical

[1]The distance matrix $D$ is symmetric, and hence the number of observations (i.e. animals) is equal to the number of dimensions (i.e. 98), and selecting either rows or columns as input vectors to SOM would yield the same result.

approach with that of GSOM.

The results suggest that the graph-theoretical approach enjoys a lower algorithmic complexity than Generalizing SOM does, and can yield SOM having significantly higher cluster validity than binary encoding does.

The paper's novelty and contribution lie in the formulation of the graph-theoretical approach, and its application as a data-preprocessing step that can extend SOM to the domain of hierarchical data.

The paper proceeds as follows. Section 2 describes briefly the SOM algorithm, binary encoding, and Generalizing SOM. Section 3 formulates the graph-theoretical approach. Section 4 outlines the design of experiments, and section 5 presents and discusses the results. Section 6 presents the conclusions.

# 2 Background and Related Work

## 2.1 The SOM Algorithm

In the context of this study, the SOM algorithm performs a non-linear projection of the probability density function of the 98-dimensional input space to an 8 x 5 2-dimensional hexagonal lattice. A neuron $i$, $i = 1, 2, \ldots 40$, is represented by XY coordinates on the lattice, and by a codevector, $\vec{m}_i \in \mathbb{R}^{98}$, in the input space. The formation of a SOM involves three processes (Haykin, 1999, p.447): (i) competition, (ii) co-operation, and (iii) adaptation. First, each input vector, $\vec{x} \in \mathbb{R}^{98}$, is compared with all codevectors, $\vec{m}_i \in \mathbb{R}^{98}$, and the best match in terms of the smallest Euclidean distance, $\| \vec{x} - \vec{m}_i \|$, is mapped onto neuron $i$, which is termed the best-matching unit (BMU):

$$BMU = \underset{i}{\operatorname{argmin}} \{ \| \vec{x} - \vec{m}_i \| \} \quad . \tag{1}$$

In the co-operation process, the BMU locates the center of the neighborhood kernel $h_{ci}$:

$$h_{ci} = a(t) \cdot \exp\left[ -\frac{\| r_c - r_i \|^2}{2\sigma^2(t)} \right] \quad . \tag{2}$$

where $r_c$, $r_i \in \mathbb{R}^2$ are the radius of BMU and node $i$ respectively, $t$ denotes discrete time, $a(t)$ is a learning rate, and $\sigma(t)$ defines the width of the kernel; $a(t)$ and $\sigma(t)$ are monotonically decreasing functions of time (Kohonen et al., 1996, p.5).

In the adaptive process, the sequence-training SOM updates the BMU codevector as follows:

$$\vec{m}_i(t+1) = \vec{m}_i(t) + h_{ci}(t) \left[ \vec{x}(t) - \vec{m}_i(t) \right] \quad . \tag{3}$$

The batch-training SOM estimates the BMU according to (1), but updates the BMU

codevector as (Vesanto et al., 2000, p.9):

$$\vec{m}_i(t+1) = \frac{\sum_{j=1}^{n} h_{ci}(t)\vec{x}_j}{\sum_{j=1}^{n} h_{ci}(t)} \quad . \tag{4}$$

To carry out the experiments, the paper uses both sequence-training (3) and batch-training (4) SOM.

## 2.2 Binary Encoding and Generalizing SOM

Binary encoding converts a categorical variable into a numerical representation consisting of values in the range $\{1, 0\}$, where 1 and 0 denote the presence and absence of an attribute respectively. The binary encoding of each categorical datum is then treated as a numerical attribute for SOM-based clustering.

To overcome the limitations associated with binary encoding, Hsu (2006) introduced Generalizing SOM (GSOM). Briefly, a domain expert extends a concept hierarchy, which describes a data domain, to a distance hierarchy by associating a weight for each link on the former. The weight represents the distance between the root and a node of a distance hierarchy. For example, a point $X$ in distance hierarchy $dh(X)$ is described by $X = (N_X, d_X)$, where $N_X$ is a leaf node and $d_X$ is the distance from the root to point $X$. The distance between points $X$ and $Y$ is defined as follows:

$$\mid X - Y \mid = d_X + d_Y - 2d_{LCP(X,Y)} \quad . \tag{5}$$

where $d_{LCP(X,Y)}$ is the distance between the root and the least common point of $X$ and $Y$.

# 3 The Graph-Theoretical Approach

## 3.1 Preliminaries

Table 1: Notations and definitions

| | |
|---|---|
| $G = (V, E, w)$ | A graph |
| $V = \{v_1, v_2, \ldots v_n\}$ | Set of vertices |
| $E = \{e_1, e_2, \ldots e_m\}$ | Set of edges |
| $w : E \to \mathbb{R}^+$ | Function assigning a positive real number to an edge |
| $\mid V \mid$ | Degree of graph, cardinality of $V$ |
| $\mid E \mid$ | Order of graph, cardinality of $E$ |
| $e = \{v_i, v_j\}$ | Edge connecting vertices $v_i$ and $v_j$ |
| $d_{ij} = w(e)$ | Distance between $v_i$ and $v_j$ |
| $D = \left(d_{ij}\right)_{nn}$ | Distance matrix |

A comprehensive review of graph theory lies beyond the scope of this paper; a textbook account on this subject can be found in Jungnickel (English edition 2002). For the

purposes of this paper, it suffices to define a tree as a special type of graph, $G = (V, E, w)$, that satisfies at least two of the following three necessary and sufficient properties: (i) $G$ is acyclic, (ii) $G$ is connected, and (iii) $| E | = | V | - 1$; any two of these properties imply the third (Jungnickel, English edition 2002, p.8). Let $T = (V, E, w)$ be a tree that is: (i) rooted, with $v_0$ the root vertex, and (ii) ordered, which means that there is a linear ordering of its vertices such that for each edge $e = \{v_i, v_j\}$ then $v_i < v_j$. It can be easily deduced that in tree $T$: (i) all vertices excluding $v_0$ have at most one "parent" vertex, (ii) at least one vertex has no "child" vertices, and (iii) there is a unique path between any two vertices. A tree can be traversed in a level-order way; such a traversal starts from the root vertex, $v_0$, and proceeds from left-to-right to visit each vertex at distance $d$ from $v_0$ before it visits any vertex at distance $d + 1$, as shown in Fig. 1.

## 3.2 Description

The graph-theoretical approach is motivated by the observation that hierarchical variables have a set of states that can be ranked in a meaningful order. For example, consider the variable "size" having five states: {very big, big, medium, small, very small}. It is obvious that {very big} matches {big} more closely than it matches {very small}. However, this piece of information is lost if binary encoding is used, because such an encoding produces a dichotomous output: a state either matches another state or does not.

The graph-theoretical approach operates in three phases. First, it encodes a set of hierarchical data in the form of a rooted and ordered tree. The root vertex represents the complete set of hierarchical data, and all other vertices are ordered in such a way that each vertex represents a sub-set of its "parent" vertex. The edges indicate the covering relation between the vertices. For example, consider a finite order set $P$; $x, y \in P$; $T = (V, E, w)$; and $v_x, v_y \in V$ correspond to $x$ and $y$ respectively. If $x$ is covered by $y$ (i.e. $x \prec y$), then $v_x$ is a "child" vertex of $v_y$. Each edge is assigned a weight, which can be any positive real number (i.e. $w : E \rightarrow \mathbb{R}^+$).

Second, the graph-theoretical approach traverses the tree in a level-order manner in order to calculate the distances between the root vertex and all other vertices. The distance between the root vertex $v_o$ and a vertex $v_i$ is the sum of the weighted-edges that exist in the unique path between $v_o$ and $v_i$. This calculation has an algorithmic complexity of $O(| V |)$. To calculate the distances for all pairs of vertices, the graph-theoretical approach designates each vertex as the root vertex and repeats the level-order traversal. Thus, the all-pairs distances can be obtained in $O\left(| V |^2\right)$. This process yields a symmetric distance matrix $D = \left(d_{ij}\right)_{nn}$, where $d_{ij}$ denotes the distance between vertex $v_i$ and vertex $v_j$, $d_{ij} > 0$ for all $i \neq j$ , $d_{ij} = 0$ if and only if $i = j$, $d_{ij} = d_{ji}$, and $d_{iz} \leq d_{ij} + d_{jz}$.

Finally, the distance matrix $D$ constitutes the numerical representation of the set of hierarchical data and each of its rows becomes an input vector to SOM.

# 4 Data and Experiments

The design of experiments consists of six steps. First, the zoo-dataset (Asuncion and Newman, 2007) contains 101 animals that are described by one numerical attribute and 15 binary attributes, and classified into seven groups. The paper eliminates the instances "girl" and "vampire" for obvious but unrelated reasons, and one instance of "frog", because it appears twice.

Second, to apply the graph-theoretical approach to the zoo-dataset, the paper uses none of the original attributes. Instead, it uses a "natural" taxonomy that classifies animals based on their "phylum", "class", and "family". This taxonomy can be expressed as a tree (Fig. 1), where the root vertex stands for the complete set of animals. For the experiments, the weight for each edge is set to 1 (i.e. $w : E \rightarrow 1$ ), though it can be any positive real number and different for each edge. The paper calculates the distances of all pairs of vertices by using a level-order traversal of the tree, and thus derives a distance matrix that makes up the numerical representation of the zoo-dataset.

Third, for each representation of the zoo-dataset, original binary encoding and the paper's graph-theoretical approach, the paper uses bootstrap to draw one hundred random samples with replacement. Fourth, for each bootstrapped sample, the paper trains two SOMs, batch and sequence, with a Guassian neighborhood and an 8 x 5 hexagonal lattice. Fifth, the paper evaluates each SOM in terms of four quality measures: (i) the entropy of clustering, (ii) quantization error, (iii) topographic error, and (iv) the Davies-Bouldin index. Sixth, based on the quality measures, the paper uses the Wilcoxon rank-sum test at the one-tailed 5% significance level to assess whether the graph-theoretical approach can yield significantly better SOMs than binary encoding does.

Further, the paper compares the algorithmic complexity of the proposed graph-theoretical approach with that of Generalizing SOM (Hsu, 2006). An experimental comparison was not possible, because GSOM was not available.[2]

## 4.1 Quality Measures

The quantization error, QE, and topographic error, TE, have been extensively reviewed in the literature pertinent to SOM. Thus, this section concentrates on two cluster validity indices: (i) the Davies-Bouldin index, and (ii) the entropy of clustering.

The Davies-Bouldin index (Davies and Bouldin, 1979), DBI, is defined as:

$$DBI = \frac{1}{C} \sum_{i=1}^{C} \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \quad . \tag{6}$$

[2]Personal correspondence with the author of Generalizing SOM.

where $C$ is the number of clusters produced by SOM, $\delta(C_i, C_j)$, and $\Delta(C_i)$ and $\Delta(C_j)$ the intercluster and intracluster distances respectively.

Following Shannon (1948), the entropy of clustering Z, H(Z), can be defined as:

$$H(Z) = -\sum_{j=1}^{C} \frac{m_j}{m} \sum_{i=1}^{K} \frac{m_{ij}}{m_j} log_2 \frac{m_{ij}}{m_j} \quad . \tag{7}$$

where C is the number of clusters produced by SOM, $K = 7$, the number of groups of animals in the zoo-dataset, $m_{ij}$ is the number of animals in group $i$ that are clustered by SOM in cluster $j$, $m_j$ is the size of cluster $j$, and $m$ is the size of all clusters.

# 5 Results and Discussion

The results (Table 2) suggest that the graph-theoretical approach yields SOMs having statistically significant lower entropy of clustering, quantization error, and Davies-Bouldin index than binary encoding does. In contrast, the difference in topographic error is not significant. Further, the results are invariant to the two SOM-training algorithms, batch and sequence.

Table 2: Wilcoxon rank-sum test

| SOM-Training | H(Z) | QE | TE | DBI |
|---|---|---|---|---|
| Batch | A<B | A<B | N.S | A<B |
| Sequence | A<B | A<B | N.S | A<B |

Referring to Table 2, A and B stand for the graph-theoretical approach and binary encoding respectively, $A < B$ denotes that the difference between the two approaches is statistically significant at the one-tailed 5% significance level, whereas N.S implies that a significant difference does not exist.

To compare the algorithmic complexity of the graph-theoretical approach with that of GSOM (Hsu, 2006), the paper assumes that GSOM is applied to the zoo-dataset, and that GSOM uses this paper's tree (Fig. 1) as its distance hierarchy. As discussed in Sect. 2.2, GSOM entails the following three tasks: (i) calculate distances from the root to all nodes, a level-order traversal of the tree has $O(\mid V \mid)$ complexity; (ii) find the all-pairs least common point (LCP), the current fastest algorithm has $O\left(\mid V \mid^{2.575}\right)$ complexity (Czumaj et al., 2007); and (iii) calculate distances from the root to all LCPs, this takes $O(l)$, where $l$ is the number of LCPs.

Therefore, the algorithmic complexity of GSOM is $O\left(\mid V \mid^{2.575}\right)$, and hence higher than the quadratic complexity, $O\left(\mid V \mid^{2}\right)$, of the graph-theoretical approach.

## 5.1 Critique

The proposed graph-theoretical approach is not impervious to criticism. Like binary encoding, it increases the dimensionality of the input space in direct proportion to the number of states a hierarchical variable has. In turn, the dimensionality of the search space increases exponentially with the dimensionality of the input space, a phenomenon aptly named "the curse of dimensionality" (Maimon and Rokash, 2005, p.160). Further, it assumes that the hierarchical data are static, and hence a deterministic approach is sufficient. To deal with this limitation, future research may explore a probabilistic variant of the graph-theoretical approach.

# 6 Conclusions

The paper's novelty and contribution lie in the development and application of a data-preprocessing step that is based on graph theory and can extend SOM to the domain of hierarchical data. The results suggest that the proposed graph-theoretical approach has lower algorithmic complexity than Generalizing SOM, and can yield SOM having significantly higher cluster validity than binary encoding does. Further, the graph-theoretical approach is not confined only to SOM, but instead it can be used by any algorithm (e.g. k-means) to process hierarchical data into a numerical representation. Future research may consider a probabilistic variant of the graph-theoretical approach as well as its application in the area of hierarchical clustering. Notwithstanding its limitations, the paper presents the first attempt that uses graph theory to process hierarchical data into a numerical representation for SOM-based clustering.

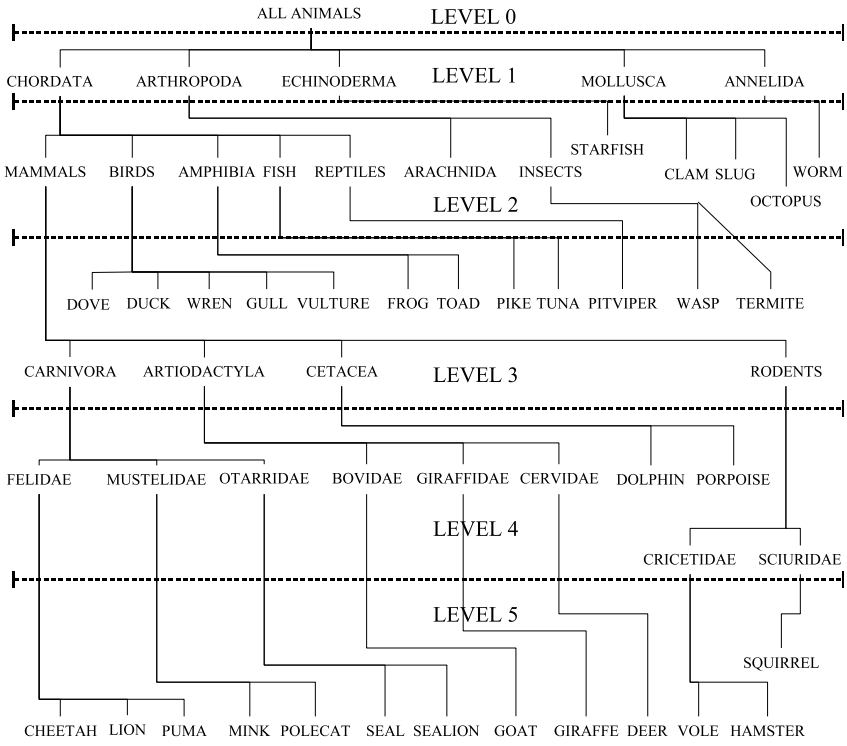



Figure 1: Extract from the graph-theoretical representation of the zoo-dataset.

**Acknowledgements and Dedication.**

I am much indebted to Mr. Tom Lindström, Prof. Anders Tallberg, Dr. Andriy Andreev, and Dr. Sofronis K. Clerides for their insightful comments and suggestions. This paper is dedicated to my first mentor, the late Mr. Marios Christou.

## REFERENCES


Asuncion, A. and Newman, D. (2007), *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml/datasets/Zoo.

Czumaj, A., Kowaluk, M. and Lingas, A. (2007), 'Faster algorithms for finding lowest common ancestors in directed acyclic graphs', *Theoretical Computer Science* **380**, 37–46.

Davies, D. and Bouldin, D. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2), 224–227.

Haykin, S. (1999), *Neural Networks. A Comprehensive Foundation*, second edn, Prentice Hall International, Upper Saddle River, New Jersey:USA.

Hsu, C.-C. (2006), 'Generalizing self-organizing map for categorical data', *IEEE Transactions on Neural Networks* **17**(2), 294–304.

Jungnickel, D. (English edition 2002), *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Volume 5, Springer, Berlin:Germany.

Kohonen, T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences, Volume 30, second edn, Springer-Verlag, Heidelberg:Germany.

Kohonen, T., Hynninen, J., Kangas, J. and Laaksonen, J. (1996), Som-pak: The self-organizing map program package., Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kohonen, T. and Somervuo, P. (1997), 'Self-organizing maps of symbol strings with application to speech recognition', *Proceedings of the First International Workshop on Self-Organizing Maps (WSOM'97)* pp. 2–7.

Kohonen, T. and Somervuo, P. (1998), 'Self-organizing maps of symbol strings', *Neurocomputing* **21**(1-3), 19–30.

Kohonen, T. and Somervuo, P. (2002), 'How to make large self-organizing maps for nonvectorial data', *Neural Networks* **15**(8-9), 945–952.

Maimon, O. and Rokash, L., eds (2005), *The Data Mining and Knowledge Discovery Handbook*, first edn, Springer-Verlag New York, Inc., Secaucus, NJ:USA.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**, 379–423 & 623–656.

Somervuo, P. J. (2004), 'Online algorithm for the self-organizing map of symbol strings', *Neural Networks* **17**(8-9), 1231–1239.

Vesanto, J. (2002), Data Exploration Process Based on the Self-Organizing Map, Doctoral dissertation, Helsinki University of Technology, Espoo, Finland.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), SOM Toolbox for Matlab 5, Technical Report A57, Helsinki University of Technology, Espoo, Finland. Available at: http://www.cis.hut.fi/somtoolbox/.

**A semi-supervised tool for clustering accounting databases with applications to internal controls**

**Argyris Argyrou and Andriy Andreev**

**Abstract**
A considerable body of literature attests to the significance of internal controls; however, little is known on how the clustering of accounting databases can function as an internal control procedure. To explore this issue further, this paper puts forward a semi-supervised tool that is based on Self-Organizing Map and the IASB XBRL Taxonomy. The paper validates the proposed tool via a series of experiments on an accounting database provided by a shipping company. Empirical results suggest the tool can cluster accounting databases in homogeneous and well-separated clusters that can be interpreted within an accounting context. Further investigations reveal that the tool can compress a large number of similar transactions, and also provide information comparable to that of financial statements. The findings demonstrate that the tool can be applied to verify the processing of accounting transactions as well as to assess the accuracy of financial statements, and thus supplement internal controls.

**Keywords:** Self-Organizing Map, clustering accounting databases, internal controls

# 1 Introduction

A number of statutory and professional pronouncements require a public company's management to implement and maintain appropriate internal controls in order to ensure the integrity and reliability of accounting transactions. In particular, the Sarbanes-Oxley Act of 2002 Section 404 (U.S. Congress, 2002) mandates a public company's CFO and CEO not only to implement and maintain internal controls over financial reporting but also to assess and certify the effectiveness of such controls on an annual basis. The Act also requires an external auditor to attest management's certification of internal controls. Further, the Statement on Auditing Standards 94 (AICPA, 2001) suggests that an auditor should not rely exclusively on substantive testing when evidence of a company's recording and processing of accounting transactions exists only in an electronic form. Instead, an auditor should assess controls over a company's information technology (i.e. I.T) environment in which the recording and processing of accounting transactions occur. These I.T controls form an integral part of a company's internal control system as they underpin the completeness, accuracy, and timeliness of a company's financial reporting (Canadian Institute of Chartered Accountants, 2004). To discharge their duties, managers and auditors can seek advice and guidance from two widely adopted frameworks, COSO (COSO, 1992) and COBIT (IT Governance Institute, 2000). However comprehensive the two frameworks are, they provide little guidance on the design and application of specific tools.

Motivated by this unexplored issue, the paper designs and validates a semi-supervised tool

for clustering accounting databases in order to supplement internal control procedures. To elaborate, the proposed tool provides a holistic snapshot of an accounting database, it supports a manager in checking whether an accounting database can process the right transactions accurately, and in assessing the accuracy of financial statements. The tool and its practical applications in the domain of accounting constitute the paper's novelty and contribution.

To design the tool, the paper uses Self-Organizing Map (i.e. SOM) (Kohonen, 1997) for its ability to cluster and visualize data as well as to map the probability density function of a multi-dimensional input space to a two-dimensional output space (i.e. SOM grid) while preserving the original topology. SOM has been applied successfully in a multitude of research domains; for example, financial benchmarking (Eklund et al., 2003), predicting corporate bankruptcy (Serrano-Cinca, 1996; Lee et al., 2005), market segmentation (Kiang et al., 2006), evaluating the creditworthiness of loan applicants (Huysmans et al., 2006), selecting an MBA program (Kiang and Fisher, 2008), improving the accuracy of a naive Bayes classifier to classify text documents (Isa et al., 2009), and identifying peer schools for AACSB accreditation (Kiang et al., 2009).

In addition to SOM, the paper incorporates into the proposed tool accounting knowledge in the form of the International Accounting Standard Board (i.e. IASB) XBRL Taxonomy (IASCF, 2009) that describes the disclosure and presentation of financial statements pursuant to the International Financial Reporting Standards (i.e. IFRSs) (IASB, 2009). Briefly, the IASB XBRL Taxonomy is a hierarchical structure describing accounting concepts (e.g. Assets, Liabilities, Equity, etc.) and the semantic relationships between them as $child \prec parent$ links. The paper opted for the IFRSs rather than some national Generally Accepted Accounting Principles (i.e. GAAPs), because IFRSs are adopted by most of the leading economies. For instance, starting in 2005, all European Union companies listed on a regulated market must adopt IFRSs for preparing their consolidated financial statements (European Parliament, 2002); the U.S. Securities and Exchange Commission (2008) proposed a roadmap that, once completed, could lead to the mandatory adoption of IFRSs by U.S publicly listed companies for fiscal periods ending on or after the $15^{th}$ of December 2014; and companies that are publicly listed in Canada must adopt IFRSs for a fiscal period ending in 2011 (AcSB, 2009).

To validate the proposed tool, the paper applies it to accounting transactions that were provided by a shipping company as a text-dump of its database covering fiscal year 2006. In the first step, the paper uses the graph-theoretical approach (Argyrou, 2009) to pre-process accounting-hierarchical data into a numerical representation for SOM-based clustering. Second, the paper uses bootstrap to select random samples with replacement.

For each bootstrapped sample, it uses SOM-Toolbox[1] for Matlab (Vesanto et al., 2000) to train a SOM in batch mode with hexagonal topology and Gaussian neighbourhood. Further, the paper evaluates each SOM by calculating the following three quality measures: (i) topographic error, T.E., (ii) quantization error, Q.E., and (iii) Davies-Bouldin Index, DBI. In addition, the paper uses the bootstrap bias-corrected with acceleration method to estimate the two-sided 95% confidence interval of the mean and standard deviation for the aforesaid quality measures. Finally, to ensure that the proposed tool satisfies its intended use, the paper benchmarks the tool's output against the case company's financial statements.

The rest of the paper proceeds as follows. The next section describes the research design and methodology, Section 3 presents and discusses the results, and Section 4 concludes as well as suggests future research perspectives.

# 2 Research Design and Methodology

## 2.1 Data description

The data were provided by a shipping company in the form of a text-dump of its accounting database. The data describe the economic activities of the case company for fiscal year 2006, and consist of 25,440 accounting transactions each of which is described by seven variables, as shown in Table 1. The accounting dataset is denoted as matrix $A = \left(a_{ij}\right)_{nm}$, where $n = 25,440$ and $m = 7$ reflecting the number of transactions and variables respectively. All variables but "Account Class" are specific to the case company, and hence they are unlikely to be used by any other company. By contrast, the variable "Account Class" conveys information on how to aggregate accounting transactions for a company to prepare financial statements. Because the presentation and disclosure of financial statements are dictated by IFRSs, "Account Class" is likely to be used by other companies that report under IFRSs, and thus it provides a link between an accounting database and IFRSs. On these grounds, the proposed tool employs "Account Class" as a lens through which a user can probe into an accounting database. To avoid duplication, we describe "Account Class" in the same table, Table 3, in which we describe the results from the benchmarking exercise.

However, "Account Class" can not be processed directly by SOM, because it lacks intrinsic quantitative information for SOM to calculate the Euclidean distance. Although SOM can use non-metric similarity measures (e.g. Levenshtein distance) in lieu of the Euclidean distance (Kohonen and Somervuo, 1998, 2002); such measures can not capture the semantic relationships between hierarchical-categorical data, and hence they

[1] SOM-Toolbox for Matlab and its documentation are available in the public domain at http://www.cis.hut.fi/somtoolbox/

are not suitable for the present study. Instead, the paper observes that "Account Class" can be represented as a directed acyclic graph (i.e. DAG) that adheres to an a priori known hierarchy, the IASB XBRL Taxonomy.

Table 1: Description of variables

| Name | Type |
|---|---|
| Account Number | Alphanumeric |
| Account Description | Text |
| Posting Date | Date |
| Debit-Credit Indicator | Binary |
| USD Amount | Numerical |
| Transaction Details | Text |
| Account Class | Hierarchical-Categorical |

## 2.2 Data pre-processing

Motivated by this observation, the paper uses the graph-theoretical approach[2] to pre-process "Account Class" into a numerical representation that takes the form of a distance matrix. In particular, the graph-theoretical approach operates in two steps. First, the paper encodes "Account Class" as a DAG, shown in Fig. 1, that in turn represents a graphical instantiation of the IASB XBRL Taxonomy. The root vertex represents the complete set of "Account Class", and all other vertices are ordered in such a way that each vertex represents a sub-set of its parent vertex. As a result, the $child \prec parent$ relationships specified in the Taxonomy are preserved. For example, "*Bank Account*" $\prec$ "*Cash in Bank*" $\prec$ "*Cash and Cash Equivalents*" $\prec$ "*Current Assets*" constitute $child \prec parent$ relationships. In the second step, the graph-theoretical approach quantifies the aforementioned relationships by using Dijkstra's algorithm (Dijkstra, 1959) to calculate the all-pairs distances between vertices. This calculation yields a symmetric distance matrix, $D = \left(d_{ij}\right)_{NN}$, where $N = 29$ denoting the number of account classes, and $d_{ij}$ the distance between a pair of account classes. The distance is the sum of the weighed-edges that exist between the path of two account classes; for the experiments, the weight for each edge is set to 1. For example, the distances between "Accounts Payable" and: {itself, "Office Expenses", "Trade Creditors"} are 0, 8, and 2 respectively. The distance metric, $d_{ij}$, satisfies the conditions of a metric space, as follows (Jungnickel, 2002, p.65): (i) $d_{ij} > 0$ for all $i \neq j$, (ii) $d_{ij} = 0$ if and only if $i = j$, (iii) $d_{ij} = d_{ji}$ for all $i$ and $j$, and (iv) $d_{iz} \leq d_{ij} + d_{jz}$ for all $i$, $j$, and $z$.

The distance matrix, $D$, thus derived forms the numerical representation of "Account Class". In essence, the distance matrix defines the semantic relationships between "Account Class", and, by extension, between accounting transactions. In doing so, it

[2]We implement the graph-theoretical approach by writing the required code in Mathematica (Wolfram Research Inc., 2007).

represents the accounting knowledge the paper incorporates into the tool.

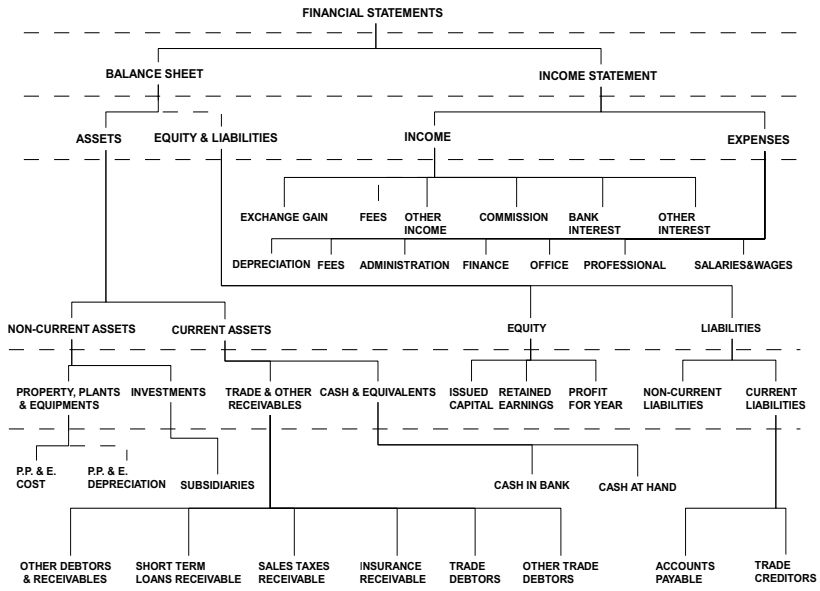



Figure 1: The Accounting Tree, a graphical instantiation of the IASB XBRL Taxonomy.

## 2.3 Self-Organizing Map

To set up the input dataset for SOM, the paper concatenates accounting dataset, $A$, with distance matrix, $D$, by using "Account Class" as the key variable. This operation produces matrix $X = (x_{ik})_{n(N+m)}$ that makes up the input dataset to SOM, where $n = 25{,}440$, $N = 29$, and $m = 7$. In the context of this study, SOM maps the probability density function of the input dataset to a two-dimensional hexagonal grid of neurons, as shown in Fig. 2. The number of neurons is set to 820 corresponding approximately to $5 \times \sqrt{n}$, where $n = 25{,}440$ reflecting the number of accounting transactions. The size of the SOM grid is set to 41 rows and 20 columns so that the ratio of the grid's sides (i.e. 2:1) matches the ratio of the two biggest eigenvalues of the covariance matrix of the input data (Vesanto et al., 2000, p.30). Each neuron has two representations, as follows: (i) a coordinate on the SOM grid, and (ii) a codevector, $\vec{m}_j \in \mathbb{R}^{36}$, in the input space, where $j = 1, 2, \ldots 820$.

The formation of SOM involves three iterative processes (Haykin, 1999, p.447). First, in the competition process, each input vector, $\vec{x}_i \in \mathbb{R}^{36}$, is compared with all codevectors, $\vec{m}_j \in \mathbb{R}^{36}$, and the best match in terms of the smallest Euclidean distance, $\| \vec{x}_i - \vec{m}_j \|$, is mapped onto neuron $j$ termed the best-matching unit (i.e. BMU) and denoted by the subscript $c$: $\| \vec{x}_i - \vec{m}_c \| = \min_j \left\{ \| \vec{x}_i - \vec{m}_j \| \right\}$ (Kohonen, 1997, p.86). Second, in the co-

operation process, the BMU locates the center of the Gaussian neighborhood: $h_{cj} = exp\left[-\frac{\|r_c - r_j\|^2}{2\sigma^2(t)}\right]$, where $r_c$, $r_j \in \mathbb{R}^2$ are the radius of BMU and neuron $j$ respectively, $t$ denotes discrete time, and $\sigma(t)$ defines the width of the kernel (Kohonen, 1997, p.87). Third, in the adaptive process, the batch-training SOM updates the BMU codevector as follows (Vesanto et al., 2000, p.9): $\vec{m}_j(t+1) = \frac{\sum_{i=1}^{n} h_{cj}(t)\vec{x}_i}{\sum_{i=1}^{n} h_{cj}(t)}$.

## 2.4 Quality Measures

The quality of the SOM grid, Fig. 2, can be evaluated with respect to its topology preservation, and resolution. The SOM grid preserves faithfully the original topology if vectors that are close in the input space are mapped onto adjacent neurons. Topographic error, T.E., quantifies topology preservation as the proportion of all input vectors whose first and second BMUs are not mapped onto adjacent neurons (Kiviluoto, 1996). Analytically, $T.E. = \frac{1}{n}\sum_{i=1}^{n} \varepsilon(\vec{x}_i)$; if the first and second BMUs of $\vec{x}_i$ are adjacent, then $\varepsilon(\vec{x}_i) = 0$, otherwise $\varepsilon(\vec{x}_i) = 1$. Further, the SOM grid exhibits high resolution when vectors that are distant in the input space are not mapped onto neighboring neurons. Quantization error, Q.E., measures resolution as the average distance between each input vector and its BMU: $Q.E. = \frac{1}{n}\sum_{i=1}^{n} \| \vec{x}_i - \vec{m}_c \|$.

The internal validity of clustering can be assessed via the Davies-Bouldin Index (Davies and Bouldin, 1979), defined as: $DBI = \frac{1}{C}\sum_{i=1}^{C} \max_{i \neq j}\left\{\frac{\Delta(C_i)+\Delta(C_j)}{\delta(C_i,C_j)}\right\}$; where $C = 7$, the number of clusters identified by SOM, $\delta(C_i, C_j)$, and $\{\Delta(C_i), \Delta(C_j)\}$ the intercluster and intracluster distances respectively. A small value indicates highly-compact clusters whose centroids are well-separated.

## 2.5 Experiments

The paper conducts the experiments in six steps. First, the paper uses bootstrap to draw one hundred random samples with replacement from the empirical distribution of the input dataset, $X = (x_{ik})_{n(N+m)}$. Second, for each bootstrapped sample, the paper uses SOM-Toolbox for Matlab (Vesanto et al., 2000) to train a SOM in batch mode with hexagonal topology and Gaussian neighborhood. Third, to cluster the neurons and their associated codevectors, the paper uses the Unified-distance matrix (i.e. U-matrix) (Ultsch and Siemon, 1990). The U-matrix calculates the average distance between a neuron's codevector and that of its neighboring neurons and then superimposes this distance as a height, Z coordinate, on the SOM grid. Because Unified-distance matrix superimposes the height between rather than at neighboring neurons, there are more than 820 neurons on the grid. The dark and light areas on the SOM grid indicate long and short distances between neurons respectively; dark areas denote cluster boundaries whereas light areas denote clusters, as shown in Fig. 2. Fourth, for each SOM, the paper calculates the following three quality measures: (i) topographic error, T.E, (ii)

quantization error, Q.E, and (iii) Davies-Bouldin Index, DBI.

Fifth, the paper uses the bootstrap bias-corrected with acceleration method, B = 1,000, (DiCiccio and Efron, 1996) in order to estimate the two-sided 95% confidence interval of the mean and standard deviation for the foregoing quality measures. The rationale behind this statistical analysis is threefold. First, the paper must ensure that SOM converges to a stable and adequate state rather than to a local minimum; second, the SOM grid, shown in Fig. 2, preserves faithfully the topology of the accounting dataset; and third, the clusters identified on the SOM grid possess good statistical properties in terms of compactness and separation, and are also relevant and meaningful within an accounting context. This analysis could provide evidence on the internal consistency of the tool, and also on the ability of the results to be generalized beyond the particular dataset. Finally, for the proposed tool to be able to benefit its intended users, it must correspond accurately with the economic reality of the case company. To this end, the paper benchmarks the tool's output against the financial statements prepared by the case company.

# 3 Results Presentation and Discussion

The proposed tool generates a two-dimensional SOM grid as depicted in Fig. 2, presenting the non-linear projection of the 25,440 accounting transactions as well as identifying seven clusters. As shown in Fig. 2, each of the 29 account classes conveys two pieces of information, as follows: (i) total number of its respective accounting transactions, shown in brackets, and (ii) total USD amount of these transactions. Positive and negative amounts denote debit and credit balances respectively. The 29 account classes are represented on the SOM grid by as many neurons identified by their corresponding numbers. For example, "Administration Expenses" consists of 1,236 transactions having a total debit balance of USD 80,509 and being represented by neuron 493.

Table 2: Two-sided 95% confidence intervals using the bootstrap bias-corrected with acceleration method, B=1,000.

| Quality | Mean | | | Std.Deviation | | |
|---|---|---|---|---|---|---|
| Measures | Lower | Upper | Std.Err. | Lower | Upper | Std.Err. |
| T.E | 0.269 | 0.336 | 0.049 | 0.150 | 0.192 | 0.029 |
| Q.E | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 |
| DBI | 1.037 | 1.107 | 0.047 | 0.153 | 0.207 | 0.036 |

The quality of SOM grid is evaluated by the topographic and quantization errors as described in Table 2. Both quality measures exhibit small values and narrow 95% confidence intervals for their respective means and standard deviations. These results provide evidence that the grid can map the input data accurately as well as preserve

faithfully the original topology; they also suggest that the tool can be generalized to cases beyond the particular accounting database.

The entry for quantization error, Q.E, in Table 2 merits additional discussion, because it points towards the convergence of SOM to a stable state. The values reported for Q.E are truncated to the third decimal place, because the corresponding original values are in the order of $10^{-5}$. A mathematical analysis of SOM convergence lies beyond the scope of this paper. Suffice it to say that in a stable state we expect $E\{h_{cj}(\vec{x}_i - \vec{m}_j^{\,*})\} = 0$, where $E\{.\}$ denotes the expectation function, and $\vec{m}_j^{\,*} = \lim_{t\to\infty} \vec{m}_j(t)$ (Kohonen, 1997, p.113); the rest of the notations are defined in Section 2.3. The minuscule Q.E values confirm that SOM converges to a stable state producing an optimal grid, Fig. 2, of the input dataset.

Furthermore, a visual inspection of the SOM grid, Fig. 2, indicates that the proposed tool can cluster the accounting transactions in seven homogeneous and well-separated clusters. This result is corroborated by statistical analysis, Table 2, that shows both the mean and standard deviation of Davies-Bouldin Index to have small values and narrow 95% confidence intervals. More importantly, as we demonstrate in Table 3, the seven clusters are meaningful and interpretable from an accounting perspective. In particular, clusters 3 and 4 represent income and expense items respectively, and their combination makes up the Income Statement. Clusters 7, 2, and 1 correspond to “Non-Current Assets”, “Trade and Other Receivables”, and “Cash and Cash Equivalents” respectively. These three clusters compose the “Assets” side of the Balance Sheet Statement. Clusters 6 and 5 comprise “Equity” and “Current Liabilities” respectively, and collectively form the “Equity and Liabilities” side of the Balance Sheet Statement.

A closer investigation reveals that the proposed tool enjoys two properties that can enhance its potential applications. First, it can preserve the semantic relationship between one account class and another; this relationship can be quantified in terms of the Euclidean distance between neurons in Fig. 2. For example, “Professional Expenses” is closer to “Finance Expenses” than either is to “Sales Taxes Receivable”; a proximity that is proportional to their relationship according to accounting theory: the first two are part of “Expenses”, whereas the third is part of “Trade and Other Receivables”. Second, it can compress a large number of similar transactions onto a single neuron. For example, neuron 46 represents all the 3,242 transactions concerning “Accounts Payable”.

Given these properties, a manager can apply the proposed tool to check whether a database can process the right transactions accurately. For example, Fig. 2 indicates that although an entry in “Cash in Bank” consists of 85 transactions, it has a nil balance, thereby signalling the occurrence of suspicious transactions that need to be followed up.

Shown in Table 3, the results from the benchmarking exercise demonstrate that the

proposed tool corresponds accurately with the financial statements prepared by the case company; this correspondence lends much credence to the tool. Indeed, the tool's output, shown in Fig. 2, conveys all the necessary and relevant information to function as financial statements.

Table 3: Results from benchmarking the tool's output against company's financial statements.

| Company's Financial Statements | | SOM grid (Fig. 2) | |
|---|---|---|---|
| **Account Class** | **USD** | **Neuron** | **Cluster** |
| **Income Statement for the year ending 31 December 2006** | | | |
| **Income** | | | |
| **Exchange Gain** | **534,475** | **803** | **3** |
| **Fees Received** | **480,000** | **797** | **3** |
| **Other Income Received** | **240,750** | **716** | **3** |
| **Insurance Commission Received** | **1,607,052** | **791** | **3** |
| **Bank Interest Received** | **449,440** | **549** | **3** |
| **Other Interest Received** | **1,166,997** | **637** | **3** |
| **Total Income** | **4,478,714** | | |
| **Expenses** | | | |
| **Depreciation Expenses** | **54,577** | **584** | **4** |
| **Fees and Commissions** | **98,871** | **703** | **4** |
| **Administration Expenses** | **80,509** | **493** | **4** |
| **Finance Expenses** | **73,430** | **336** | **4** |
| **Office Expenses** | **103,575** | **247** | **4** |
| **Professional Expenses** | **96,308** | **539** | **4** |
| **Salaries and Wages** | **649,258** | **780** | **4** |
| **Total Expenses** | **1,156,528** | | |
| **Profit for the year** | **3,322,186** | | |
| **Balance Sheet as at 31 December 2006** | | | |
| **Assets** | | | |
| **Non-Current Assets** | | | |
| **Property, Plants and Equipment_Costs** | **330,053** | **390** | **7** |
| **Property, Plants and Equipment_Depr.** | **-201,395** | **518** | **7** |
| **Investment in Subsidiaries** | **35,250** | **185** | **7** |
| **Current Assets** | | | |
| **Trade and Other Receivables** | | | |
| **Short-term Loans Receivable** | **22,148,046** | **239** | **2** |
| **Other Debtors and Receivables** | **185,158** | **287** | **2** |
| **Sales Taxes Receivable** | **56,133** | **688** | **2** |
| **Other Trade Debtors** | **563,750** | **441** | **2** |
| **Insurance Receivable** | **-425,391** | **816** | **2** |
| **Trade Debtors** | **-6,880,761** | **573** | **2** |
| **Cash and Cash Equivalents** | | | |
| **Cash in Bank** | **6,737,785** | **33** | **1** |
| **Cash at Hand** | **3,485** | **191** | **1** |
| **Total Assets** | **22,552,113** | | |
| **Equity and Liabilities** | | | |
| **Equity** | | | |
| **Issued Capital** | **21,503** | **135** | **6** |
| **Retained Earnings** | **13,935,370** | **342** | **6** |
| **Profit for the year** | **3,322,186** | **221** | **6** |
| **Total Equity** | **17,279,059** | | |
| **Liabilities** | | | |
| **Current Liabilities** | | | |
| **Accounts Payable** | **244,690** | **46** | **5** |
| **Trade Creditors** | **5,028,364** | **17** | **5** |
| **Total Equity and Liabilities** | **22,552,113** | | |

Further, during the benchmarking exercise two unexpected issues emerged. First, the tool may reduce the time and effort involved in preparing financial statements; an issue that could be the focus of future research. Second, to prepare financial statements, the case company relies on SQL queries and spreadsheets. As both procedures are prone to human error, the tool provided a much-needed method for the case company to assess the accuracy of their reporting function. Specifically, the tool clusters correctly "Insurance Receivable" and "Trade Debtors" in cluster 2, "Trade and Other Receivables", as shown in Fig. 2 and Table 3. However, both items have credit balances, whereas their normal balances are debit; a discrepancy that points towards erroneous transactions warranting further investigation.

Although coding is still in progress, the paper deems it necessary to estimate the algorithmic complexity of the tool in order to provide a machine-independent measure for evaluating the tool. As discussed in Sections 2.2 and 2.3, the tool performs two main operations, as follows: (i) it uses Dijkstra's algorithm to derive the distance matrix, and (ii) it uses SOM to process the input dataset. Dijkstra's algorithm has a $O(N^3)$ complexity, and $O(N^2)$ memory consumption for storing the distance matrix, $D$, where $N$ denotes the number of account classes. SOM has $O(nld)$ complexity for searching BMUs and updating codevectors, where $n$ and $l$ denote the number of input vectors and neurons respectively, and $d$ is the dimensionality of input space. If the number of neurons is set to be proportional to $\sqrt{n}$, then the complexity becomes $O(n^{1.5}d)$. The memory consumption of SOM is $O((n + l)d)$ for storing the input and codevector matrices, and $O(l^2)$ for storing the inter-neuron distance matrix.

Nonetheless, the proposed tool suffers from certain limitations that may restrict its potential applications. First, it is deterministic, and as such it does not address uncertainty; a limitation that prevents a user from applying it to budgeting and forecasting. Second, once coding has been completed, applying the tool ought to be straightforward; interpreting its output, however, requires a user to have some familiarity with SOM. Finally, the quality measures, Table 2, should be interpreted with caution, because topographic and quantization errors are data-dependent, and small values may be the result of SOM overfitting the input data rather than projecting them accurately.

# 4 Conclusions and Future Research

Based on the IASB XBRL Taxonomy and SOM, this paper presents a semi-supervised tool for clustering accounting databases in order to supplement internal control procedures. While a wealth of pronouncements require a company's management to implement and maintain adequate internal controls, these pronouncements provide little guidance on the design and application of specific tools. Further, in contrast to published

financial statements, existing literature has paid insufficient attention to accounting databases. These issues suggest that the clustering of accounting databases as an internal control procedure has not been fully explored.

Empirical analyses reveal that the proposed tool can cluster accounting databases in homogeneous and well-separated clusters that can be interpreted from an accounting perspective. Additional investigations indicate that the tool can preserve the semantic relationships between account classes, and also compress a large number of similar accounting transactions. Further, benchmarking results suggest that the tool's output corresponds accurately with the financial statements prepared by the case company. The paper's findings demonstrate that the tool can have practical applications in the domain of accounting, such as: (i) providing a holistic picture of an accounting database at a given point in time, (ii) assisting a manager to supervise the processing of accounting transactions, and (iii) enabling a manager to assess the accuracy of financial statements.

Ongoing research aims at coding the proposed tool as a toolbox for Matlab, and subsequently releasing it in the public domain. Additional experiments could investigate the tool's potential to produce financial statements. A further development could be to apply the proposed tool at consecutive time-intervals (e.g. monthly, quarterly) and use the results thus obtained for financial benchmarking or for moving-window comparisons. A promising line of future research could be the validation of the paper's results against those of independent auditors as well as the evaluation of the proposed tool within the end-user computing satisfaction framework (i.e. EUCS) (Doll et al., 1988).

**Acknowledgements**

We are especially indebted to the company's Finance Director, who wishes to remain anonymous, for providing us with data without which we could not have pursued this paper. We are grateful to the HANKEN Foundation for their financial support, and we also thank Mr. Tom Lindström and Prof. Anders Tallberg for their insightful suggestions.

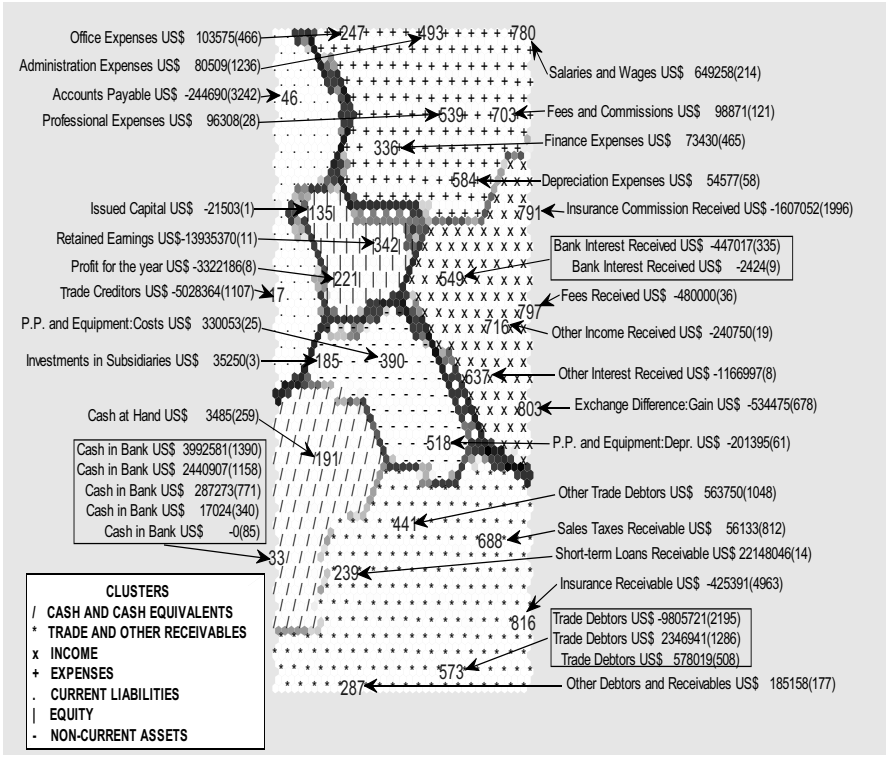



Figure 2: **The tool's output, SOM grid.** The 29 account classes are represented on the SOM grid by as many neurons identified by their corresponding numbers. For each account class, the total number of transactions is shown in brackets; positive and negative US$ amounts denote debit and credit balances respectively. The cluster boundaries are defined by the U-matrix.

## REFERENCES

AcSB (2009), *Exposure Draft: Adopting IFRSs in Canada, II*, Canadian Accounting Standard Board (AcSB), Toronto, Canada.

AICPA (2001), 'Statement On Auditing Standards 94 (SAS 94): Effect of information technology on the auditor's consideration of internal control in a financial statement audit', *American Institute of Certified Public Accountants* .

Argyrou, A. (2009), Clustering hierarchical data using Self-Organizing map: A Graph-Theoretical approach, *in* 'WSOM 2009: Proceedings of the 7th International Workshop on Advances in Self-Organizing Maps', LNCS 5629, Springer-Verlag, pp. 19–27.

Canadian Institute of Chartered Accountants (2004), *IT Control Assessments in the context of CEO/CFO Certification*, Toronto, Canada.

COSO (1992), *Internal Control - An Integrated Framework*, The Committee of Sponsoring Organizations of the Treadway Commission (COSO). AICPA, New Jersey, USA.

Davies, D. and Bouldin, D. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2), 224–227.

DiCiccio, T. J. and Efron, B. (1996), 'Bootstrap confidence intervals', *Statistical Science* **11**(3), 189–228.

Dijkstra, E. W. (1959), 'A note on two problems in connexion with graphs', *Numerische Mathematik* **1**, 269–271.

Doll, W. J., Doll, W. J. and Torkzadeh, G. (1988), 'The measurement of End-User computing satisfaction.', *MIS Quarterly* **12**(2), 259–274.

Eklund, T., Back, B., Vanharanta, H. and Visa, A. (2003), Financial benchmarking using self-organizing maps studying the international pulp and paper industry, *in* 'Data mining: opportunities and challenges', IGI Publishing, pp. 323–349.

European Parliament (2002), 'Regulation (EC) No 1606/2002 of the European Parliament and of the Council of 19 July 2002 on the application of international accounting standards.', Official Journal L 243/1.

Haykin, S. (1999), *Neural Networks. A Comprehensive Foundation*, second edn, Prentice Hall International, Upper Saddle River, New Jersey, USA.

Huysmans, J., Baesens, B., Vanthienen, J. and van Gestel, T. (2006), 'Failure prediction with self organizing maps', *Expert Systems with Applications* **30**(3), 479–487.

IASB (2009), *International Financial Reporting Standards (IFRS) 2009*, International Accounting Standards Committee Foundation (IASCF), London, United Kingdom.

IASCF (2009), *IFRS Taxonomy Guide 2009 (XBRL)*, International Accounting Standards Committee Foundation (IASCF), London, United Kingdom.

Isa, D., Kallimani, V. and Lee, L. H. (2009), 'Using the self organizing map for clustering of text documents', *Expert Systems with Applications* **36**(5), 9584–9591.

IT Governance Institute (2000), *Governance, Control and audit for information technology*, COBIT 3rd edn, IT Governance Institute, Rolling Meadows, IL, USA.

Jungnickel, D. (2002), *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Volume 5, English edn, Springer, Berlin, Germany.

Kiang, M. Y. and Fisher, D. M. (2008), 'Selecting the right MBA schools - an application of self-organizing map networks', *Expert Systems with Applications* **35**(3), 946–955.

Kiang, M. Y., Fisher, D. M., Chen, J. V., Fisher, S. A. and Chi, R. T. (2009), 'The application of SOM as a decision support tool to identify AACSB peer schools', *Decision Support Systems* **47**(1), 51–59.

Kiang, M. Y., Hu, M. Y. and Fisher, D. M. (2006), 'An extended self-organizing map network for market segmentation–a telecommunication example', *Decision Support Systems* **42**(1), 36–47.

Kiviluoto, K. (1996), Topology preservation in Self-Organizing maps, *in* 'Proceeding of International Conference on Neural Networks (ICNN'96)', pp. 294–299.

Kohonen, T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences, Volume 30, second edn, Springer-Verlag, Heidelberg, Germany.

Kohonen, T. and Somervuo, P. (1998), 'Self-organizing maps of symbol strings', *Neurocomputing* **21**(1-3), 19–30.

Kohonen, T. and Somervuo, P. (2002), 'How to make large self-organizing maps for nonvectorial data', *Neural Networks* **15**(8-9), 945–952.

Lee, K., Booth, D. and Alam, P. (2005), 'A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms', *Expert Systems with Applications* **29**(1), 1–16.

Serrano-Cinca, C. (1996), 'Self organizing neural networks for financial diagnosis', *Decision Support Systems* **17**(3), 227–238.

Ultsch, A. and Siemon, H. (1990), Kohonen's self organizing feature maps for exploratory data analysis, *in* 'Proceedings International Neural Network Conference', Kluwer Academic Press, Dordrecht, Netherlands, pp. 305–308.

U.S. Congress (2002), 'Sarbanes-Oxley Act of 2002, H.R.3763'.

U.S. Securities and Exchange Commission (2008), 'Roadmap for the Potential Use of Financial Statements Prepared in Accordance With International Financial Reporting Standards by U.S. Issuers', *Federal Register* **73:226 (November 21, 2008)**, 70816–70856.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), SOM Toolbox for Matlab 5, Technical Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland.

Wolfram Research Inc. (2007), *Mathematica, Version 6.0*, Wolfram Research Inc., Champaign, IL, USA.

**Using Self-Organizing Map for data mining:**
**A synthesis with accounting applications**

**Andriy Andreev and Argyris Argyrou**

**Abstract**
The self-organizing map (i.e. SOM) has inspired a voluminous body of literature in a number of diverse research domains. We present a synthesis of the pertinent literature as well as demonstrate, via a case study, how SOM can be applied in clustering accounting databases. The synthesis explicates SOM's theoretical foundations, presents metrics for evaluating its performance, explains the main extensions of SOM, and discusses its main financial applications. The case study illustrates how SOM can identify interesting and meaningful clusters that may exist in accounting databases. The paper extends the relevant literature in that it synthesises and clarifies the salient features of a research area that intersects the domains of SOM, data mining, and accounting.

# 1 Introduction

This paper aims to present a coherent synthesis of the literature pertinent to self-organizing map (SOM; Kohonen, 1982, 1997) as well as demonstrate how SOM can be applied as a data-mining tool in the domain of accounting. The motivation behind the paper emanates from the considerable academic interest and body of literature SOM has inspired in a multitude of research domains; until 2005, SOM has paved the way for 7,718 publications [1]. The paper differs from previous reviews of SOM (Kohonen, 1998, 2008), and (Yin, 2008), in that it addresses SOM from a data-mining perspective, and places much emphasis on the main financial applications of SOM. The contribution and novelty of the paper lie in it synthesising an expansive and fragmented literature pertinent to SOM, focusing on how SOM can perform certain data-mining tasks, and demonstrating the performance of such tasks via a case study that considers the clustering of accounting databases.

In essence, SOM performs a non-linear projection of a multi-dimensional input space to a two-dimensional regular grid that consists of spatially-ordered neurons, and preserves the topology of the input space as faithfully as possible. SOM has been applied successfully in numerous research domains for clustering, data visualization, and feature extraction.

To carry out the synthesis, we adopt the following four organizing principles. First, we select and review papers that explain the SOM algorithm in sufficient details, and exclude those papers that delve into the mathematical complexities and subtleties of SOM. Further, we review only the most prevalent criteria that evaluate the performance of SOM, because there is neither an accepted global criterion nor a consensus about which criteria are the most informative. As the literature abounds with extensions of SOM, we delimit the

[1] A complete bibliography is available at: http://www.cis.hut.fi/research/som-bibl/

synthesis to those extensions that enhance the ability of SOM to perform data-mining tasks (e.g. clustering of non-vectorial data). Finally, to review the financial applications of SOM, we pay particular attention to the subjects of bankruptcy prediction, financial benchmarking, and clustering of hedge funds.

To conduct the case study, the paper uses a set of accounting transactions that describe the economic activities of an international shipping company for fiscal year 2007. It first pre-processes the accounting transactions for SOM-based processing, and then uses bootstrap to select random samples with replacement from the empirical distribution of the transactions. For each bootstrapped sample, the paper trains a SOM, and subsequently evaluates the performance of each SOM by calculating three metrics: (i) quantization error, (ii) topographic error, and (iii) Davies-Bouldin index. Finally, it estimates the two-sided 95% confidence interval of the mean and standard deviation of the foregoing metrics.

The rest of the paper is organized as follows. Section 2 introduces data pre-processing; an activity that precedes most of data-mining tasks. Section 3 elaborates on the SOM algorithm and its main constituents, and Section 4 presents three metrics for evaluating the performance of SOM as well as a criterion for assessing the internal validity of clustering. Section 5 discusses the main extensions of SOM, and Section 6 reviews the main financial applications of SOM. Section 7 demonstrates, via a case study, how SOM can be applied in identifying and visualizing meaningful clusters that may exist in accounting databases.

# 2 Data pre-processing

For SOM to operate efficiently and yield meaningful results, a researcher must pay attention to data pre-processing; this activity involves three main tasks: (i) understanding the different types of variables, (ii) selecting an appropriate and valid distance metric, and (iii) rescaling input variables.

## 2.1 Types of variables

Understanding the various types of variables guides a researcher in selecting mathematical and statistical operations that are valid for each type as well as in choosing permissible transformations that preserve the original meaning of variables.

Four types of variables can be identified, as follows (Stevens, 1946): (i) nominal, (ii) ordinal, (iii) interval, and (iv) ratio. The order is cumulative, which means that each type subsumes the properties of its predecessor. Nominal variables take as values different names or labels (e.g. name of employees), and hence they are not amenable to any mathematical operation other than a simple function of equality. Further, ordinal or hierarchical variables can be ranked by order (e.g. examination grades, hardness of

minerals). Interval variables take values whose differences are meaningful, but ratios are not. The reason being that interval variables lack a "true" zero point; a zero value does not entail the absence of a variable (e.g. temperature in Celsius). In contrast, ratio variables (e.g. length) take values whose differences, and ratios are meaningful. For completeness, nominal and ordinal variables are collectively described as categorical data, whereas interval and ratio variables as numerical data.

A further distinction can be made between discrete and continuous data; the former is associated with counting and can take a value from a finite set of real integers, whereas the latter is associated with physical measurements and thus it can take any numerical value within an interval.

## 2.2 Distance metrics

Selecting a valid distance metric takes on added importance in an unsupervised-learning task (e.g. clustering), because in performing such task, a researcher has no recourse to information concerning the class labels of input data. On the other hand, in a supervised-learning task (e.g. classification), the classification error can form an external criterion that can be optimised to yield a valid distance metric. A distance metric over $\mathbb{R}^n$ is considered to be valid only if it can assign distances that are proportionate to the similarities between data points.

In particular, given three vectors $\vec{x}$, $\vec{y}$, $\vec{z} \in \mathbb{R}^n$, a distance metric $d(\vec{x},\vec{y})$ must satisfy the conditions of a metric space, as follows (Jungnickel, 2002, p.65):

1. $d(\vec{x},\vec{y}) > 0$ for all $\vec{x} \neq \vec{y}$ (non-negativity),
2. $d(\vec{x},\vec{y}) = 0$ if and only if $\vec{x} = \vec{y}$ (distinguishability),
3. $d(\vec{x},\vec{y}) = d(\vec{y},\vec{x})$ for all $\vec{x}$ and $\vec{y}$ (symmetry),
4. $d(\vec{x},\vec{z}) \leq d(\vec{x},\vec{y}) + d(\vec{y},\vec{z})$ for all $\vec{x}$, $\vec{y}$, $\vec{z}$ (triangular inequality).

In a metric space, the most prevalent distance metric is the Euclidean distance defined as:

$$d_E(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n} (\vec{x}_i - \vec{y}_i)^2} \quad . \tag{1}$$

Further, Mahalanobis distance takes into account the correlations between data, and hence it can yield more accurate results than Euclidean distance does. However, any benefits derived from the improved accuracy may be outweighed by the computational costs involved in calculating the covariance matrix.

$$d_M(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})S^{-1}(\vec{x}-\vec{y})^T} \quad . \tag{2}$$

Mahalanobis distance reduces to the Euclidean if the covariance matrix, $S$, is equal to the identity matrix, and to the normalized Euclidean if $S$ is diagonal.

### 2.3 Rescaling input variables

The rationale behind rescaling input variables is threefold (Bishop, 1995, p.298): (i) to ensure that input variables reflect their relative importance, (ii) different variables may have different units of measurement, and hence their typical values may differ considerably, and (iii) to facilitate the initialisation of codevectors. For example, consider a dataset having two variables: "weight" and "height". The former takes values in the range $\{30Kg, 40Kg, \ldots, 100Kg\}$, whereas the latter in the range $\{1.3m, 1.4m, \ldots, 2m\}$; without rescaling, the "weight" variable is going to dominate the distance calculations, and thus it would bias SOM. Linear and non-linear operations can rescale input variables so that they could take values in the interval $[0, 1]$, or values having zero mean and unit variance. The latter is particularly appropriate when the Euclidean distance is used as a distance metric.

## 3 Self-Organizing Map

### 3.1 Introduction to SOM

SOM performs two closely coupled operations: (i) vector quantization, and (ii) non-linear vector projection. First, SOM regresses a set of codevectors into input data in a non-parametric and non-linear manner. In doing so, SOM partitions input data into a finite number of disjoint and contiguous spaces each of which is represented by a codevector. The codevectors are estimated in such a way that minimises the distance between an input vector and its closest codevector; this estimation process continues iteratively until it converges to a stationary state at which codevectors are not updated any further. The vector-quantization operation reduces input data to a much smaller, albeit representative, set of codevectors.

Second, SOM projects the codevectors onto a regular two-dimensional grid of neurons; a grid can be either hexagonal or rectangular depending on whether a neuron has either six or four neighbours. In either case, the neurons are spatially-ordered, and thus the grid can preserve the neighbourhood relations or topology between input data as faithfully as possible. A neuron $k$ is described by the tuple $(\vec{m}_k, \vec{p}_k)$, where $\vec{m}_k \in \mathbb{R}^{\mathsf{d}}$ is a codevector, $\vec{p}_k \in \mathbb{R}^2$ is a location vector on the SOM grid, $d$ denotes the dimensionality of input space, and $k = 1, 2, \ldots, K$ represents the number of neurons. The codevectors are used for vector quantization, and their corresponding locations on the SOM grid are used for vector projection.

To elaborate, given an input dataset, $X = \left(\vec{x}_{ij}\right)_{nd}$, SOM constructs a set of codevectors,

$M = \left(\vec{m}_{kj}\right)_{Kd}$, where $\vec{x}_i$, $\vec{m}_k \in \mathbb{R}^d$ are row vectors of $X$ and $M$ respectively, and $n$ represents the number of input vectors. The number of neurons can be estimated by using the heuristic formula: $K = 5\sqrt{n}$ (Vesanto et al., 2000, p.30). As a first approximation to input data, SOM can initialise the codevectors by using either random vectors or vectors derived from the hyperplane that is spanned by the two largest principal components of input data (Kohonen, 1998). Although SOM is very robust to either of the two initialisation approaches, the latter approach is preferable, because it enables SOM to converge more efficiently than random initialisation does. The reason for this is twofold: (i) SOM gets organized approximately from the beginning, and (ii) a researcher can start with a narrower neighbourhood kernel and a smaller learning rate (Kohonen, 1998).

## 3.2 Formation of SOM

The formation of SOM involves the following three iterative processes: (i) competition, (ii) co-operation, and (iii) adaptation (Haykin, 1999, p.447). First, in the competition process, each input vector, $\vec{x}_i \in \mathbb{R}^d$, is compared with all codevectors, $\vec{m}_k \in \mathbb{R}^d$, and the best match in terms of the smallest Euclidean distance, $\| \vec{x}_i - \vec{m}_k \|$, is mapped onto neuron $k$ that is termed the best-matching unit (i.e. BMU) and denoted by the subscript $c$ (Kohonen, 1997, p.86):

$$\| \vec{x}_i - \vec{m}_c \| = \min_k \{ \| \vec{x}_i - \vec{m}_k \| \} \quad , \tag{3}$$

or equivalently:

$$c = \underset{k}{\operatorname{argmin}} \{ \| \vec{x}_i - \vec{m}_k \| \} \quad . \tag{4}$$

Second, in the co-operation process, the BMU locates the centre of a neighbourhood kernel, $h_{ck}(t)$, which is usually a Gaussian function defined as:

$$h_{ck}(t) = exp\left[ -\frac{\| \vec{p}_c - \vec{p}_k \|^2}{2\sigma^2(t)} \right] \quad , \tag{5}$$

where $\vec{p}_c$, $\vec{p}_k \in \mathbb{R}^2$ are the location vectors of BMU and neuron $k$ respectively, $t$ denotes discrete time, and $\sigma(t)$ defines the width of the kernel; $\sigma(t)$ is a monotonically decreasing function of time (Kohonen, 1997, p.87). The motivation behind the neighbourhood kernel is that an activated neuron, BMU, excites adjacent neurons to a greater degree than that to which it excites distant neurons. It can be deduced from Eq. 5 that the neighbourhood kernel decays proportionately with discrete time $t$ as well as with the lateral distance between the BMU and the "excited" neurons, that is: $h_{ck}(t) \to 0$ as $\| \vec{p}_c - \vec{p}_k \|$ increases.

The neighbourhood kernel enables the SOM grid to preserve the topology, or neighbourhood relations, between input data by allowing the codevectors to be updated

according to their respective proximity to the BMU; the closer to the BMU a codevector is, the greater the extent of its updating is, whereas codevectors lying outside the neighbourhood of BMU are not updated at all. As a result, the neurons on the SOM grid become spatially-ordered in the sense that neighbouring neurons have similar codevectors, and thus they represent similar areas in the input space.

A SOM grid enjoys two main properties (Ritter and Kohonen, 1989): (i) it preserves the neighbourhood or topological relations between input data as faithfully as possible, and (ii) it represents a mapping of input vectors that is determined by their density function, whereby more frequent vectors are mapped to a larger area on the grid. Because of these properties, SOM can map similar and high-frequent input vectors to a localised area, capture the overall topology of cluster arrangements, and perform non-linear dimensionality reduction.

Third, in the adaptive process, the sequence-training SOM updates recursively codevector $\vec{m}_k$ as follows:

$$\vec{m}_k(t+1) = \vec{m}_k(t) + a(t)h_{ck}(t)\left[\vec{x}_i(t) - \vec{m}_k(t)\right] \quad , \tag{6}$$

where $0 < a(t) \leqslant 1$ is a learning rate at discrete time $t$; $a(t)$ is a non-increasing function of time, for example: $a(t) = a_0\left(1 - \frac{t}{T}\right)$ or $a(t) = a_0\left(\frac{0.005}{a_0}\right)^{\frac{t}{T}}$, where $a_0$ is the initial learning rate, and $T$ is the training length. The updating rule, described in Eq. 6, is motivated by the Hebbian law for synaptic modification, and includes a non-linear forgetting process for synaptic strength (Ritter and Kohonen, 1989).

Batch-SOM is a variant of the updating rule, described in Eq. 6. It estimates the BMU according to Eq. 3, but updates the codevectors only at the end of each epoch, which is a complete presentation of input data, rather than recursively, as follows (Vesanto et al., 2000, p.9):

$$\vec{m}_k(t+1) = \frac{\sum_{i=1}^{n} h_{ck}(t)\vec{x}_i}{\sum_{i=1}^{n} h_{ck}(t)} \quad . \tag{7}$$

Batch-SOM does not contain a learning rate, and estimates codevectors as averages of input vectors being weighed by neighbourhood kernel $h_{ck}(t)$. It can also be expressed in terms of Voronoi cells: $V_k = \{\vec{x}_i \mid \| \vec{x}_i - \vec{m}_k \| < \| \vec{x}_i - \vec{m}_j \| \; \forall\, k \neq j\}$, as follows (Vesanto et al., 2000, p.11):

$$\vec{m}_k(t+1) = \frac{\sum_{k=1}^{K} h_{ck}(t)\vec{s}_k(t)}{\sum_{k=1}^{K} h_{ck}(t)N_k} \quad , \tag{8}$$

where $K$ is the number of codevectors, and hence that of Voronoi cells, $N_k$ and $\vec{s}_k(t) = \sum_{\vec{x} \in V_k} \vec{x}$ denote the number and sum of input vectors that belong in Voronoi cell $V_k$, respectively.

Batch-SOM enjoys certain advantages vis-a-vis sequence-SOM (Lawrence et al., 1999): (i) it is less computationally expensive, (ii) because the updating of codevectors is not

recursive, there is no dependence on the order in which input vectors appear, and (iii) it mitigates concerns that input vectors that appear at a later iteration may affect the result disproportionately.

SOM converges to a stationary state when the codevectors do not get updated any further. This entails that at a stationary state we require that $E\{\vec{m}_k(t+1)\}$ must be equal to $E\{\vec{m}_k(t)\}$ for $t \to \infty$, even if $h_{ck}(t)$ was non-zero (Kohonen, 1998). The foregoing leads to the convergence criterion: $E\{h_{ck}(\vec{x}_i - \lim_{t\to\infty} \vec{m}_k(t))\} = 0$, where $E\{.\}$ denotes the expectation function (Kohonen, 1997, p.113).

# 4 Performance metrics and cluster validity

In the general case, the updating rule, described in Eq. 6, is not the gradient of any cost function that can be optimised for SOM to converge to at least a local optimum (Erwin et al., 1992). The lack of a cost function has hindered researchers in their efforts to derive a global criterion that can assess the quality and performance of SOM. In the absence of a global criterion, we confine the discussion to the following three criteria, each evaluating a distinct facet of SOM: (i) Quantization error, Q.E, (ii) Topographic error, T.E, and (iii) Distortion error, D.E.

First, Q.E quantifies the resolution of a SOM grid. A som grid exhibits high resolution if the input vectors that are near to one another are projected nearby on the SOM grid, and the input vectors that are farther apart are projected farther apart on the SOM grid as well. The average Q.E for a SOM grid is defined as: $Q.E = \frac{1}{n}\sum_{i=1}^{n} \parallel \vec{x}_i - \vec{m}_c \parallel$; and the lower its value is, the higher the SOM's resolution is. Second, T.E quantifies the topology preservation of SOM mapping as follows: $T.E = \frac{1}{n}\sum_{i=1}^{n} \varepsilon(\vec{x}_i)$; if the first and second BMUs of $\vec{x}_i$ are adjacent, then $\varepsilon(\vec{x}_i) = 0$, otherwise $\varepsilon(\vec{x}_i) = 1$ (Kiviluoto, 1996). Further, T.E indicates whether a mapping from $\mathbb{R}^{\mathsf{d}}$ to $\mathbb{R}^{2}$ is continuous; if the first and second BMUs of an input vector are not adjacent on the SOM grid, then the mapping is not continuous near that input vector.

Although Q.E and T.E are the most prevalent performance metrics, a number of caveats apply to their interpretation. First, the two metrics are not independent, instead there is a trade-off between them being moderated by the size of the kernel width (i.e. $\sigma(t)$). Second, Q.E and T.E are dependent on input data; a feature that precludes their use from comparing SOMs that are trained on different datasets. Third, good results may be achieved by SOM overfitting the data rather than it enjoying high resolution and preserving topology faithfully. Finally, Q.E takes its minimum value when the neighbourhood kernel, $h_{ck}(t)$, becomes equivalent to Kronecker's delta, that is: $h_{ck}(t) = \begin{cases} 1 & for \quad c = k, \\ 0 & for \quad c \neq k, \end{cases}$. However, in that case SOM reduces to the classical k-means algorithm, and thus it does not possess any self-organizing capabilities.

Further, the distortion error can act as a local cost function, provided that the input data are discrete and the neighbourhood kernel is fixed (Vesanto et al., 2003). Given these assumptions, the sequence-training rule in Eq.6 estimates approximately the gradient of the distortion error (Graepel et al., 1997). The distortion error is defined as:

$$D.E = \sum_{i=1}^{n} \sum_{j=1}^{m} h_{cj} \left( \vec{x}_i - \vec{m}_j \right)^2 \quad , \tag{9}$$

and it can be decomposed into the following three terms (Vesanto et al., 2003):

$$D.E = E_{qx} + E_{nb} + E_{nv} \quad . \tag{10}$$

The term $E_{qx}$ measures the quantization quality of SOM in terms of the variance of input data that belong in each Voronoi set; $E_{nb}$ can be interpreted as the stress or link between the quantizing and ordering properties of SOM; and $E_{nv}$ quantifies the topological quality of SOM. Analytically,

1. $E_{qx} = \sum_{j=1}^{m} N_j H_j Var\{x|j\}$ ,
2. $E_{nb} = \sum_{j=1}^{m} N_j H_j \parallel n_j - \bar{m}_j \parallel^2$ ,
3. $E_{nv} = \sum_{j=1}^{m} N_j H_j Var_h\{m|j\}$ ,

where $Var\{x|j\} = \sum_{x \in V_j} \parallel x - n_j \parallel^2 / N_j$ is the local variance, $Var_h\{m|j\} = \sum_k h_{jk} \parallel m_k - \bar{m}_j \parallel^2 / H_j$ is the weighed variance of the codevectors, and $\bar{m}_j = \sum_k h_{jk} m_k / H_j$ is the weighed mean of the codevectors.

Although the aforesaid criteria have been used extensively in evaluating the performance of SOM, they do not provide any information about the validity of clustering. For this reason, we describe the Davies-Bouldin index (Davies and Bouldin, 1979) that evaluates the internal validity of clustering as follows:

$$DBI = \frac{1}{C} \sum_{i=1}^{C} \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \quad , \tag{11}$$

where $C$ is the number of clusters produced by SOM, $\delta(C_i, C_j)$ denotes inter-cluster distances, and $\Delta(C_i)$ and $\Delta(C_j)$ represent intra-cluster distances. A small value indicates highly-compact clusters whose centroids are well-separated.

# 5 Extensions of SOM

SOM has been extended so that it could address data-mining tasks in the following domains: (i) non-metric spaces, (ii) temporal sequence processing, (iii) clustering, and (iv) visualizing high-dimensional data.

## 5.1 Non-metric spaces

In the domain of non-metric spaces, extensions of SOM include models for clustering text documents, symbol strings, categorical data, and hierarchical data.

### 5.1.1 WEBSOM

WEBSOM extends SOM to clustering, visualizing, and filtering a large collection of text documents; tasks that are essential in the domains of text-mining and information retrieval. WEBSOM consists of two steps: (i) vector space model (Salton et al., 1975), and (ii) dimensionality reduction. In the first step, WEBSOM sets up a word-document matrix whose elements are the weighed frequencies of a word in each document. The frequencies are weighed by using the inverse document frequency (i.e. IDF), so that rare words could get a higher weight than frequent words. The justification for this scoring mechanism being that words that appear rarely enjoy more discriminatory power than words that appear frequently do (Manning et al., 2009, p.119). If a body of documents is classified into known topics, then a word can be weighed by using entropy (Shannon, 1948). For example, let the entropy of word "$\omega$" be: $H_{(\omega)} = -\sum_g \frac{N_g(\omega)}{N_g} log_2 \frac{N_g(\omega)}{N_g}$, and the total entropy of a body of documents be: $H_{max} = log_2 N_g$, where $N_g(\omega)$ denotes the number of times word "$\omega$" occurs in topic $g$, and $N_g$ represents the number of topics. Then, $W_{(\omega)} = H_{max} - H_{(\omega)}$ becomes the weight for word "$\omega$".

However, the vector space model increases dimensionality, because each word has to be represented by a dimension in the document vector. To reduce dimensionality, WEBSOM uses random mapping (Kaski, 1998) rather than the more established technique of latent semantic indexing (i.e. LSI) (Deerwester et al., 1990). The reason is the computational complexity of the former, $O(Nl) + O(n)$, is much lower than that of the latter, $O(Nld)$; where $N$ is the number of documents, $l$ is the average number of different words in a document, $n$ represents the original dimensionality, and $d$ denotes the dimensionality that results from LSI performing singular value decomposition. Further, experimental results have suggested that the classification accuracy of random mapping is very close to that of LSI (Kohonen et al., 2000).

WEBSOM has been applied[2] in clustering and visualizing collections of text documents, such as: articles from Encyclopaedia Britannica (Lagus et al., 2004), a collection of 7 million patent abstracts that are available in electronic form and written in English (Kohonen et al., 2000), a number of scientific abstracts[3] (Lagus, 1997), and articles from Usenet newsgroups (Honkela et al., 1996*a*,*b*).

[2]Some of these applications are demonstrated at: http://websom.hut.fi/websom/

[3]This WEBSOM is available at: http://www.cis.hut.fi/wsom97/abstractmap/

### 5.1.2 SOM for symbol strings

Unlike numerical data, symbol strings and other non-vectorial data lack intrinsic quantitative information for a distance metric (e.g. Euclidean) to be used as a similarity metric. To extend SOM to symbol strings and other non-vectorial data, Kohonen (1996), and Kohonen and Somervuo (1998) put forward the following principles: (i) define learning as a succession of conditional averages over subsets of strings, (ii) use batch-SOM, Eq. 7, (iii) define a valid similarity measure over strings (e.g. Levenshtein distance), and (iv) averages over such data are computed as generalized means or medians (Kohonen, 1985). Based on these principles, SOM for symbol strings has been applied in constructing a pronunciation dictionary for speech recognition (Kohonen and Somervuo, 1997), clustering protein sequences (Kohonen and Somervuo, 2002), and identifying novel and interesting clusters from a selection of human endogenous retroviral sequences (Oja et al., 2005).

### 5.1.3 SOM for categorical data

To extend SOM to categorical data, Hsu (2006) proposed the Generalized SOM. In brief, a domain expert first describes input data in terms of a concept hierarchy, and then extends it to a distance hierarchy by assigning a weight to each link that exists on the concept hierarchy. For example, a data point $X$ is represented on a distance hierarchy by the tuple $(N_x, d_x)$, where $N_x$ denotes the leaf node corresponding to $X$, and $d_x$ stands for the distance between the root node and $N_x$. The distance between two points, $X$ and $Y$, can then be defined as:

$$| X - Y | = d_x + d_y - 2d_{LCP(X,Y)} \ , \tag{12}$$

where $d_{LCP(X,Y)}$ is the distance between the root node and the least common point of $X$ and $Y$.

### 5.1.4 SOM for hierarchical data

The graph-theoretical approach (Argyrou, 2009) is founded on graph theory and is decoupled from SOM. It functions as a data pre-processing step that transforms hierarchical data into a numerical representation, and thereby renders them suitable for SOM-based processing. In essence, it operates in two steps. First, it encodes hierarchical data in the form of a directed acyclic graph (i.e. DAG), whereby the root vertex represents the complete set of data, and all other vertices are ordered in such a way that each vertex is a subset of its parent vertex. In doing so, the graph-theoretical approach preserves the semantical relationships, (i.e. $child \prec parent$ relationship), that exist between hierarchical data. Second, it uses Dijkstra's algorithm (Dijkstra, 1959) in order to calculate all pairwise distances between vertices. The calculation yields a distance

matrix that satisfies the conditions of a metric space, specified in Section 2.2, and thus it can form the input dataset to SOM.

## 5.2 SOM for temporal sequence processing

SOM can be applied only to static data, because it ignores the temporal ordering of input data. To elaborate, we draw on the updating rules, Eq. 6 and Eq. 7, that represent the output of a neuron at iteration $t$; iteration $t$ acts as a surrogate for discrete time. At each iteration, the updating rules update codevector $\vec{m}_k$ towards input vector $\vec{x}_i$ by considering the value of $\vec{m}_k$ at the previous iteration, but ignoring any changes to the value of $\vec{x}_i$. As a result, SOM can not capture any temporal context that may exist between consecutive input vectors.

### 5.2.1 Recurrent SOM

Recurrent SOM (i.e. RSOM) incorporates leaky integrators to maintain the temporal context of input vectors, and thus extends SOM to temporal sequence processing (Koskela et al., 1998). It also constitutes an improvement to Temporal Kohonen Map (Chappell and Taylor, 1993) in terms of learning and convergence (Varsta et al., 2001).

RSOM modifies SOM's updating rule, Eq. 6, and models the leaky integrators as follows:

$$\vec{m}_k(t+1) = \vec{m}_k(t) + a(t)h_{ck}(t)\vec{\psi}_k(t,\beta) \quad , \tag{13}$$

$$\vec{\psi}_k(t,\beta) = (1-\beta)\vec{\psi}_k(t-1,\beta) + \beta(\vec{x}_i(t) - \vec{m}_k(t)) \quad , \tag{14}$$

where $\vec{\psi}_k(t,\beta)$ is the leaked difference vector of neuron $k$ at iteration $t$, and $0 < \beta \leqslant 1$ is the leaky coefficient that determines how quickly memory decays. A large $\beta$ entails fast memory decay, i.e. short memory, whereas a small $\beta$ a slower memory loss, i.e. long memory. The leaked difference vector forms the feedback to RSOM; and because the feedback is a vector rather than a scalar, it allows the updating rule, Eq. 13, to capture information about changes in both the magnitude and direction of input vectors.

### 5.2.2 Recursive SOM

Recursive SOM (i.e. RecSOM) uses the output of the entire SOM as its feedback to the next iteration (Voegtlin, 2002). RecSOM defines neuron $k$ by means of two codevectors: (i) a feed-forward codevector, $\vec{m}_k \in \mathbb{R}^{\mathsf{d}}$, which is the same as the codevector defined by SOM, and (ii) a recurrent codevector, $\vec{w}_k(t) \in \mathbb{R}^{|\mathsf{N}|}$ that represents the output of the entire SOM at iteration $t$, where $| N |$ denotes the number of iterations. The distance calculation becomes:

$$d_k(t) = \alpha \parallel \vec{x}_i(t) - \vec{m}_k(t) \parallel^2 + \beta \parallel \vec{y}(t-1) - \vec{w}_k(t) \parallel^2 \quad , \tag{15}$$

where $\alpha,\ \beta > 0,\ \vec{y}(t) = [\exp(-d_1(t), \exp(-d_2(t), \ldots, \exp(-d_K(t)],$ and $k = 1, 2, \ldots, K$ represents the number of neurons. It follows form Eq.15 that the best-matching unit is given by $c = \underset{k}{\operatorname{argmin}}\{d_k(t)\}$.

The updating rule for the feed-forward codevector is the same as SOM's updating rule described in Eq. 6, that is:

$$\vec{m}_k(t+1) = \vec{m}_k(t) + a(t)h_{ck}(t)\left[\vec{x}_i(t) - \vec{m}_k(t)\right] \ , \tag{16}$$

and the updating rule for the recurrent codevector is:

$$\vec{w}_k(t+1) = \vec{w}_k(t) + a(t)h_{ck}(t)\left[\vec{y}(t-1) - \vec{w}_k(t)\right] \ . \tag{17}$$

## 5.3 SOM for cluster analysis

Cluster analysis aims at identifying subsets or clusters in data for which no information about their class-labels is known in advance. Cluster analysis groups data into subsets so that a distance metric (e.g. Euclidean) within a subset is minimized, and that between one subset and another is maximized. This process ensures that the intra-cluster similarity is maximized, and the inter-cluster similarity is minimized, as the distance between two data points is inversely proportional to their similarity.

While k-means algorithm clusters data directly, SOM performs a two-step clustering. First, it regresses a set of codevectors to input data in a non-parametric and non-linear manner; second, the codevectors can be clustered by using either a partitive or an agglomerative algorithm. An input vector belongs to the same cluster as its nearest codevector does.

SOM enjoys a lower computational complexity than the algorithms that cluster data directly (Vesanto and Alhoniemi, 2000). In particular, the computational complexity of k-means algorithm is proportionate to $\sum_{C=2}^{Cmax} NC$, where $Cmax$ denotes the maximum number of clusters, $N$ stands for the number of input vectors, and $C$ represents the number of clusters. In contrast, when codevectors are used as as intermediary step, the computational complexity becomes proportionate to $NM + \sum_C MC$, where $M$ represents the number of codevectors. Further, an agglomerative algorithm starts with either $N$ or $M$ clusters depending on whether it is applied directly on input data or on codevectors. The latter is much less computationally expensive than the former, because the number of codevectors is usually chosen to be approximately equal to the square root of the number of input vectors, $M \approx \sqrt{n}$.

SOM has an additional benefit in that it is tolerant to outliers and noise that may be present in input data; the reason is SOM estimates codevectors as weighed averages of

input data, as shown in Eq. 7. In addition, SOM is robust to missing values, as it performs the distance calculations, $\| \vec{x}_i - \vec{m}_k \|$, by excluding them. This approach yields a valid solution, because the same variables are excluded at each distance calculation (Vesanto et al., 2000, p.7).

For SOM to be able to identify clusters that may exist in input data, the probability density function (i.e. pdf) of codevectors must approximate that of input data; in this case, small and large distances between codevectors indicate dense and sparse areas in the input data, respectively.

For vector quantization algorithms (e.g. k-means), the pdf of codevectors (i.e. $p(\vec{m}_k)$) approximates asymptotically that of the input data, (i.e. $p(x)$), thus $p(\vec{m}_k) \propto p(x)^{\frac{d}{d+r}}$, where $d$ denotes the dimensionality of the input space, and $r$ is the distance norm. Such an approximation has been derived only for the one-dimensional SOM (Ritter, 1991):

$$p(\vec{m}_k) \propto p(x)^{\frac{2}{3} - \frac{1}{3\sigma^2 + 3(\sigma+1)^2}} \quad . \tag{18}$$

The approximation, Eq. 18, is valid only if the neighbourhood width (i.e. $\sigma$) is very large and the number of codevectors tends to infinity. Nonetheless, experimental results have suggested that $p(\vec{m}_k)$ approximates $p(x)$ or some monotone function of $p(x)$ (Kaski and Kohonen, 1996; Kohonen, 1999).

### 5.3.1 Unified-distance matrix

The Unified-distance matrix (i.e. U-matrix) enables SOM to identify and visualize clusters that may be present in the input data (Ultsch and Siemon, 1990). The U-matrix first calculates the average distance between a neuron's codevector and that of its immediate neighbouring neurons:

$$h(k) = \frac{1}{| N_k |} \sum_{i \in N_k} d(\vec{m}_k, \vec{m}_i) \quad , \tag{19}$$

where $N_k$ stands for the neighbourhood of neuron $k$, and $| N_k |$ represents the number of neurons in that neighbourhood. It then superimposes distance $h(k)$ as a height on the SOM grid, and thereby transforms the latter into a three-dimensional landscape of a multi-dimensional input space. To represent the additional information (i.e. distances), U-matrix augments the SOM grid by inserting an additional neuron between each pair of neurons. The distances are depicted on the SOM grid as shades of grey; dark and light areas indicate large and small distances respectively, and denote cluster boundaries and clusters in that order.

U-matrix may not be sufficient to delineate clusters that are either overlapping or not well-separated, because it does not take into account the density of input data. To

overcome this limitation, U*-matrix combines both distances and densities in order to improve the visualization and clustering of input data (Ultsch, 2003*a*,*c*). It uses Pareto Density Estimation (i.e. PDE) to estimate the local density of input data as follows (Ultsch, 2003*b*):

$$p(k) = \{\vec{x}_i \in \mathbb{R}^{\mathsf{d}} \,|\, d(\vec{x}_i, \vec{m}_k) < r\} \quad , \tag{20}$$

where $p(k)$ denotes the density of input data in the vicinity of codevector $\vec{m}_k$, and $r$ denotes the Pareto radius of a hypersphere. If $r$ is kept constant, then the number of data points that are included in a hypersphere is proportionate to the underlying density of the data.

### 5.3.2 U-matrix refined

To refine the resolution of U-matrix, Kaski et al. (2003) proposed the following modification:

$$G_{ki} = \| (\vec{m}_k - \vec{m}_i) - (\vec{c}_k - \vec{c}_i) \| \quad . \tag{21}$$

The first term calculates the distance between codevectors $\vec{m}_k$ and $\vec{m}_i$; the second term is estimated from the input data, where $\vec{c}_k = \frac{1}{|N_k|} \sum_{\vec{x} \in N_k} \vec{x}$ denotes the centroid of the input data that belong in the neighbourhood of neuron $k$. The magnitude of $G_{ki}$ is inversely proportional to the density of the corresponding area, the lower the density the larger the magnitude of $G_{ki}$ would be, and vice versa. In a manner similar to U-matrix, the values of $G_{ki}$ can be converted to colour by using index-colour coding, and subsequently be depicted on the SOM grid in order to delineate inter-cluster boundaries. Experimental results have suggested that this visualization method can identify important and meaningful clusters the U-matrix can not identify (Kaski et al., 2003).

### 5.3.3 SOM interpretation

In order to evaluate how well a variable can explain the formation of cluster borders, Kaski et al. (1998) suggested the following metric:

$$\Phi_j = \frac{\| (m_{ij} - m_{kj}) \|}{\| (\vec{m}_i - \vec{m}_k) \|} \quad , \tag{22}$$

where the denominator represents the distance between codevectors $\vec{m}_i$ and $\vec{m}_k$, $m_{ij}$ and $m_{kj}$ denote the $j^{th}$ variable of $\vec{m}_i$ and $\vec{m}_k$, respectively. A large $\Phi_j$ value means that variable $j$ explains well the formation of the cluster border, and data points that belong on either side of the border differ predominantly in the value of variable $j$.

Two further measures assess the relative importance a variable possesses within, and between clusters (Siponen et al., 2001).

$$S_{ij} = \frac{mean_{ij} - min_j}{max_j - min_j} \quad . \tag{23}$$

The term $S_{ij}$ measures the weight variable $j$ has within cluster $i$ relative to that variable's range, $mean_{ij}$ is the mean value of variable $j$ in cluster $i$, $min_j$ and $max_j$ denote the minimum and maximum values of variable $j$, respectively.

The quantity $Q_{ij}$ measures the weight of variable $j$ in cluster $i$ with respect to the values of that variable in clusters other than $i$:

$$Q_{ij} = \frac{S_{ij}}{\frac{1}{C-1}\sum_{k \neq i} S_{kj}} \quad , \tag{24}$$

where $C$ represents the number of clusters.

## 5.4 SOM for visualizing high-dimensional data

As we elaborated in Section 3, SOM preserves the topology or neighbourhood relations between input data as faithfully as possible. However, SOM does not perform a point-to-point mapping, and hence it can not reproduce the pairwise distances between input data. Thus, the distances between neurons on the SOM grid are not proportionate to their corresponding distances in the input space.

ViSOM (Yin, 2002) modifies SOM's updating rule, described in Eq. 6, so that SOM could preserve not only the topology but also the pairwise distances between the input data. It decomposes the term $F_{ik} = \| \vec{x}_i(t) - \vec{m}_k(t) \|$ into two parts: (i) an expansion force, $F_{ic} = \| \vec{x}_i(t) - \vec{m}_c(t) \|$, where $\vec{m}_c(t)$ stands for the BMU codevector of input vector $\vec{x}_i(t)$, and (ii) a lateral force $F_{ck} = \| \vec{m}_c(t) - \vec{m}_k(t) \|$. ViSOM's updating rule is defined as:

$$\vec{m}_k(t+1) = \vec{m}_k(t) + a(t) h_{ck}(t) [\| \vec{x}_i(t) - \vec{m}_c(t) \| + \| \vec{m}_c(t) - \vec{m}_k(t) \| (\frac{d_{ck}}{\lambda \Delta_{ck}} - 1)] \, \forall k \in N_c \quad , \tag{25}$$

where $N_c$ is the neighbourhood of neuron $c$, $k$ denotes the neurons that belong in $N_c$, $d_{ck}$ represents the distance between codevectors $\vec{m}_c$ and $\vec{m}_k$, $\Delta_{ck}$ represents the distance between their location vectors, $\vec{p}_c$ and $\vec{p}_k$, on the SOM grid, and $\lambda$ is a resolution parameter. The aim of Eq. 25 is to adjust inter-neuron distances on the SOM grid, (i.e. $\| \vec{p}_c - \vec{p}_k \|$), so that they could be proportionate to their corresponding distances in the input space, (i.e. $\| \vec{m}_c - \vec{m}_k \|$).

# 6 Financial applications of SOM

To demonstrate how a researcher can apply SOM in order to understand complex data sets, Kaski and Kohonen (1996) applied SOM to 39 statistical indicators that described aspects of the welfare of a number of countries. Based on the 39 statistical indicators, the resulting SOM grid revealed a clustering of countries as well as maintained the similarity between one country and another. For example, OECD and most African countries were clustered on opposite corners of the map indicating extreme welfare and poverty, respectively.

SOM has been applied in predicting the event of bankruptcy as well as in identifying those financial characteristics that are positively correlated with bankruptcy. In order to gain insight into the Spanish banking crisis of 1977-1985, Martín-del-Brío and Serrano-Cinca (1993) selected a sample of 37 solvent banks and 29 bankrupt banks. For each bank, they calculated the following nine ratios: (i) Current Assets/Total Assets, (ii) (Current Assets - Cash and Banks)/Total Assets, (iii) Current Assets/Loans, (iv) Reserves/Loans, (v) Net Income/Total Assets, (vi) Net Income/Total Equity Capital, (vii) Net Income/Loans, (viii) Cost of Sales/Sales, and (ix) Cash-Flow/Loans. Based on these ratios, the study developed a SOM that was able to cluster solvent banks separately from their bankrupt counterparts, and identify how well each ratio could discriminate between the two sets of banks. As expected, small profitability and large debts were positively correlated with the event of bankruptcy. A closer examination revealed that SOM grouped the seven biggest banks as a sub-cluster of the solvent banks, although no information about the sizes of the banks was available.

In addition, Serrano-Cinca (1996) selected 64 solvent, and 65 bankrupt companies from the Moody's Industrial Manual as well as five ratios that described the financial performance of the companies from 1975 to 1985. The ratios were: (i) Working Capital/Total Assets, (ii) Retained Earnings/Total Assets, (iii) Earnings before Interest and Tax/Total Assets, (iv) Market value of Equity, and (v) Sales/Total Assets. SOM clustered the two sets of companies into separate clusters, and also revealed that solvent and bankrupt companies were characterised by high and low levels of earnings, respectively.

Further, to predict the event of bankruptcy, Kiviluoto (1998) applied SOM in a semi-supervised manner in that the input vector, and by extension the codevector, consisted of two parts. The first part contained four financial indicators: (i) Operating Margin, (ii) Net Income before Depreciation and Extraordinary items, (iii) the same as (ii) for the previous year, and (iv) Equity ratio. The second part was a binary indicator, $\{1,0\}$, denoting bankrupt and solvent companies, respectively. The first part was used only for finding the BMU, whereas the whole codevector got updated. The results suggested that companies having low profitability were more likely to go bankrupt.

To investigate the potential of SOM in performing financial benchmarking, Back et al. (1996, 1997) applied SOM to a set of financial ratios that described the performance of a number of international companies operating in the pulp and paper industry from 1985 to 1989. The results suggested that SOM could cluster companies according to their financial similarities, and identify regularities and patterns that were present in the data. This line of research was applied by Karlsson et al. (2001) in order to analyse the quarterly financial performance of a number of international companies that operated in the telecommunication industry.

A similar methodology was adopted by Eklund et al. (2003). They first developed two sets of SOM: the first was based on financial ratios of international companies operating in the pulp and paper industry, and the second was based on the financial ratios having been averaged on a country by country basis. They then compared the two sets in order to identify which financial characteristics were specific to a country. For example, the results revealed the propensity of Finnish companies to finance their activities via debt as well as the economic downturn the Japanese companies experienced in 1997 as a result of the Asian financial crisis. Lansiluoto et al. (2004) extended the foregoing body of literature by using SOM to perform financial benchmarking at both company and industry levels; they did so in order to find out how changes at the industry level (e.g. labour and material costs) might affect the financial performance of companies.

SOM was used by (Baghai-Wadj et al., 2005) to derive a taxonomy of hedge funds based on their monthly returns in order to address the following three issues: (i) whether hedge funds change their self-declared styles over time, (ii) whether self-declared styles are useful or misleading, and (iii) which types of hedge funds are prone to misclassify their own styles. The study selected the monthly returns of 2,442 hedge funds from the Centre for International Securities and Derivatives Markets (i.e. CISDM). The monthly returns covered the period from April 1995 to April 2004, and the sample included both active and inactive hedge funds. SOM was able to group hedge funds into nine clusters that were labelled as follows: (i) convertible arbitrage and fixed income, (ii) emerging markets, (iii) futures, (iv) merger arbitrage and distressed securities, (v) sector financial, (vi) sector health care, (vii) sector technology, (viii) short-selling, and (ix) other hedge funds. The study found that 23% of the examined hedge funds changed their self-declared style over the period under investigation, and that hedge funds having high consistency in their self-declared styles were less prone to change their styles over time.

Further financial applications of SOM include: selecting shares suitable for portfolio management (Deboeck and Ultsch, 2000), and (Khan et al., 2009); assessing the creditworthiness of loan applicants (Tan et al., 2002), and (Huysmans et al., 2006); and performing market segmentation as part of customer relationship management (Hung and Tsai, 2008).

# 7 Case study: Clustering accounting databases

The case study illustrates how SOM can be applied in clustering accounting databases, and discusses potential applications such clustering may have in the discipline of accounting. The motivation behind the case study is twofold. First, existing literature has paid insufficient attention to the clustering of accounting databases as an internal control procedure; and second, the importance and relevance of this issue are underlined by a number of statutory and professional pronouncements that have proliferated in the wake

of several cases of financial malfeasance (e.g. Enron, WorldCom).

## 7.1 Data description

The data were provided by an international shipping company as a text-dump of its accounting database, and consisted of 16,300 journal entries covering fiscal year 2007. Journal entries implement the system of double-entry bookkeeping by recording accounting transactions in a "Debit" and "Credit" format; it entails that an accounting transaction is recorded twice: as a "Debit" and as an equivalent "Credit", hence the name "double-entry". This system of recording accounting transactions serves to preserve the accounting equality: $Assets = Liabilities + Equity$. To be considered as an accounting transaction and thus be recorded, an economic event must cause either an increase or a decrease in a company's assets, liabilities, or equity.

For example, Table 1 depicts the aforesaid company's journal entry that consists of two accounting transactions, and records the payment of salaries for August 2007. Each accounting transaction is described by the following seven variables: (i) Account Number (hierarchical), (ii) Account Description (text), (iii) Date (date), (iv) Debit-Credit (binary), (v) US$ Amount (numerical), (vi) Details (text), and (vii) Account Class (hierarchical). The types of variables are shown in parentheses.

Table 1: Journal Entry

| Account Number | Account Description | Date | Debit-Credit | US$ Amount | Details | Account Class |
|---|---|---|---|---|---|---|
| 60000 | Salaries | 27/Aug/2007 | Debit | 56,070 | Payment of salaries | Expenses |
| 30000 | Bank | 27/Aug/2007 | Credit | 56,070 | Payment of salaries | Cash & Cash Equivalents |

To carry out the case study, we select three variables: (i) Debit-Credit, (ii) US$ Amount, and (iii) Account Class. We incorporate "Debit-Credit" into "US$ Amount" by expressing the latter as positive and negative amounts denoting "Debit" and "Credit" balances, respectively.

## 7.2 Data pre-processing

As we discussed in Section 2.2, we must define a valid similarity metric over "Account Class" for SOM to yield meaningful results; "Account Class" is a hierarchical variable and as such it lacks intrinsic quantitative information, and hence SOM can not calculate the Euclidean distance. To this end, we opt for the graph-theoretical approach, described in Section 5.1.4, in order to transform the values of "Account Class" into a numerical representation, and thus render them amenable for SOM-based processing.

To elaborate, we observe that the values of "Account Class" follow a hierarchy, the IASB XBRL Taxonomy (IASCF, 2009); it describes accounting concepts and their semantical

relationships as *child* $\prec$ *parent* links according to the International Financial Reporting Standards (IASB, 2009). Based on the aforementioned taxonomy, we first encode "Account Class" as a directed acyclic graph (i.e. DAG), whereby the root vertex represents the complete set of "Account Class", and all other vertices are ordered in such a way that each vertex represents a subset of its parent vertex. Consequently, the DAG can preserve the *child* $\prec$ *parent* relationships specified in the taxonomy. For example, "Salaries"$\prec$ "Expenses" $\prec$ "Income Statement" forms a path on the DAG, and serves to induce the corresponding *child* $\prec$ *parent* relationships.

Second, we calculate all pairwise distances between the vertices by using Dijkstra's algorithm (Dijkstra, 1959). This operation yields distance matrix $X = (\vec{x}_{ij})_{nd}$, where $d = 30$ denotes the dimensionality, and $n = 36,510$ represents the number of accounting transactions. The distance matrix can form the input dataset to SOM, as it satisfies the conditions of a metric space specified in Section 2.2. There are more accounting transactions than journal entries, because a journal entry can be composed of any number of accounting transactions provided that the "Debit" and "Credit" balances are always equal.

## 7.3 Experiments

To illustrate how SOM can be applied in clustering accounting databases, the paper adopts a five-step approach. First, it uses bootstrap to draw two hundred random samples with replacement from the empirical distribution of the input dataset, $X = (\vec{x}_{ij})_{nd}$. Second, for each bootstrapped sample, it uses SOM-Toolbox for Matlab[4] (Vesanto et al., 2000) to train a SOM in batch mode, Eq. 7, with hexagonal topology, and Gaussian neighbourhood, Eq. 5. Third, to identify clusters on the SOM grid, it uses U-matrix, described in Section 5.3.1. Fourth, to evaluate the performance of SOM, it calculates the quantization and topographic errors, and to assess the internal validity of clustering, it calculates the Davies-Bouldin index; these metrics were explained in Section 4. Finally, the paper estimates the two-sided 95% confidence intervals of the mean and standard deviation of the foregoing metrics.

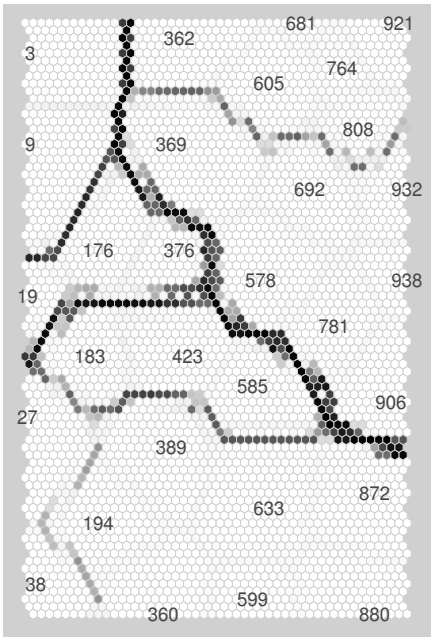
## 7.4 Results presentation and discussion

The mean and standard deviation of the topographic error have small values and narrow 95% confidence intervals: $CI_{0.95} = 0.3336 - 0.3984$ and $CI_{0.95} = 0.1478 - 0.1880$, respectively. The corresponding values for the quantization error are zero to the $4^{th}$ decimal point. These results suggest that the SOM grid can map the input data accurately as well as preserve their topology faithfully.

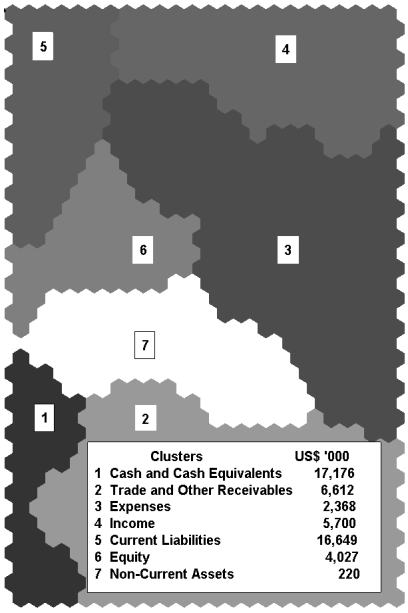
[4]SOM-Toolbox for Matlab and its documentation are available at:
http://www.cis.hut.fi/somtoolbox

The results of the case study are depicted in the form of two SOM grids in Fig. 1. A visual inspection of Fig. 1a reveals how the U-matrix can identify seven clusters. As we discussed in Section 5.3.1, dark and light areas indicate long and short distances, respectively; the former delineates cluster borders, whereas the latter represents clusters. The internal validity of clustering is evaluated by statistical analysis; both the mean and standard deviation of Davies-Bouldin index have small values and narrow 95% confidence intervals: $CI_{0.95} = 1.4905 - 1.5988$, and $CI_{0.95} = 0.2445 - 0.3563$, in this order. The statistical results imply that the seven clusters thus identified are highly -compact, and their centroids are well-separated.

More importantly, the seven clusters can be interpreted within an accounting context, as shown in Fig. 1b. In particular, clusters 3 and 4 represent expense and income items respectively, and the Income Statement collectively. Further, clusters 7, 2, and 1 correspond to "Non-Current Assets", "Trade and Other Receivables", and "Cash and Cash Equivalents", in that order. These three clusters make up the "Assets" side of the Balance Sheet Statement. In addition, clusters 6 and 5 stand for "Equity" and "Current Liabilities" respectively, and their total forms the "Equity and Liabilities" side of the Balance Sheet Statement. Finally, for each cluster, Fig. 1b conveys the total US$ amount in thousands for fiscal year 2007. This piece of information may serve useful purposes, as it provides a user with the aggregated US$ amounts for each category of the financial statements. Indeed, the SOM grid, Fig. 1b, preserves the accounting equality: $Assets(220 + 6,612 + 17,176) = Liabilities(16,649) + Equity(4,027 + 3,332)$. The amount of US$ 3,332 (i.e. 5,700 -2,368) represents the profit for the fiscal year 2007.




(a) SOM grid: the numbers denote those neurons that are BMU, and the U-matrix identifies seven clusters. The dark areas represent the borders of the clusters.




(b) SOM grid: it labels the seven clusters, and also conveys their respective total US$ amounts for fiscal year 2007.

Figure 1: The SOM grids.

## REFERENCES

Argyrou, A. (2009), Clustering hierarchical data using self-organizing map: A graph-theoretical approach, *in* J. Príncipe and R. Miikkulainen, eds, 'Advances in Self-Organizing Maps', Vol. 5629 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin / Heidelberg, pp. 19–27.

Back, B., Sere, K. and Vanharanta, H. (1996), Data mining accounting numbers using self-organizing maps, *in* J. Alander, T. Honkela and M. Jakobsson, eds, 'Genes, Nets and Symbols (STeP '96). Finnish Artificial Intelligence Society', University of Vaasa, Vaasa, Finland, pp. 35–47.

Back, B., Sere, K. and Vanharanta, H. (1997), Analyzing financial performance with self-organizing maps, *in* 'Proceedings of the First International Workshop on Self-Organizing Maps (WSOM'97)', Espoo, Finland, pp. 356–361.

Baghai-Wadj, R., El-Berry, R., Klocker, S. and Schwaiger, M. (2005), 'The Consistency of Self-Declared Hedge Fund Styles - A Return-Based Analysis with Self-Organizing Maps', *Central Bank of Austria: Financial Stability Report* (9), 64–76.

Bishop, C. M. (1995), *Neural networks for pattern recognition*, Oxford University Press, Oxford, United Kingdom.

Chappell, G. J. and Taylor, J. G. (1993), 'The temporal kohonen map', *Neural Networks* **6**(3), 441–445.

Davies, D. and Bouldin, D. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2), 224–227.

Deboeck, G. and Ultsch, A. (2000), Picking stocks with emergent Self-Organizing value maps, *in* 'Neural Networks World', Vol. 10 of *1-2*, pp. 203–216.

Deerwester, S., Dumais, S., Furnas, G., Thomas and Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science* **41**, 391–407.

Dijkstra, E. W. (1959), 'A note on two problems in connexion with graphs', *Numerische Mathematik* **1**, 269–271.

Eklund, T., Back, B., Vanharanta, H. and Visa, A. (2003), Financial benchmarking using self-organizing maps studying the international pulp and paper industry, *in* 'Data mining: opportunities and challenges', IGI Publishing, pp. 323–349.

Erwin, E., Obermayer, K. and Schulten, K. (1992), 'Self-Organizing maps: Ordering, convergence properties and energy functions', *Biological Cybernetics* **67**, 47–55.

Graepel, T., Burger, M. and Obermayer, K. (1997), 'Phase transitions in stochastic self-organizing maps', *Physical Review E* **56**(4), 3876–3890.

Haykin, S. (1999), *Neural Networks. A Comprehensive Foundation*, second edn, Prentice Hall International, Upper Saddle River, New Jersey, USA.

Honkela, T., Kaski, S., Lagus, K. and Kohonen, T. (1996*a*), Exploration of full-text databases with self-organizing maps, *in* 'Proceedings of the International Conference on Neural Networks (ICNN '96)', Vol. I, IEEE Service Center, Piscataway, NJ, USA, pp. 56–61.

Honkela, T., Kaski, S., Lagus, K. and Kohonen, T. (1996*b*), Newsgroup exploration with WEBSOM method and browsing interface, Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Hsu, C. (2006), 'Generalizing self-organizing map for categorical data', *IEEE Transactions on Neural Networks* **17**(2), 294–304.

Hung, C. and Tsai, C. (2008), 'Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand', *Expert Systems with Applications* **34**(1), 780–787.

Huysmans, J., Baesens, B., Vanthienen, J. and van Gestel, T. (2006), 'Failure prediction with self organizing maps', *Expert Systems with Applications* **30**(3), 479–487.

IASB (2009), *International Financial Reporting Standards (IFRS) 2009*, International Accounting Standards Committee Foundation (IASCF), London,United Kingdom.

IASCF (2009), *IFRS Taxonomy Guide 2009 (XBRL)*, International Accounting Standards Committee Foundation (IASCF), London,United Kingdom.

Jungnickel, D. (2002), *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Volume 5, English edn, Springer, Berlin, Germany.

Karlsson, J., Back, B., Vanharanta, H. and Visa, A. (2001), Analysing financial performance with quarterly data using Self-Organising Maps, TUCS Technical Report No 430, Turku Centre for Computer Science, Turku, Finland. Available at: http://tucs.fi/publications/attachment.php?fname=TR430.pdf.

Kaski, S. (1998), Dimensionality reduction by random mapping: Fast similarity computation for clustering, *in* 'Proceedings of the International Joint Conference on Neural Networks (IJCNN '98)', Vol. 1, IEEE Service Center, Piscataway, NJ, USA, pp. 413–418.

Kaski, S. and Kohonen, T. (1996), Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world, *in* A. N. Refenes, Y. Abu-Mostafa, J. Moody and A. Weigend, eds, 'Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, London, England, 11 - 13 October, 1995', World Scientific, Singapore, pp. 498–507.

Kaski, S., Nikkilä, J. and Kohonen, T. (1998), Methods for interpreting a Self-Organized map in data analysis, *in* 'Proceedings of ESANN'98, 6th European Symposium on Artificial Neural Networks', D-Facto, Brussels, Belgium, pp. 185–190.

Kaski, S., Nikkilä, J. and Kohonen, T. (2003), Methods for exploratory cluster analysis, *in* 'Intelligent exploration of the web', Studies In Fuzziness And Soft Computing, Physica-Verlag GmbH, Heidelberg, Germany, pp. 136–151.

Khan, A. U., Bandopadhyaya, T. K. and Sharma, S. (2009), 'Classification of stocks using self organizing map', *International Journal of Soft Computing Applications* (4), 19–24.

Kiviluoto, K. (1996), Topology preservation in self-organizing maps, *in* 'Proceeding of the International Conference on Neural Networks (ICNN '96)', Vol. 1, IEEE Service Center, Piscataway, NJ, USA, pp. 294–299.

Kiviluoto, K. (1998), 'Predicting bankruptcies with the self-organizing map', *Neurocomputing* **21**(1-3), 191–201.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**(1), 59–69.

Kohonen, T. (1985), 'Median strings', *Pattern Recognition Letters* **3**(5), 309 – 313.

Kohonen, T. (1996), Self-Organizing Maps of symbol strings, Technical Report A42, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kohonen, T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences,Volume 30, second edn, Springer-Verlag, Heidelberg, Germany.

Kohonen, T. (1998), 'The self-organizing map', *Neurocomputing* **21**(1-3), 1–6.

Kohonen, T. (1999), 'Comparison of SOM point densities based on different criteria', *Neural Computation* **11**(8), 2081–2095.

Kohonen, T. (2008), Data management by Self-Organizing Maps, *in* 'Computational Intelligence: Research Frontiers', Vol. 5050/2008 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 309–332.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000), 'Self organization of a massive document collection', *IEEE Transactions on Neural Networks* **11**(3), 574–585.

Kohonen, T. and Somervuo, P. (1997), Self-organizing maps of symbol strings with application to speech recognition, *in* 'Proceedings of the First International Workshop on Self-Organizing Maps (WSOM'97)', Espoo, Finland, pp. 2–7.

Kohonen, T. and Somervuo, P. (1998), 'Self-organizing maps of symbol strings', *Neurocomputing* **21**(1-3), 19–30.

Kohonen, T. and Somervuo, P. (2002), 'How to make large self-organizing maps for nonvectorial data', *Neural Networks* **15**(8-9), 945–952.

Koskela, T., Varsta, M., Heikkonen, J. and Kaski, K. (1998), Temporal sequence processing using recurrent SOM, *in* 'Proceedings of the Second International Conference on Knowledge-Based Intelligent Electronic Systems, (KES '98)', Vol. 1, pp. 290–297.

Lagus, K. (1997), Map of WSOM'97 Abstracts Alternative Index, *in* 'Proceedings of the First International Workshop on Self-Organizing Maps (WSOM'97)', Espoo, Finland, pp. 368–372.

Lagus, K., Kaski, S. and Kohonen, T. (2004), 'Mining massive document collections by the WEBSOM method', *Information Sciences* **163**(1-3), 135–156.

Lansiluoto, A., Eklund, T., Back, B., Vanharanta, H. and Visa, A. (2004), 'Industry-specific cycles and companies' financial performance comparison using self-organizing maps', *Benchmarking: An International Journal* **11**(3), 267–286.

Lawrence, R., Almasi, G. and Rushmeier, H. (1999), 'A scalable parallel algorithm for Self-Organizing Maps with applications to sparse data mining problems', *Data Mining and Knowledge Discovery* **3**(2), 171–195.

Manning, C. D., Raghavan, P. and Shutze, H. (2009), *An Introduction to Information Retrieval*, Online edn, Cambridge University Press, New York, USA. Available at: http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

Martín-del-Brío, B. and Serrano-Cinca, C. (1993), 'Self-organizing neural networks for the analysis and representation of data: Some financial cases', *Neural Computing & Applications* **1**(3), 193–206.

Oja, M., Sperber, G., Blomberg, J. and Kaski, S. (2005), 'Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups', *International Journal of Neural Systems* **15**(3), 163–179.

Ritter, H. (1991), 'Asymptotic level density for a class of vector quantization processes', *IEEE Transactions on Neural Networks* **2**(1), 173–175.

Ritter, H. and Kohonen, T. (1989), 'Self-Organizing Semantic Maps', *Biological Cybernetics* **61**(4), 241–254.

Salton, G., Wong, A. and Yang, C. S. (1975), 'A vector space model for automatic indexing', *Communications of ACM* **18**(11), 613–620.

Serrano-Cinca, C. (1996), 'Self organizing neural networks for financial diagnosis', *Decision Support Systems* **17**(3), 227–238.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**, 379–423 & 623–656.

Siponen, M., Vesanto, J., Simula, O. and Vasara, P. (2001), An approach to automated interpretation of SOM, *in* N. Allinson, H. Yin, L. Allinson and J. Slack, eds, 'In proceedings of Workshop on Self-Organizing Map (WSOM 2001)', Springer, pp. 89–94.

Stevens, S. S. (1946), 'On the theory of scales of measurement', *Science* **103**(2684), 677–680.

Tan, R., van den Berg, J. and van den Bergh, W. (2002), Credit rating classification using Self-Organizing Maps, *in* 'Neural Networks in Business: Techniques and Applications', Idea Group Publishing, pp. 140–153.

Ultsch, A. (2003*a*), Maps for the visualization of high-dimensional data spaces, *in* 'Proceedings Workshop on Self-Organizing Maps (WSOM 2003)', Hibikino, Kitakyushu, Japan, pp. 225–230.

Ultsch, A. (2003*b*), Pareto density estimation: A density estimation for knowledge discovery, *in* 'Innovations in Classification, Data Science, and Information Systems - Proceedings of 27th Annual Conference of the German Classification Society (GfKL 2003)', Springer, pp. 91–100.

Ultsch, A. (2003*c*), U*-Matrix: a tool to visualize clusters in high dimensional data, Technical Report 36, Department of Mathematics and Computer Science, University of Marburg, Germany.

Ultsch, A. and Siemon, H. (1990), Kohonen's self organizing feature maps for exploratory data analysis, *in* 'Proceedings International Neural Networks', Kluwer Academic Press, pp. 305–308.

Varsta, M., Heikkonen, J., Lampinen, J. and Millán, J. D. (2001), 'Temporal kohonen map and the recurrent self-organizing map: analytical and experimental comparison', *Neural Processing Letters* **13**(3), 237–251.

Vesanto, J. and Alhoniemi, E. (2000), 'Clustering of the self-organizing map', *IEEE Transactions on Neural Networks* **11**(3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), SOM Toolbox for Matlab 5, Technical Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland. Available at: http://www.cis.hut.fi/somtoolbox.

Vesanto, J., Sulkava, M. and Hollmén, J. (2003), On the decomposition of the Self-Organizing Map distortion measure, *in* 'Proceedings of the Workshop on Self-Organizing Maps (WSOM 2003)', Hibikino, Kitakyushu, Japan, pp. 11–16.

Voegtlin, T. (2002), 'Recursive self-organizing maps', *Neural Networks* **15**(8-9), 979–991.

Yin, H. (2002), 'Data visualisation and manifold mapping using the ViSOM', *Neural Networks* **15**(8-9), 1005–1016.

Yin, H. (2008), The Self-Organizing Maps: Background, theories, extensions and applications, *in* J. Fulcher and L. C. Jain, eds, 'Computational Intelligence: A Compendium', Vol. 115 of *Studies in Computational Intelligence (SCI)*, Springer-Verlag Berlin / Heidelberg, pp. 715–762.

**Auditing journal entries using Self-Organizing Map**

**Argyris Argyrou**

**Abstract**
A considerable body of regulatory pronouncements attests to the significance of auditing journal entries for ensuring that financial statements are free of material misstatements; however, existing empirical studies have paid insufficient attention to the auditing of journal entries. To explore this issue further, this paper proposes a model based on self-organizing map as well as validates this model by performing experiments on a dataset containing journal entries. Empirical results suggest that the proposed model can detect "suspicious" and legitimate transactions with a high degree of accuracy. Further investigations reveal that the performance of the model is robust to varying prior probabilities of "suspicious" journal entries occurring in the population. The findings indicate that the model can assist auditors to detect "suspicious" journal entries.

**Keywords:** Auditing, journal entries, self-organizing map

# 1 Introduction

Auditing journal entries has gained prominence as a result of statute and auditing standards that have been promulgated in the wake of financial malfeasance (e.g. Enron, WorldCom). In essence, Sarbanes and Oxley Act (U.S. Congress, 2002) Sections 302 and 404 require the management of companies to establish and maintain internal controls as well as to assess the effectiveness of these controls at the end of a fiscal year. Auditing Standard No.5 (Public Company Accounting Oversight Board (PCAOB), 2007) requires external auditors to express an opinion on the effectiveness of a company's internal controls. Further, auditors are required to test the appropriateness of journal entries recorded in the general ledger and other adjustments made in preparing financial statements in order to obtain reasonable assurance about whether the financial statements are free of material misstatements, whether caused by errors or fraud (PCAOB, 2002). In addition, the Standing Advisory Group (PCAOB, 2004) considers "...significant or unusual journal entries ..." to be an area that poses a high risk of material misstatement due to fraud. Auditors may be exposed to litigation (Palmrose, 1991) and subject to regulatory enforcement releases (Bonner et al., 1998), if they fail to detect material misstatements to financial statements.

However, absent a study by Debreceny and Gray (2010), existing literature has paid little attention to the auditing of journal entries. Indeed, a review of the pertinent literature acknowledges the paucity of studies that examine the link between reviewing journal entries and detecting financial statement fraud (Hogan et al., 2008).

To investigate this unexplored issue further, this paper proposes and then validates a model in order to assist auditors to detect "suspicious" journal entries. To this end, the paper adopts an one-class classification approach (Chandola et al., 2009; Juszczak et al., 2008),

because, in an audit engagement, a dataset does not contain journal entries that are pre-labelled as "suspicious" nor as legitimate. As a result, detecting "suspicious" journal entries can be achieved only with reference to information describing legitimate journal entries. The implicit assumption is the vast majority of journal entries are legitimate and generated by a probability distribution that is different form that generating "suspicious" journal entries.

Given this assumption, the paper first employs self-organizing map (SOM; Kohonen, 1997), an unsupervised-learning algorithm, to derive a reference model that can describe the behaviour of legitimate journal entries. Subsequently, the paper calculates the quantization errors, a distance metric, between journal entries and the SOM-based reference model. Finally, the paper considers a journal entry to be "suspicious", if the journal entry has a quantization error that exceeds an optimum threshold.

To validate the proposed SOM-based model, the paper conducts a series of experiments on a dataset that has been made available by an international shipping company; the dataset contains the complete set of their journal entries for fiscal year 2006. The paper incorporates certain characteristics of an audit engagement in order to carry out experiments that are as realistic as possible. First, it simulates four categories of potential errors that may affect a journal entry; the number of categories is restricted by the number of variables describing a journal entry. Second, it considers three prior probabilities, or prevalence, of "suspicious" journal entries occurring in the population, because the number of "suspicious" journal entries is likely to be orders of magnitude smaller than that of legitimate. Third, it examines four cost-ratios of Type I to Type II errors. The rationale is the cost of a Type II error, identifying a "suspicious" journal entry as legitimate, tends to be much higher than that of a Type I error, identifying a legitimate journal entry as "suspicious". Fourth, the paper initialises a range of thresholds and selects that which minimises the total misclassification cost to be the optimum threshold.

In the following section, the paper reviews related studies; subsequently, it describes the data, introduces SOM and the reference model as well as elaborates on the simulation of "suspicious" journal entries and the experiments. Further, the paper presents and discusses the results, provides conclusions, and suggests directions for additional research.

# 2 Background and related studies

The Statement On Auditing Standards 99 (SAS 99): Consideration of Fraud in a Financial Statement Audit (AICPA, 2002), which is the interim standard AU Section 316 of PCAOB, requires an auditor to obtain reasonable assurance that financial statements are free of material misstatements, whether caused by error or fraud. Further, AU

Section 316 0.58 requires an auditor, among other things, to test the appropriateness of journal entries recorded in a general ledger as well as other adjustments made in preparing financial statements. The reason is management can misstate financial statements by recording fictitious and inappropriate journal entries, especially towards the end of a fiscal year (AU Section 316 .08, and .58). Financial statement fraud entails considerable legal and economic costs for companies, directors, and shareholders (COSO, 2010) as well as may expose auditors to litigation (Palmrose, 1991).

A review of the literature pertinent to financial statement fraud acknowledges the lack of empirical studies that address the relationship between unusual or top-level journal entries and financial statement fraud (Hogan et al., 2008). A noteworthy exception is a study by Debreceny and Gray (2010) who applied digit analysis, or Benford's Law, in order to detect fraudulent journal entries. In brief, the paper compared the observed distribution of the first digit of US$ amounts against that expected by Benford's Law; if the difference is statistically significant under a chi-square test, then the US$ amount is deemed to be suspicious. The results have suggested that, for all entities in the sample, the observed distribution of the first digit of US$ amounts were significantly different form that expected by Benford's Law. However, the results may have been an artefact of the chi-square test, as a large number of observations can induce statistically significant results (Grabski, 2010). A further explanation is that either fraudulent journal entries were the norm in the sample, or Benford's Law is not applicable to journal entries (Grabski, 2010).

Self-organizing map has been applied to detect "anomalous" behaviour in a number of diverse domains where information describing this type of behaviour does not exist in advance, or is rare, or expensive to collect. Consequently, detecting "anomalous" behaviour can be achieved only with reference to the information describing "legitimate" behaviour. For example, SOM has been used to detect fraudulent insurance claims (Brockett et al., 1998), to develop risk-adjustment models for estimating medical expenditures (Hsu et al., 2008), and detect fraudulent credit-card transactions (Juszczak et al., 2008).

Further, Ypma and Duin (1997) detected faults in rotating mechanical machinery and leaks in pipelines by using a variant of the SOM-based metric proposed by Kaski and Lagus (1996). SOM has been used as an intrusion detection tool for detecting "anomalous" network traffic (e.g. TCP/IP packets) that may signal a possible Denial of Service attack (Labib and Vemuri, 2002), detecting instances of buffer-overflow attacks based on a model of legitimate network traffic (Rhodes et al., 2000), and identifying "anomalous" behaviour that may exist in network services (e.g. web, e-mail, telnet) (Ramadas et al., 2003).

# 3 Research design and methodology

## 3.1 Data description

The raw data have been provided by an international shipping company in the form of a text file containing the complete set of their journal entries, $n = 6,404$, for fiscal year 2006. The text file consists of 25,422 lines and eight columns, representing accounting transactions and variables, respectively. The number of accounting transactions is greater than that of journal entries, because a journal entry can be made up of any number of transactions as long as the "Debit" and "Credit" sides are equal. The eight variables are as follows: (i) "Account Number" (alphanumerical), (ii) "Account Name" (text), (iii) "Posting Date" (date), (iv) "US$ amount" (numerical), (v) "Debit-Credit Indicator" (binary), (vi) "Description" (text), (vii) "Account Class" (hierarchical), and (viii) "Code" (numerical); the types of the variables are shown in parentheses.

Table 1: Descriptive statistics

| | | | | | Interquartile range | |
|---|---|---|---|---|---|---|
| Account Class | Code | N | Mean | MAD(*) | 0.75 | 0.25 |
| Property, Plants and Equipment: Costs | 1000 | 25 | 13,202 | 913 | 982 | -526 |
| Property, Plants and Equipment: Depr. | 1100 | 61 | -3,302 | 386 | -98 | -1,139 |
| Investments in Subsidiaries | 1200 | 3 | 11,750 | 9,800 | 21,413 | 2,550 |
| Cash in Bank | 2110 | 3,744 | 1,800 | 4,913 | 13,983 | -247 |
| Cash at Hand | 2140 | 259 | 13 | 84 | 39 | -113 |
| Sales Taxes Receivable | 2170 | 812 | 69 | 8 | 42 | 3 |
| Trade Debtors | 2200 | 3,989 | -1,725 | 4,037 | 1,725 | -8,812 |
| Other Trade Debtors | 2204 | 1,048 | 538 | 1,549 | 1,566 | -1,495 |
| Accounts Receivable | 2207 | 309 | 0 | 36 | 7 | -60 |
| Short-term Loans Receivable | 2300 | 14 | 1,582,003 | 110,434 | 2,500,000 | -5,000 |
| Insurance Receivable | 2400 | 4,963 | -86 | 1,869 | 1,843 | -2,000 |
| Other Debtors and Receivables | 2600 | 177 | 1,046 | 979 | 762 | -1,302 |
| Accounts Payable | 3455 | 3,242 | -75 | 296 | 112 | -546 |
| Trade Creditors | 3480 | 1,107 | -4,542 | 1,743 | -135 | -6,165 |
| Other Income Received | 6000 | 19 | -12,671 | 186 | -139 | -430 |
| Fees Received | 6100 | 36 | -13,333 | 15,000 | -1,500 | -28,500 |
| Insurance Commission Received | 6260 | 1,996 | -805 | 332 | -89 | -1,125 |
| Bank Interest Received | 6300 | 344 | -1,307 | 712 | -215 | -2,367 |
| Other Interest Received | 6305 | 8 | -145,875 | 10,238 | -130,668 | -153,653 |
| Exchange Difference: Gain | 6440 | 678 | -788 | 18 | 1 | -42 |
| Administration Expenses | 8500 | 1,236 | 65 | 57 | 114 | -6 |
| Office Expenses | 8501 | 466 | 222 | 176 | 332 | -54 |
| Salaries and Wages | 8550 | 214 | 3,034 | 4,372 | 7,392 | -2,516 |
| Fees and Commissions | 8600 | 121 | 817 | 107 | 592 | 17 |
| Professional Expenses | 8610 | 28 | 3,440 | 583 | 1,757 | -52 |
| Finance Expenses | 8700 | 465 | 158 | 14 | 36 | 10 |
| Depreciation Expenses | 8800 | 58 | 941 | 388 | 1,139 | 165 |
| | | 25,422 | | | | |

(*) Median absolute difference. Amounts are expressed in US$.

The paper uses accounting transactions as the unit of analysis and the following three

variables: (i) "US$ amount", (ii) "Debit-Credit Indicator", and (iii) "Code". The paper aggregates the transactions at the "Account Class" level, $n = 27$, because there are not enough transactions at the "Account Number" level, $n = 360$, for the paper to perform statistical analyses. Descriptive statistics are shown in Table 1.

## 3.2 Self-organizing map

SOM performs two operations: first, it performs vector quantization by representing input vectors with a much smaller, albeit representative, set of codevectors; and second, it carries out a non-linear mapping or projection from a high-dimensional input space to a regular two-dimensional grid of neurons, while preserving the original topology as faithfully as possible. For the purpose of this study, SOM is employed to perform only vector quantization.

In the context of this study, the input dataset to SOM is denoted by $X = (\vec{x}_{ij})_{nd}$, where $n = 25,422$ represents the number of transactions, and $d = 3$ denotes the dimensionality of the dataset. Given this input dataset, SOM constructs a set of codevectors, $M = (\vec{m}_{kj})_{Kd}$, where $d = 3$ as noted above, and $k = 1, 2, 3, \ldots, 810$ denotes the number of codevectors, which is approximately equal to 5 x $\sqrt{25,422}$ (Vesanto et al., 2000, p.30).

SOM is formed in three iterative processes. First, in the competition process, each input vector, $\vec{x}_i \in \mathbb{R}^3$, is compared with all codevectors, $\vec{m}_k \in \mathbb{R}^3$, and the best match in terms of the smallest Euclidean distance, $\| \vec{x}_i - \vec{m}_k \|$, is mapped onto neuron $k$, termed the best-matching unit (i.e. BMU), and denoted by the subscript $c$: $\| \vec{x}_i - \vec{m}_c \| = \min_k \{\| \vec{x}_i - \vec{m}_k \|\}$ (Kohonen, 1997, p.86).

Second, in the co-operation process, the BMU locates the centre of a neighbourhood kernel, $h_{ck}(t)$, which is usually a Gaussian function defined as: $h_{ck}(t) = exp\left[-\frac{\|\vec{p}_c - \vec{p}_k\|^2}{2\sigma^2(t)}\right]$, where $\vec{p}_c$, $\vec{p}_k \in \mathbb{R}^2$ are the location vectors of BMU and neuron $k$ respectively, $t$ denotes discrete time, and $\sigma(t)$ defines the width of the kernel (Kohonen, 1997, p.87).

Third, in the adaptive process, the sequence-training SOM updates recursively codevector $\vec{m}_k$ as follows: $\vec{m}_k(t+1) = \vec{m}_k(t) + a(t)h_{ck}(t)\left[\vec{x}_i(t) - \vec{m}_k(t)\right]$, where $0 < a(t) \leqslant 1$ is a learning rate at discrete time $t$, and $a(t)$ is a non-increasing function of time. Batch-training SOM updates the codevectors only at the end of each epoch, which is a complete presentation of input data, rather than recursively, as follows (Vesanto et al., 2000, p.9): $\vec{m}_k(t+1) = \frac{\sum_{i=1}^{n} h_{ck}(t)\vec{x}_i}{\sum_{i=1}^{n} h_{ck}(t)}$.

Finally, SOM converges to a stable state when the codevectors do not get updated any further; the convergence criterion is $E\{h_{ck}(\vec{x}_i - \lim_{t\to\infty} \vec{m}_k(t))\} = 0$, where $E\{.\}$ denotes the expectation function (Kohonen, 1997, p.113).

## 3.3 Reference model

The stable-state codevectors, $M = (\vec{m}_{kj})_{Kd}$, can function as a reference model of the input data, $X = (\vec{x}_{ij})_{nd}$, because the probability density function of the codevectors approximates that of the input data (Kohonen, 1999, 1997, p.48). The reference model enjoys a number of properties that can enhance its application. First, it does not make any assumptions concerning the probability distribution of the transactions, as the codevectors are essentially non-parametric regressors of the transactions. Second, it is robust to changes to input data, because all the relevant operations (e.g. calculating distances) are performed in the input space; thus, any changes in the input data would propagate corresponding changes to the estimation and update of codevectors. Third, it is derived directly from the data, and hence does not entail encoding domain-expert knowledge, as the case would have been had the paper adopted a model-driven approach. Fourth, the paper can detect transactions deviating from the reference model without having to form expectations about these transactions in advance. Fifth, the computational complexity of the model scales well with large datasets, because the number of codevectors is often chosen to be approximately equal to the square root of the number of input vectors, $K \approx \sqrt{n}$.

The paper evaluates the degree of affinity between the reference model and a transaction by calculating the quantization error: $R_i = \| \vec{x}_i - \vec{m}_c \|^2$; the quantization error represents the Euclidean distance between the $i^{th}$ transaction, $\vec{x}_i$, and the codevector corresponding to its BMU, $\vec{m}_c$. This scoring mechanism is feasible and valid, because quantization error is monotonically related to the degree of "suspiciousness" (Bolton and Hand, 2002). The larger a transaction's quantization error is, the farther away from the reference model the transaction is going to be, and hence the more likely it could be "suspicious". An equivalent interpretation is transactions having quantization errors that exceed an optimum threshold, and hence are considered to be "suspicious", are generated by a probability distribution that is different from that generating legitimate transactions.

## 3.4 Simulating "suspicious" transactions

To represent an audit engagement as faithfully as possible, the paper addresses the following four issues: (i) categories of potential errors that may affect a transaction, (ii) prior probability, or prevalence, of "suspicious" transactions occurring in the population, (iii) asymmetrical misclassification costs of Type I and Type II errors, and (iv) optimum threshold that can distinguish "suspicious" from legitimate transactions.

***Four categories of potential errors***

In the present case, a transaction is described by three variables and can be considered to

be "suspicious" if either of these variables contains an error, given the rest do not contain any errors, or all of the variables contain an error. This combination yields four categories of potential errors that may affect a transaction. The paper assumes that the four categories of potential errors occur with an equal probability.

In order to simulate errors in "US$ amount", the paper adds a noise to this variable. The noise is equal to the average of median-absolute-difference of an "Account Class" that is selected randomly excluding the "Account Class" of the transactions whose "US$ amount" are to be modified. Second, to simulate errors in "Debit - Credit Indicator", the paper reverses the binary indicator, {1,0}, thereby changing "Debit", denoted by 1, to "Credit", denoted by 0, and vice versa. This operation is equivalent to multiplying the "US$ amount" by $(-1)$, thus converting positive amounts, Debit balances, to negative amounts, Credit balances, and conversely. Third, the paper replaces the value of "Account Class" by a different value selected randomly; the combination of the aforesaid three categories forms the fourth category of potential errors.

***Prior probabilities***

The paper investigates the following three prior probabilities of "suspicious" transactions: $p1 = 5\%$, $p2 = 3\%$, and $p3 = 1\%$. For each of the three probabilities, the paper sets up a dataset that contains both legitimate and simulated "suspicious" transactions; the former type is selected randomly from the input dataset, whereas the latter is seeded according to the foregoing probabilities. The paper elaborates on this procedure in the section describing the experiments.

***Asymmetrical misclassification costs***

The model produces a binary output, $T \in \{1,0\}$, where 1 and 0 denote "suspicious" and legitimate transactions, respectively; similarly, $D \in \{1,0\}$ represents the actual classes of transactions. The model can make two types of errors: (i) Type I, or false positive, when it identifies a legitimate transaction as "suspicious", and (ii) Type II, or false negative, when it fails to identify a "suspicious" transaction as such.

Table 2 describes the two types of errors, the two correct outcomes, and their respective costs; for example, $C_{s/l}$ and $C_{l/s}$ represent the costs of false negative and false positive, respectively. The paper assumes that no costs are incurred in identifying "suspicious" and legitimate transactions correctly (i.e. $C_{s/s} = C_{l/l} = 0$). Although actual costs may be difficult to estimate, in an audit engagement the cost of a false negative, $C_{s/l}$, tends to be much higher than that of a false positive, $C_{l/s}$. Consequently, the paper investigates four cost-ratios of Type I to Type II errors: $C_{l/s}/C_{s/l}$: 1:1, 1:10, 1:20, and 1:30.

Table 2: Classification and cost matrix

| Actual | Model | |
|---|---|---|
| | suspicious (1) | legitimate (0) |
| suspicious (1) | true positive $C_{s/s}$ | false negative $C_{s/l}$ |
| legitimate (0) | false positive $C_{l/s}$ | true negative $C_{l/l}$ |

***Optimum threshold***

In order to estimate the optimum threshold, the paper minimises the total misclassification cost that is calculated as follows:

$$C = C_{l/s}\,P(T=1|D=0)\,P(D=0) + C_{s/l}\,P(T=0|D=1)\,P(D=1). \tag{1}$$

Where $P(T=1|D=0)$, the false positive rate, denotes the probability that the model identifies a transaction as "suspicious", given that it is legitimate; $P(T=0|D=1)$, the false negative rate, represents the probability that the model identifies a transaction as legitimate, given that it is "suspicious"; $P(D=1)$ represents the prior probability of "suspicious" transactions occurring in the population; and, by definition, $P(D=0) = 1 - P(D=1)$ represents the prior probability of legitimate transactions.

The binary output of the model, $T \in \{1,0\}$, can be expressed as: $T_i = \begin{cases} 1, R_i \geq u \\ 0, otherwise. \end{cases}$

Where $u$ is a threshold, $R_i = \| \vec{x}_i - \vec{m}_c \|^2$ represents the "suspiciousness" score the model estimates for each transaction, and $i = 1,2,\ldots,25,422$ denotes the transactions. A range of thresholds is initialised, and that which minimises the total misclassification cost, Equation 1, is declared to be the optimum threshold.

## 3.5 Experiments

The paper conducts the experiments in five steps depicted as an input-process-output diagram in Figure 1. First, it uses non-parametric bootstrap, $B = 100$, to select one hundred random samples with replacement from the empirical distribution of the input dataset, $X = (\vec{x}_{ij})_{nd}$. Second, for each bootstrap dataset (i.e. $X^1,\ldots,X^{100}$) the paper simulates and then seeds "suspicious" transactions according to three prior probabilities: $p1 = 5\%$, $p2 = 3\%$, and $p3 = 1\%$. For example, $Y^1_{p1}$, $Y^1_{p2}$, and $Y^1_{p3}$ denote the three datasets that correspond to the first bootstrap dataset, $X^1$, and contain simulated "suspicious" transactions in the order of $5\%, 3\%$, and $1\%$, respectively.

Third, for each bootstrap dataset, the paper uses SOM-toolbox for Matlab (Vesanto et al., 2000) to train a SOM in batch mode with hexagonal grid of neurons and Gaussian neighbourhood. Each SOM produces a set of codevectors that constitutes the reference

model describing the behaviour of the corresponding bootstrap dataset. For example, $M^2 = (\vec{m}_{kj})_{Kd}$ forms the reference model of dataset $X^2 = (\vec{x}_{ij})_{nd}$, where $K = 810$ denoting the number of neurons and that of codevectors, $n = 25,422$ reflecting the number of transactions, and $d = 3$ representing the dimensionality of input dataset. It is worth repeating that the paper trains SOMs on the bootstrap datasets that contain only legitimate transactions, and hence the sets of codevectors do not include any information concerning "suspicious" transactions.

In the fourth step, the paper applies the sets of codevectors, derived in the preceding training step, on the datasets containing both legitimate and simulated "suspicious" transactions in order to calculate the quantisation error, $R_i^b = \| \vec{y}_i{}^b - \vec{m}_c^{\,b} \|^2$; this error represents the "suspiciousness" score of each transaction. Finally, the paper initialises a range of candidate thresholds, and selects that which minimises the total misclassification cost, Equation 1, to be the optimum threshold.

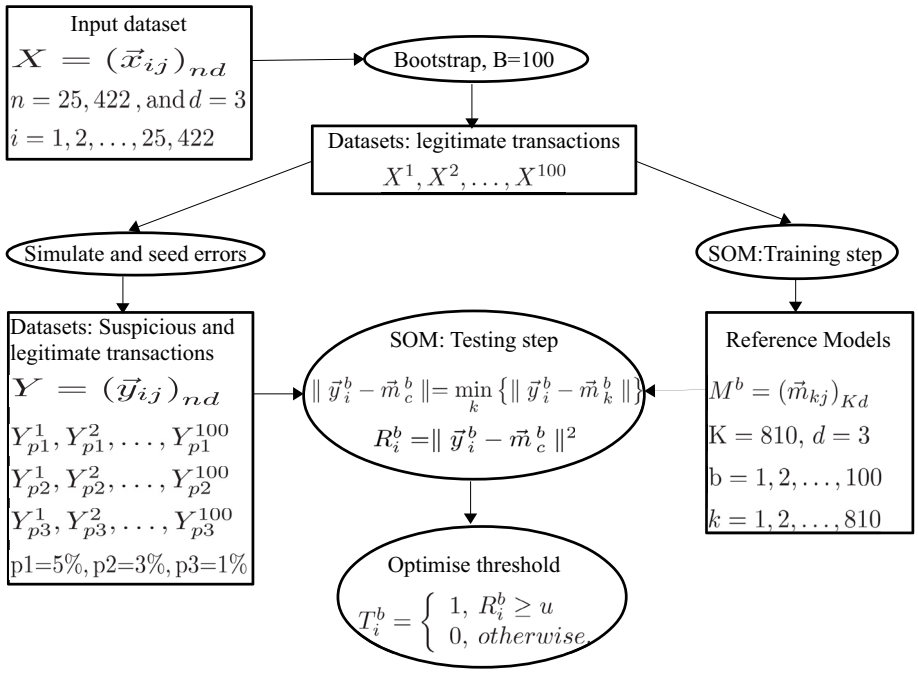



Figure 1: An input-process-output representation of the experiments

# 4 Results presentation and discussion

To recapitulate, the paper proposes and validates via experiments a SOM-based model in order to detect "suspicious" transactions. To conduct the experiments, the paper considers twelve scenarios by combining three prior probabilities of "suspicious" transactions and four cost-ratios of Type I to Type II errors. For each scenario, the paper

addresses four categories of potential errors that may affect a transaction and performs one hundred experiments by using bootstrap.

Table 3: Statistical analyses of results

| Prior | Cost Ratios | True Negative Rate | | | | | | True Positive Rate | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | | | Stdev | | | Mean | | | Stdev | | |
| | | | 95% C.I | | | 95% C.I | | | 95% C.I | | | 95% C.I | |
| 5% | 1:1 | 0.957 | 0.954 | 0.960 | 0.015 | 0.013 | 0.019 | 0.820 | 0.788 | 0.847 | 0.151 | 0.131 | 0.176 |
| | 1:10 | 0.910 | 0.861 | 0.938 | 0.185 | 0.102 | 0.281 | 0.883 | 0.854 | 0.907 | 0.130 | 0.105 | 0.158 |
| | 1:20 | 0.898 | 0.840 | 0.927 | 0.211 | 0.139 | 0.311 | 0.889 | 0.859 | 0.911 | 0.130 | 0.104 | 0.156 |
| | 1:30 | 0.897 | 0.844 | 0.930 | 0.211 | 0.136 | 0.299 | 0.891 | 0.859 | 0.912 | 0.130 | 0.108 | 0.163 |
| 3% | 1:1 | 0.953 | 0.940 | 0.959 | 0.041 | 0.019 | 0.079 | 0.881 | 0.843 | 0.908 | 0.168 | 0.136 | 0.207 |
| | 1:10 | 0.914 | 0.876 | 0.938 | 0.149 | 0.087 | 0.248 | 0.936 | 0.907 | 0.955 | 0.116 | 0.086 | 0.159 |
| | 1:20 | 0.893 | 0.836 | 0.920 | 0.191 | 0.124 | 0.278 | 0.945 | 0.918 | 0.963 | 0.111 | 0.073 | 0.153 |
| | 1:30 | 0.885 | 0.831 | 0.916 | 0.206 | 0.136 | 0.290 | 0.946 | 0.919 | 0.964 | 0.111 | 0.075 | 0.159 |
| 1% | 1:1 | 0.962 | 0.959 | 0.965 | 0.016 | 0.013 | 0.020 | 0.812 | 0.781 | 0.840 | 0.149 | 0.136 | 0.172 |
| | 1:10 | 0.917 | 0.871 | 0.943 | 0.181 | 0.099 | 0.271 | 0.860 | 0.830 | 0.885 | 0.148 | 0.128 | 0.177 |
| | 1:20 | 0.905 | 0.849 | 0.937 | 0.204 | 0.137 | 0.298 | 0.862 | 0.834 | 0.889 | 0.147 | 0.127 | 0.178 |
| | 1:30 | 0.897 | 0.840 | 0.931 | 0.216 | 0.137 | 0.298 | 0.862 | 0.829 | 0.890 | 0.148 | 0.127 | 0.177 |

The results are depicted in Table 3; each line corresponds to a scenario and represents the outcome of one hundred experiments. For example, the second line describes the performance of the model when the prior probability of "suspicious" transactions is 5% and the cost of a Type II error is ten times that of a Type I error. In this scenario, the model can detect on average 91% ($C.I_{0.95} = 0.861 - 0.938$) of the legitimate transactions as being legitimate as well as 88.3% ($C.I_{0.95} = 0.854 - 0.907$) of the "suspicious" transactions as being "suspicious"; the 95% confidence intervals are shown in parentheses.

A closer examination of the results reveals that within each of the three prior probabilities, as the cost of Type II error increases relative to that of Type I error, the mean true positive rate increases, whereas the mean true negative rate decreases. This pattern is expected, because a more costly Type II error would shift the threshold in favour of the model identifying more transactions as "suspicious". Consequently, the true positive rate would be higher, albeit only at the expense of a higher false positive rate.

To analyse further the performance of the proposed model, the paper presents the averaged Receiver Operating Characteristic Curve (i.e. ROC curve), Figure 2, that corresponds to the four scenarios examining a 5% prior probability of "suspicious" transactions. In brief, the ROC curve plots the model's true positive rate (y-axis), or sensitivity, against the false positive rate (x-axis), or 1 - specificity. The ascending diagonal stands for the random or non-informative classifier; any classifier that appears below this diagonal performs worse than chance. Further, points (0,0) and (1,1) represent classifiers that identify all transactions as "suspicious" and as legitimate, respectively; whereas, point (0,1) marks the perfect classifier that can identify all "suspicious" and legitimate transactions correctly. The area under the ROC curve, $AUROC = 0.975$, summarises the discriminatory power

of the model (Hanley and McNeil, 1982) in that it denotes the probability with which the model can identify correctly legitimate and "suspicious" transactions.

The presence of asymmetrical misclassification costs induces different operating points on the ROC curve that identify different trade-offs between true positive and false positive rates. For example, the operating points (0.0635, 0.9567) and (0.0256, 0.7798) minimise the total misclassification cost, Equation 1, when the cost-ratios of Type I to Type II errors are 1:30 and 1:1, respectively.

The uncertainty about cost-ratios can be incorporated in the ROC curve by considering the range the two operating points delineate. The area corresponding to this range can function as a measure of the model's performance that is more robust than the AUROC, because only this segment of the ROC curve is useful for decision making. In a process analogous to estimating the AUROC, the paper defines $Ratio_{segment}$ as the ratio of the area under the segment of ROC curve that is of interest to the area of the corresponding rectangle of unit length. The model exhibits a $Ratio_{segment} = 0.906$, whereas, by definition, the perfect classifier would have $Ratio_{segment} = 1$.

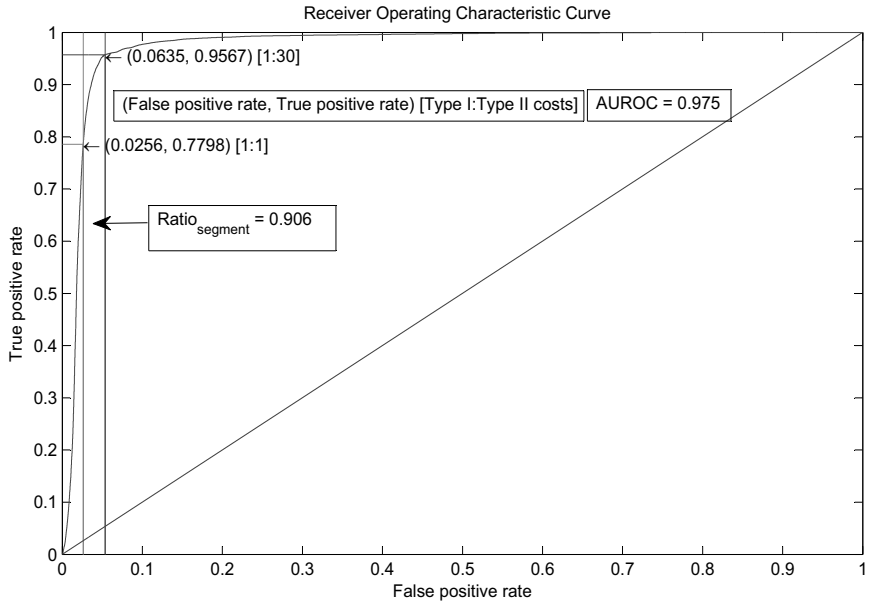



Figure 2: Averaged ROC curve for the four scenarios using a 5% probability of "suspicious" transactions.

# 5 Conclusions and directions for further research

This paper proposes a SOM-based model for detecting "suspicious" journal entries. While statute and auditing standards require auditors to consider the complete set of

journal entries in planning and performing audits, existing literature, absent a study by (Debreceny and Gray, 2010), has paid insufficient attention to the auditing of journal entries.

Empirical analyses suggest that the proposed model enjoys a high true positive rate in detecting "suspicious" journal entries. Further investigations reveal that the performance of the model is robust to varying prior probabilities of "suspicious" journal entries occurring in a population as well as to asymmetrical misclassification costs of Type I and Type II errors. These findings allow the paper to infer that the model can be generalised to datasets beyond the present case. As a result, the model can have a practical application in the domain of accounting in that it can aid auditors to detect "suspicious" journal entries.

Ongoing research aims at developing a probabilistic model that can assign a probability, rather than a score, to a journal entry. This probability would indicate the degree to which a journal entry is considered to be "suspicious", and be updated by using Bayesian analysis, given different estimates of audit risks. An additional line of research could be examining whether the model can detect "suspicious" journal entries auditors fail to do so. A further focal point of research could be coding and implementing the model in a database environment.

## REFERENCES


AICPA (2002), 'Statement On Auditing Standards 99 (SAS 99): Consideration of Fraud in a Financial Statement Audit', *American Institute of Certified Public Accountants* .

Bolton, R. J. and Hand, D. J. (2002), 'Statistical fraud detection: A review', *Statistical Science* **17**(3), 235–249.

Bonner, S., Palmrose, Z. and Young, S. (1998), 'Fraud type and auditor litigation: An analysis of SEC accounting and auditing enforcement releases', *The Accounting Review* **73**(4), 503–532.

Brockett, P. L., Xia, X. and Derrig, R. A. (1998), 'Using kohonen's Self-Organizing feature map to uncover automobile bodily injury claims fraud', *The Journal of Risk and Insurance* **65**(2), 245–274.

Chandola, V., Banerjee, A. and Kumar, V. (2009), 'Anomaly detection: A survey', *ACM Computing Surveys* **41**(3), 1–58.

COSO (2010), *Fraudulent Financial Reporting: 1998-2007. An Analysis of U.S. Public Companies*, The Committee of Sponsoring Organizations of the Treadway Commission (COSO). Available at:
http://www.coso.org/documents/COSOFRAUDSTUDY2010_001.PDF.

Debreceny, R. S. and Gray, G. L. (2010), 'Data mining journal entries for fraud detection: An exploratory study', *International Journal of Accounting Information Systems* **11**(3), 157–181.

Grabski, S. (2010), 'Discussion of "Data mining journal entries for fraud detection: An exploratory study"', *International Journal of Accounting Information Systems* **11**(3), 182–185.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology* **143**(1), 29–36.

Hogan, C. E., Rezaee, Z., Riley, R. A. and Velury, U. K. (2008), 'Financial statement fraud: Insights from the academic literature', *Auditing: A Journal of Practice and Theory* **27**(2), 231–252.

Hsu, S., Lin, C. and Yang, Y. (2008), 'Integrating neural networks for Risk-Adjustment models', *The Journal of Risk and Insurance* **75**(3), 617–642.

Juszczak, P., Adams, N. M., Hand, D. J., Whitrow, C. and Weston, D. J. (2008), 'Off-the-peg and bespoke classifiers for fraud detection', *Computational Statistics & Data Analysis* **52**(9), 4521–4532.

Kaski, S. and Lagus, K. (1996), Comparing Self-Organizing maps, *in* 'Proceedings of the 1996 International Conference on Artificial Neural Networks', Vol. 1112 of *Lecture Notes in Computer Science*, Springer-Verlag, Bochum, Germany, pp. 809–814.

Kohonen, T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences,Volume 30, second edn, Springer-Verlag, Heidelberg,Germany.

Kohonen, T. (1999), 'Comparison of SOM point densities based on different criteria', *Neural Computation* **11**(8), 2081–2095.

Labib, K. and Vemuri, R. (2002), 'NSOM: a Real-Time Network-Based intrusion detection system using Self-Organizing maps', *Network Security* .

Palmrose, Z. (1991), 'An analysis of auditor litigation disclosures.', *Auditing: A Journal of Practice and Theory* **10**(Supplement), 54–71.

Public Company Accounting Oversight Board (PCAOB) (2002), 'AU Section 316: Consideration of Fraud in a Financial Statement Audit'.

Public Company Accounting Oversight Board (PCAOB) (2004), 'Standing Advisory Group Meeting: Financial Fraud', *Available at: http://pcaobus.org/News/Events/Documents/09082004_SAGMeeting/Fraud.pdf* .

Public Company Accounting Oversight Board (PCAOB) (2007), 'Auditing Standard No. 5: An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements'.

Ramadas, M., Ostermann, S. and Tjaden, B. (2003), Detecting anomalous network traffic with self-organizing maps, *in* 'Recent Advances in Intrusion Detection', pp. 36–54.

Rhodes, B. C., Mahaffey, J. A. and Cannady, J. D. (2000), Multiple self-organizing maps for intrusion detection, *in* 'Proceedings of the 23rd national information systems security conference', pp. 16–19.

U.S. Congress (2002), 'Sarbanes-Oxley Act of 2002, H.R.3763'.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), SOM Toolbox for Matlab 5, Technical Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland. Available at: http://www.cis.hut.fi/somtoolbox/.

Ypma, A. and Duin, R. P. W. (1997), Novelty detection using self-organizing maps, *in* 'Proceedings of International Conference on Neural Information Processing. ICONIP '97', Vol. 2, Springer-Verlag, Dunedin, New Zealand, pp. 1322–1325.

**Auditing journal entries using extreme value theory**

**Argyris Argyrou**

**Abstract**
While a wealth of statutory and auditing pronouncements attest to the importance of the auditing of journal entries for preventing and detecting material misstatements to financial statements, existing literature has so far paid inadequate attention to this line of research. To explore this line of research further, this paper proposes a bipartite model that is based on extreme value theory and Bayesian analysis of Poisson distributions. The paper assesses the veracity of the model via a series of experiments on a dataset that contains the journal entries of an international shipping company for fiscal years 2006 and 2007. Empirical results suggest the model can detect journal entries that have a low probability of occurring and a monetary amount large enough to cause financial statements to be materially misstated. Further investigations reveal that the model can assist auditors to form expectations about the journal entries thus detected as well as update their expectations based on new data. The findings indicate that the model can be applied for the auditing of journal entries, and thus supplement existing procedures.

**Keywords:** auditing, journal entries, extreme value theory, Bayesian analysis

# 1 Introduction

The area of journal entries is deemed to pose a high risk of material misstatements to financial statements (Public Company Accounting Oversight Board (PCAOB), 2004); an egregious example is that of WorldCom who recorded false journal entries to artificially achieve expected revenue growth (Beresford et al., 2003). The example demonstrates that controls over the processing and recording of journal entries underpin the completeness and timeliness of financial reporting (Canadian Institute of Chartered Accountants, 2004). The importance of journal entries is attested by a wealth of auditing standards that require auditors to test the appropriateness of journal entries recorded in a general ledger (AICPA, 2002; IFAC, 2009).

However, absent a study by Debreceny and Gray (2010), existing literature has so far provided tenuous empirical evidence on how the auditing of journal entries can prevent and detect material misstatements to financial statements (Hogan et al., 2008; Grabski, 2010). Further, although literature suggests numerous procedures for detecting "anomalous" observations (Chandola et al., 2009), these procedures may not be sufficient for detecting those journal entries that can materially misstate financial statements.

Motivated by these issues, this paper proposes a bipartite model for detecting "suspicious" journal entries in order to assist auditors to identify and asses the risk of material misstatement to financial statements. The paper defines "suspicious" journal entries as having both a large monetary amount and a low probability of occurring; in

other words, “suspicious” journal entries are rare and have a monetary amount that is large enough to materially misstate financial statements.

The first component of the model employs the peaks-over-threshold method (i.e. POT), a subset of extreme value theory, to estimate an optimum threshold that can differentiate the distribution of legitimate from that of “suspicious” journal entries. The second component models the number of monthly “suspicious” journal entries in terms of a univariate Poisson distribution and uses Bayesian analysis to draw inferences. The Bayesian analysis allows auditors to update their expectations concerning “suspicious” journal entries given new data as well as transfer their knowledge from an audit engagement to the next.

The paper applies the proposed model to a dataset provided by an international shipping company. The dataset contains the complete set of the company’s journal entries for fiscal years 2006 and 2007; it was exported from the company’s database to a text file that consists of 55,350 lines and eight columns representing accounting transactions and variables, respectively. Exploratory data analysis has revealed that the transactions have different distributions between debit and credit balances as well as between one account category and another. For this reason, the paper combines the two variables, “Debit-Credit” and “Account Category”, to partition the transactions into twelve experimental cells.

In the following section, the paper sets out the background and reviews procedures for anomaly detection. Section 3 provides the motivation behind the paper’s using extreme value theory, describes the data, and introduces the model. Section 4 presents and analyses the results, and Section 5 discusses the main limitations of the paper. Section 6 draws conclusions and suggests possible directions for further research.

# 2 Background and procedures for anomaly detection

## 2.1 Background

The Statement on Auditing Standards 99: Consideration of Fraud in a Financial Statement Audit (AICPA, 2002) requires auditors, among other things, to test the appropriateness of journal entries recorded in a general ledger and other adjustments made in preparing financial statements. The Standing Advisory Group of the PCAOB (2004) considers journal entries to be an area that poses a high risk of material misstatement to financial statements due to fraud. Further, controls over the recording and processing of journal entries underpin the completeness, accuracy, and timeliness of financial reporting (Canadian Institute of Chartered Accountants, 2004).

The fraud perpetrated at WorldCom exemplifies how journal entries can be manipulated

to achieve, albeit artificially, expected revenue growth. False and unsupported journal entries were recorded to reduce operating line costs by capitalising these costs and by improperly releasing accruals; these journal entries had an estimated value of about US$ 7.3 billion (Beresford et al., 2003, pp.17 and 56).

Existing literature acknowledges the dearth of empirical evidence on how the reviewing of journal entries can detect and prevent financial statement fraud (Hogan et al., 2008; Grabski, 2010). A noteworthy exception is a study by Debreceny and Gray (2010) who used Benford's Law, or digit analysis, to detect fraudulent journal entries. In essence, the study compared the observed distribution of the first digit of USD amounts against that expected by Benford's Law; if the difference was statistically significant under a chi-square test, then the USD amount was deemed to be fraudulent.

Debreceny and Gray (2010) suggested the observed distributions were significantly different from that expected by Benford's Law for all entities in the sample. However, the results may have been an artefact of the chi-square test, as a large number of observations can induce significant results (Grabski, 2010). An additional explanation is that either fraudulent journal entries were the norm in the sample, or Benford's Law is not applicable to journal entries (Grabski, 2010).

## 2.2 Procedures for anomaly detection

Although procedures for anomaly and novelty detection abound in the literature, they may not be sufficient for detecting "suspicious" journal entries, because they make restrictive assumptions concerning data. These procedures can be classified into three broad categories: (i) two-class classification or supervised, (ii) one-class classification or semi-supervised, and (iii) unsupervised. An extensive and thorough review of these procedures can be found in (Bolton and Hand, 2002; Chandola et al., 2009).

A two-class classification procedure assumes that a dataset contains observations labelled either as "legitimate" or as "anomalous". In this case, a model (e.g. neural networks based on supervised learning) is first trained on the dataset, and then used to determine the class (i.e. "legitimate" or "anomalous" ) to which a previously unseen observation belongs. Two issues arise: first, the prevalence, or prior probability, of "anomalous" observations occurring in the population may be orders of magnitude smaller than that of the "legitimate" observations; and second, it may be difficult to obtain accurate and representative class descriptions, especially for the "anomalous" class.

An one-class classification procedure first develops a reference model that can describe the behaviour of legitimate journal entries. It then estimates a similarity metric (e.g. Euclidean distance) between the reference model and novel journal entries; the similarity metric is monotonically related to the degree of suspiciousness. Finally, it considers a

journal entry to be “suspicious”, if the journal entry has a similarity metric in excess of an optimum threshold.

However, an one-class classification is prone to a high false-positive rate. The reason is that, unless it can encompass all possible instances of legitimate behaviour, it could classify a large number of legitimate journal entries as being “suspicious”. Further, the status of legitimate behaviour is likely to change over time, and hence has to be updated as well. Additional shortcomings stem from the uneven sizes of legitimate and “suspicious” classes as well as the asymmetrical misclassification costs of Type I and Type II errors.

An unsupervised approach first estimates the probability density of data and then selects a threshold in such a way that the probability of a journal entry exceeding the threshold is very small (e.g. $P(X > u) = 10^{-4}$). A journal entry having such a small probability of occurring is deemed to be “suspicious”.

However, this approach has three main limitations. First, it implicitly assumes that “suspicious” journal entries, observations occurring beyond the threshold, follow a uniform distribution; this assumption may be restrictive or invalid in practice (Lee and Roberts, 2008). Second, it has a valid probabilistic interpretation only for classification tasks whereby a single observation is being compared against a model describing legitimate behaviour (Clifton et al., 2010). Finally, it does not provide any guidance on how a threshold should be selected; instead, the threshold is selected in a heuristic manner based on past experience and knowledge.

The literature review has revealed a lack of knowledge on how the auditing of journal entries can prevent and detect material misstatements to financial statements. Further, it has indicated that procedures for anomaly detection may not suffice to detect “suspicious” journal entries. Motivated by these issues, the paper proposes an alternative model for detecting “suspicious” journal entries that is based on extreme value theory and Bayesian analysis of Poisson distributions.

A review of extreme value theory lies beyond the scope and confines of the paper; a thorough and comprehensive treatment of this subject can be found in (Embrechts P., Kluppelberg C., and Mikosch T., 1997; Coles, 2001; Reiss and Thomas, 2007). At this juncture, it suffices to note that extreme value theory has been applied extensively in the discipline of Finance. For example, it has been applied to estimate Value-at-Risk (Longin, 2000) and expected shortfall (McNeil and Frey, 2000), to investigate contagion risk in the international banking sector (Ong et al., 2007), and to examine risk-based allocation of assets (Bensalah, 2002).

# 3 Research Design and Methodology

## 3.1 Motivation

The paper follows (Rohrbach, 1993) to conjecture that "suspicious" journal entries exhibit two distinguishing characteristics: they are rare, which means they have a low probability of occurring; and, they have a monetary amount that is sufficiently large to cause financial statements to be materially misstated. The corollary of this conjecture is that, if monetary amounts follow a unimodal distribution, then the amounts that are maxima in magnitude are also minima in probability values, and vice versa; in this case, these amounts would concentrate in the tail, or extreme-quantiles, of the unimodal distribution.

This insight motivates the paper to employ extreme value theory in order to model "suspicious" journal entries, as it is the appropriate statistical framework for studying observations that pertain to the tails, or extreme-quantiles, of a distribution.

## 3.2 Data description

The dataset has been provided by an international shipping company and consists of their journal entries for fiscal years 2006 and 2007. The dataset was exported from the database of the company to a text file that contains 55,350 lines and eight columns representing accounting transactions and variables, respectively; the variables are described in Table 1. For example, "Account Class" takes thirty values, such as: "Interest Received", "Office Expenses", "Trade Debtors" etc. In the present case, "Account Category" takes eight values: (i) "Non-Current Assets", (ii) "Cash and Cash Equivalents", (iii) "Trade and Other Receivables", (iv) "Income", (v) "Expenses", (vi) "Current Liabilities", (vii) "Non-Current Liabilities", and (viii) "Equity".

Table 1: Description of variables

| Name | Type | Values |
|---|---|---|
| Account Number | Alphanumeric | 360 |
| Account Description | Text | 360 |
| Posting Date | Date | 24 |
| Debit-Credit | Binary | 2 |
| USD Amount | Numerical | |
| Transaction Details | Text | |
| Account Class | Categorical-hierarchical | 30 |
| Account Category | Categorical-hierarchical | 8 |

The variables "Account Number", "Account Class", and "Account Category" group transactions in an ascending order of aggregation. For example, the account category "Trade and Other Receivables" consists of 23, 886 transactions; this number represents the aggregation of six account classes: "Sales Taxes Receivable" (1,410), "Trade

Debtors" (6,217), "Other Debtors" (5,634), "Loans Receivable" (164), "Insurance Receivables" (9,361), and "Other Receivables" (1,100). The number of transactions is shown in parentheses.

The paper groups the transactions according to the "Account Category" variable, because there are not enough transactions at lowers levels of aggregation (i.e. "Account Number", "Account Class") to estimate the parameters of the proposed model. For the same reason, the paper excludes "Non-Current Assets" and "Equity" completely.

In addition, the paper excludes those transactions auditors would select as a standard procedure in the normal course of an audit; for example, transactions that record transfers to reserves, year-end consolidation, and closing Profit and Loss items to the Balance Sheet. Although the paper has not investigated the counter-factual, including these transactions would cause the model to estimate a higher threshold than otherwise; the reason is this type of transactions tends to have large monetary amounts and occur infrequently, often at the end of a fiscal year. As a result, the model would select transactions, which are selected anyway, but ignore transactions that may warrant further investigation.

Exploratory data analysis suggests the distributions of transactions are different between one account category and another as well as between debit and credit balances. As a result, the paper combines the two variables, "Debit-Credit" and "Account Category", to partition the transactions into twelve experimental cells, as shown in Table 2.

Table 2: Descriptive statistics for fiscal years 2006 and 2007

| Account Category | N | Mean | Mode | Median | Var |
|---|---|---|---|---|---|
| | USD:Credit | | | | |
| Cash and Cash Equivalents | 432 | -2,449 | -26 | -215 | 13,424,043 |
| Trade and Other Receivables | 12,322 | -2,879 | -1,052 | -1,053 | 14,819,048 |
| Income | 5,567 | -1,081 | 0 | -422 | 2,348,775 |
| Expenses | 860 | -757 | -11 | -187 | 2,304,802 |
| Current Liabilities | 11,087 | -1,270 | -203 | -272 | 5,604,960 |
| Non-Current Liabilities | 143 | -1,228 | -2,677 | -624 | 2,974,177 |
| | 30,411 | | | | |
| | USD:Debit | | | | |
| Cash and Cash Equivalents | 2,655 | 6,154 | 1,274 | 4,168 | 29,997,396 |
| Trade and Other Receivables | 11,564 | 2,985 | 1,052 | 1,188 | 16,255,837 |
| Income | 1,663 | 1,225 | 0 | 441 | 4,944,962 |
| Expenses | 4,162 | 614 | 10 | 72 | 2,806,104 |
| Current Liabilities | 4,763 | 968 | 203 | 293 | 3,751,541 |
| Non-Current Liabilities | 89 | 2,000 | 109 | 735 | 9,538,657 |
| | 24,896 | | | | |

## 3.3 Modelling "suspicious" transactions using the peaks-over-threshold method

In order to introduce the peaks-over-threshold method, the paper uses a concrete example that is based on the results and depicted in Fig. 1. Let variable $X = (x_1, \ldots, x_n)$ denote the monetary amounts of the transactions belonging to the "Debit" side of "Trade and Other Receivables", where $n = 11,564$ representing the number of transactions. Variable $X$ can be assumed to be an independent and identically distributed (i.e. iid) random variable that follows an unknown distribution. The distribution is unimodal, and hence its probability density function decreases monotonically with increasing distance from the single mode, as shown in Fig. 1a.

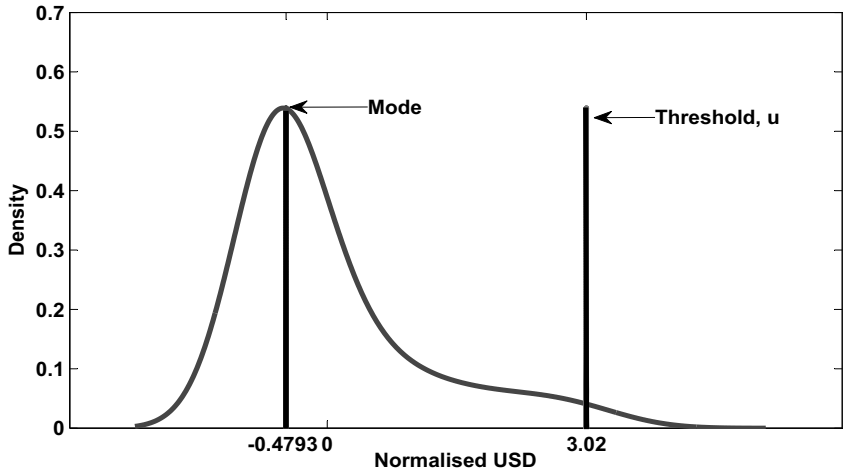



Fig. 1a : Probability density function: Trade and Other Receivables - Debit side

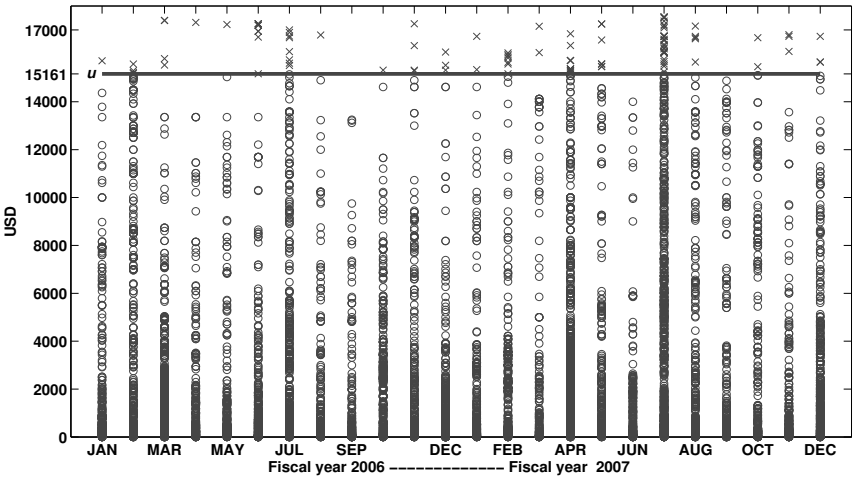



Fig. 1b : Peaks-over-threshold: Trade and Other Receivables - Debit side

Because the distribution is unimodal, the further from the mode a monetary amount is, the larger its magnitude and the lower its probability of occurring would be. In other words, transactions that are extreme in monetary amounts are also minima in probability density; the converse is also true. This insight motivates the paper to estimate an optimum threshold, $u$, that can differentiate the distribution of legitimate from that of "suspicious" transactions. Transactions whose monetary amounts exceed the optimum threshold are considered to be "suspicious", because these amounts are both rare and large.

For a sufficiently high threshold, $u$, Pickand's theorem (Pickands, 1975) describes the distribution of excesses over threshold $u$ conditional on the threshold being exceeded, i.e. $(X-u|X>u)$, in terms of a distribution within the Generalized Pareto (GP) family, as follows (Coles, 2001, p.75):

$$F(x-u;\tilde{\sigma},\xi)=\begin{cases}1-\left[1+\xi\left(\frac{x-u}{\tilde{\sigma}}\right)\right]_{+}^{\frac{-1}{\xi}} & \text{if } \xi\neq 0,\\ 1-exp\left[-\frac{x-u}{\tilde{\sigma}}\right] & \text{if } \xi=0.\end{cases} \tag{1}$$

The distribution, Eq. 1, has the following probability density function:

$$f(x-u;\tilde{\sigma},\xi)=\begin{cases}\left(\frac{1}{\tilde{\sigma}}\right)\left[1+\xi\left(\frac{x-u}{\tilde{\sigma}}\right)\right]_{+}^{\frac{-1-\xi}{\xi}} & \text{if } \xi\neq 0,\\ \left(\frac{1}{\tilde{\sigma}}\right)exp\left[-\frac{x-u}{\tilde{\sigma}}\right] & \text{if } \xi=0.\end{cases} \tag{2}$$

Where $x-u>0$, $\tilde{\sigma}>0$, and $\tilde{\sigma}$, $\xi$ stand for the scale and the shape of a GP distribution, respectively. The shape parameter, $\xi$, describes the tail of a GP distribution; if $\xi<0$ then the distribution of excesses, $x-u$, has an upper bound at $u-\frac{\tilde{\sigma}}{\xi}$, which means that the probability density for $x>u-\frac{\tilde{\sigma}}{\xi}$ is zero; if $\xi>0$ then the distribution decreases polynomially and has a lower bound at $u-\frac{\tilde{\sigma}}{\xi}$; and if $\xi=0$ then the distribution decreases exponentially and has no lower nor upper bound.

Selecting a threshold that is sufficiently high so that Pickand's theorem could apply is fraught with difficulties (Embrechts P., Kluppelberg C., and Mikosch T., 1997, p. 355). Suffice it to note that a threshold strikes a trade-off between bias and variance in a model. A too low threshold could lead to sampling from the main body of a distribution, non-extremal values, and thus induce bias in estimating the parameters of a model. On the other hand, as a threshold increases the number of excesses with which parameters can be estimated decreases, and hence the standard errors of the parameters would increase.

In order to estimate an optimum threshold, the paper follows a three-step approach for each of the twelve experimental cells, described in Table 2. First, the paper initialises a set of candidate thresholds that take values between the 95% and 99% percentiles of the amounts. Second, at each candidate threshold, the paper fits a GP distribution, Eq. 1, and estimates the parameters, $\theta=(\tilde{\sigma},\xi)$, by maximising the log-likelihood function (Coles,

2001, p.80):

$$LL(\tilde{\sigma},\xi;x-u)=\begin{cases} -Nlog(\tilde{\sigma})-\left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{N} log\left[1+\xi\left(\frac{x_i-u}{\tilde{\sigma}}\right)\right]_+ & \text{if } \xi\neq 0, \\ -Nlog(\tilde{\sigma})-\frac{1}{\tilde{\sigma}}\sum_{i=1}^{N}(x_i-u) & \text{if } \xi=0. \end{cases} \quad (3)$$

Where $N$ denotes the number of excesses over threshold $u$. Finally, the paper selects the threshold that corresponds to the GP distribution having the maximum log-likelihood.

This threshold becomes the decision boundary that distinguishes legitimate from "suspicious" transactions. For example, Fig.1b depicts the threshold, $u = 15,161$ USD, that separates legitimate from "suspicious" transactions; in this example, there are $11,348$ legitimate and 216 "suspicious" transactions.

## 3.4 Bayesian analysis of Poisson distributions

The paper models the number of monthly "suspicious" transactions in terms of a univariate Poisson distribution and draws inference via the Bayes' rule. Let the observed number of monthly "suspicious" transactions be denoted by the discrete variable $V$ that follows a Poisson distribution having a probability mass function $\Pr(V = v) = \frac{\lambda^v e^{-\lambda}}{v!}$, where $\lambda > 0$ and $v \geqslant 0$.

Let $\lambda$ be the unobserved and unknown average number of "suspicious" transactions over the 24-month period under investigation; and, let the prior distribution of $\lambda$, $p(\lambda)$, denote the degree of certainty, or inductive bias, about $\lambda$ in the absence of any observed evidence.

The probability of a "suspicious" transaction occurring is assumed to depend on $\lambda$. This dependence can be formalised as $p(V|\lambda)$, which is the conditional probability of the observed number of "suspicious" transactions for each possible value of $\lambda$; it is also termed the likelihood function of $\lambda$, $L(\lambda)$. The Bayes' rule combines prior probability and likelihood function to estimate the conditional probability, $p(\lambda|V)$, for different values of $\lambda$ taking into account observed evidence, $V$.

Formally, the Bayes' rule states:

$$p(\lambda|V)=\frac{p(V|\lambda)\,p(\lambda)}{\sum_{\lambda} p(V|\lambda)\,p(\lambda)}, \qquad \text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Marginalising factor}} \ . \quad (4)$$

The likelihood function of $\lambda$ is given by: $p(V|\lambda) = \prod_{i=1}^{m} \frac{\lambda^{v_i} e^{-\lambda}}{v_i!}$, where $v_i$ represents the number of "suspicious" transactions for the $i^{th}$ month and $m = 24$ reflects the number of months.

The paper chooses the Gamma distribution as a conjugate prior, $\lambda \sim Gamma(\alpha,\beta)$, and sets the hyper-parameters equal to one, $\alpha = \beta = 1$, for the prior to be non-informative.

Because of the conjugacy property, the posterior distribution of the Poisson parameter follows a Gamma distribution as well: $\lambda|V \sim Gamma(\sum_{i=1}^{m} v_i + \alpha, \beta + m)$.

The posterior distribution has a closed-form solution, as follows:

$$p(\lambda|V) = \frac{\tilde{\beta}^{\tilde{\alpha}} \lambda^{\tilde{\alpha}-1} \exp(-\tilde{\beta}\lambda)}{\Gamma(\tilde{\alpha})} \quad , \tag{5}$$

where $\tilde{\alpha} = \sum_{i=1}^{m} v_i + \alpha$, $\tilde{\beta} = \beta + m$, and $\Gamma(\tilde{\alpha}) = (\tilde{\alpha} - 1)!$ is the Gamma function of $\tilde{\alpha}$.

# 4 Results presentation and discussion

## 4.1 Modelling "suspicious" transactions

The results, summarised in Table 3, identify the parameters of the best-fitted GP model, the thresholds, $u$, and the number of "suspicious" transactions, $N(u)$. For example, one hundred GP models at varying thresholds are fitted to the transactions that belong to the "Debit" side of "Trade and Other Receivables". The best-fitted GP model occurs when the threshold is USD 15,161, and hence 216 transactions are considered to be "suspicious", as depicted in Fig. 1b.

Table 3: Generalized Pareto models

| | | | | | Gener. Pareto Distribution | | |
|---|---|---|---|---|---|---|---|
| Account Category | N | $N(u)$ | % | u | Shape | Scale | N.Log.Likel. |
| | | | USD:Credit | | | | |
| Cash and Cash Equivalents | 432 | 5 | 1.16% | -13,551 | -1.421 | 0.300 | -17.549 |
| Trade and Other Receivables | 12,322 | 554 | 4.5% | -12,695 | -0.562 | 0.325 | -379.392 |
| Income | 5,567 | 11 | 0.19% | -12,821 | 1.635 | 0.003 | -34.214 |
| Expenses | 860 | 2 | 0.23% | -11,627 | -4.399 | 6.151 | -25.513 |
| Current Liabilities | 11,087 | 23 | 0.21% | -13,629 | 1.524 | 0.021 | -31.134 |
| Non-Current Liabilities | 143 | 27 | 18.88% | -2,677 | 4.560 | 0.000 | -100.799 |
| | 30,411 | 622 | 2.04% | | | | |
| | | | USD:Debit | | | | |
| Cash and Cash Equivalents | 2,655 | 225 | 8.47% | 15,137 | -0.98 | 0.43 | -184.75 |
| Trade and Other Receivables | 11,564 | 216 | 1.86% | 15,161 | -0.51 | 0.34 | -129.32 |
| Income | 1,663 | 23 | 1.38% | 12,811 | 1.47 | 0.01 | -51.24 |
| Expenses | 4,162 | 3 | 0.07% | 15,657 | -2.21 | 1.21 | -20.42 |
| Current Liabilities | 4,763 | 8 | 0.17% | 16,366 | -1.75 | 0.64 | -22.32 |
| Non-Current Liabilities | 89 | 2 | 2.25% | 11,482 | -3.65 | 4.44 | -24.37 |
| | 24,896 | 477 | 1.91% | | | | |

The results suggest the proposed model can perform more efficiently than simply selecting the largest $x\%$ (e.g. 5%) of the amounts. For example, the thresholds estimated for the debit and credit sides of "Trade and Other Receivables" select about 1.86% and 4.5% of the corresponding transactions, respectively. The reason for this efficiency is the model can estimate a threshold that is a function of three variables: (i) "USD Amount", (ii) "Debit-Credit", and (iii) "Account Category". On the other hand, a uniform threshold is a function of only "USD Amount".

The proposed model can estimate a threshold that, at least in principle, can be interpreted in the context of extreme value theory, whereas a heuristic threshold lacks any interpretation. In this respect, the model can mitigate the subjectivity and bias that may occur when auditors select a threshold only on the basis of their past experience and knowledge (Tversky and Kahneman, 1974; Trotman et al., 2011).

## 4.2 Bayesian analysis

Figure 2 depicts the Bayesian analysis of the monthly "suspicious" transactions that belong to the "Debit" side of "Trade and Other Receivables". The maximum a posteriori (MAP), the mode of the posterior distribution, denotes the number of monthly "suspicious" transactions, $\lambda = 9$, that has the highest probability of occurring. Further, the MAP has a 95% credible interval of $8 - 10$, which means that, given the data, there is a 95% probability that the number of monthly "suspicious" transactions is either 8, or 9, or 10. In this case, the MAP is the same as the maximum likelihood estimate (MLE), which is the average calculated form the data (i.e. 216 "suspicious" transactions / 24 months). The reason is the posterior distribution is estimated only on the basis of the data, as the prior probability has been chosen to be non-informative.

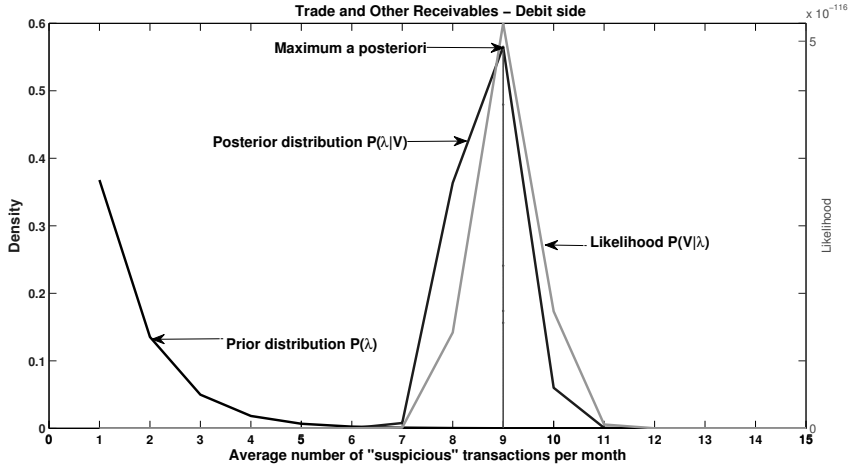



Figure 2: Bayesian analysis for Trade and Other Receivables - Debit side

Auditors can apply the Bayesian analysis in two ways. First, they can select those monthly transactions that have the $8 - 10$ largest monetary amounts, provided the monetary amounts follow a unimodal distribution. Second, they can update their evidence sequentially, as the current posterior distribution will become the prior distribution in next year's audit.

# 5 Caveats and limitations

In the context of this paper, the monetary amounts follow a unimodal distribution, and hence those transactions that are extreme in monetary amounts are also minima in probability density values. However, if the monetary amounts follow a multimodal distribution, then the correspondence between large amounts and low probability does not hold, because there is not a single mode from which distances can be defined. This possibility limits the applicability and veracity of the results. Nonetheless, this limitation is not irredeemable, as a future study could extend the paper to multimodal distributions by employing an appropriate methodology; for example, Clifton et al. (2010) describe such methodology.

The model does not include two important variables: time, and audit risk. In particular, the model is static in that it derives a time-invariant threshold, as shown in Fig. 1b. However, it may not be realistic to use a constant threshold throughout a fiscal year, because there is much more scope for manipulating journal entries at year-end than there is during the rest of the year. Ongoing research examines how a temporal covariate can incorporate information about cycles and trends that may be present in the distribution of journal entries.

Further, the model does not take into account the concept of audit risk, as it implicitly assumes that all audit engagements pose the same level of audit risk. An extension to the model could include auditors' assessment of audit risk and its components (i.e. inherent, detection, control). For example, an audit engagement having a high inherent risk would be assigned a much lower threshold than otherwise, other things being equal.

# 6 Conclusions and directions for future research

Existing literature, absent a study by Debreceny and Gray (2010), has so far provided little empirical evidence on how the auditing of journal entries can prevent and detect material misstatements to financial statements. The lack of evidence becomes more pronounced, given that a wealth of auditing standards require auditors to consider the complete set of journal entries in planning and performing audits. The auditing of journal entries becomes problematic considering that established procedures for anomaly detection may not suffice for detecting those journal entries that may cause financial statements to be materially misstated. Motivated by these issues, the paper proposes a bipartite model in order to assist auditors to detect such journal entries.

The results suggest the model can detect journal entries that are both rare and have a monetary amount large enough to materially misstate financial statements. Further, the Bayesian analysis indicate how auditors can form as well as update expectations about "suspicious" journal entries.

The paper has raised some questions that may support additional research. Ongoing research aims at incorporating a temporal covariate for the model to estimate a threshold that can capture potential cycles and trends existing in the distribution of journal entries; a further extension to the model could include a covariate for auditors' assessment of audit risk. An additional study could compare the performance of the model against that of unaided auditors who rely only on their past experience and knowledge.

## REFERENCES


AICPA (2002), 'Statement On Auditing Standards 99 (SAS 99): Consideration of Fraud in a Financial Statement Audit', *American Institute of Certified Public Accountants* .

Bensalah, Y. (2002), Asset allocation using extreme value theory, Working Papers 2002-2, Bank of Canada. Available at: http://www.bankofcanada.ca/wp-content/uploads/2010/02/wp02-2.pdf.

Beresford, D. R., Katzenbach, N. d. and Rogers, C. B. J. (2003), *Report of Investigation by the Special Investigative Committee of the Board of Directors of WorldCom, INC.*, Available at: http://news.findlaw.com/wsj/docs/worldcom/bdspcomm60903rpt.pdf.

Bolton, R. J. and Hand, D. J. (2002), 'Statistical fraud detection: A review', *Statistical Science* **17**(3), 235–249.

Canadian Institute of Chartered Accountants (2004), *IT Control Assessments in the context of CEO/CFO Certification*, Toronto,Canada.

Chandola, V., Banerjee, A. and Kumar, V. (2009), 'Anomaly detection: A survey', *ACM Computing Surveys* **41**(3), 1–58.

Clifton, D., Hugueny, S. and Tarassenko, L. (2010), 'Novelty detection with multivariate extreme value statistics', *Journal of Signal Processing Systems* pp. 1–19.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, Springer-Verlag, London, UK.

Debreceny, R. S. and Gray, G. L. (2010), 'Data mining journal entries for fraud detection: An exploratory study', *International Journal of Accounting Information Systems* **11**(3), 157–181.

Embrechts P., Kluppelberg C., and Mikosch T. (1997), *Modelling Extremal Events for Insurance and Finance*, Vol. 33 of *Stochastic Modelling and Applied Probability*, Springer-Verlag.

Grabski, S. (2010), 'Discussion of "Data mining journal entries for fraud detection: An exploratory study"', *International Journal of Accounting Information Systems* **11**(3), 182–185.

Hogan, C. E., Rezaee, Z., Riley, R. A. and Velury, U. K. (2008), 'Financial statement fraud: Insights from the academic literature', *Auditing: A Journal of Practice and Theory* **27**(2), 231–252.

IFAC (2009), 'International Standards on Auditing 240 (ISA 240). The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements', *International Federation of Accountants* .

Lee, H. and Roberts, S. (2008), On-line novelty detection using the kalman filter and extreme value theory, *in* 'Pattern Recognition 2008', Tampa, FL,USA, pp. 1–4.

Longin, F. M. (2000), 'From value at risk to stress testing: The extreme value approach', *Journal of Banking & Finance* **24**(7), 1097–1130.

McNeil, A. J. and Frey, R. (2000), 'Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach', *Journal of Empirical Finance* **7**(3–4), 271–300.

Ong, L. L., Mitra, S. and Chan-Lau, J. A. (2007), Contagion Risk in the International Banking System and Implications for London as a Global Financial Center, IMF Working Papers WP-07-74, International Monetary Fund.
Available at: http://www.imf.org/external/pubs/ft/wp/2007/wp0774.pdf.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *The Annals of Statistics* **3**(1), 119–131.

Public Company Accounting Oversight Board (PCAOB) (2004), 'Standing Advisory Group Meeting: Financial Fraud', *Available at: http://pcaobus.org/News/Events/Documents/09082004_SAGMeeting/Fraud.pdf* .

Reiss, R. and Thomas, M. (2007), *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*, third edn, Birkhäuser, Basel, Switzerland.

Rohrbach, K. J. (1993), 'Variance Augmentation to Achieve Nominal Coverage Probability in Sampling from Audit Populations.', *Auditing: A Journal of Practice and Theory* **12**(2), 79–97.

Trotman, K. T., Tan, H. C. and Ang, N. (2011), 'Fifty-year overview of judgment and decision-making research in accounting', *Accounting and Finance* **51**(1), 278–360.

Tversky, A. and Kahneman, D. (1974), 'Judgment under uncertainty: Heuristics and biases', *Science* **185**(4157), 1124–1131.

**EKONOMI OCH SAMHÄLLE**
Skrifter utgivna vid Svenska handelshögskolan

**ECONOMICS AND SOCIETY**
Publications of the Hanken School of Economics

222. KATARINA HELLÉN: A Continuation of the Happiness Success Story: Does Happiness Impact Service Quality? Helsinki 2010.

223. MARIA JAKUBIK (Maria Jakubikne Toth): Becoming to Know: Essays on Extended Epistemology of Knowledge Creation. Helsinki 2011.

224. JOHANNA GUMMERUS: Customer Value in E-Service. Conceptual Foundation and Empirical Evidence. Helsinki 2011.

225. MARIA SOLITANDER: When Sharing Becomes a Liability: An Intellectual Capital Approach to Describing the Dichotomy of Knowledge Protection versus Sharing in Intra- and Interorganizational Relationships. Helsinki 2011.

226. BEATA SEGERCRANTZ: "... The Walls Fell Down but the Blokes Just Coded...". Varieties of Stability in Software Product Development during Organizational Restructurings. Helsinki 2011.

227. ERIC BREIT: On the Discursive Construction of Corruption: A Critical Analysis of Media Texts. Helsinki 2011.

228. MICHAEL WAGNER: Inventory Routing. A Strategic Management Accounting Perspective. Helsinki 2011.

229. NIKODEMUS SOLITANDER: Designing Creativity Through Clusters: A Periodisation of Cluster Discourse. Helsinki 2011.

230. ANDREAS PERSSON: Profitable Customer Management: a Study in Retail Banking. Helsinki 2011.

231. MATHIAS HÖGLUND: Microfoundations of HRM Effects: Individual and Collective Attitudes and Performance. Helsinki 2011.

232. ANETTE SÖDERQVIST: Opportunity Exploration and Exploitation in International New Ventures. A Study of Relationships' Involvement in Early Entrepreneurial and Internationalisation Events. Helsinki 2011.

233. OSKARI LEHTONEN: An Examination of How Entrepreneurs Can Improve Their Position in Relationship to Investors. Helsinki 2011.

234. IMOH ANTAI: Operationalizing Supply Chain vs. Supply Chain Competition. Helsinki 2011.

235. DMITRI MELKUMOV: Towards Explaining the Tasks and Roles of the Boards of Directors: The Role of Contextual, Behavioural and Social Identification Factors. Helsinki 2011.

236. CHARLOTTA NIEMISTÖ: Work/Family Reconciliation: Corporate Management, Family Policies, and Gender Equality in the Finnish Context. Helsinki 2011.

237. PAUL VIIO: Strategic Sales Process Adaptation: Relationship Orientation of the Sales Process in a Business-to-Business Context. Helsinki 2011.

238. KHALID BHATTI: Factors Affecting Knowledge Sharing in Strategic Alliances: The Role of Knowledge Sharing as Strategic Control Behavior among Multinational Enterprises. Helsinki 2011.

239. STEFAN GRANQVIST: Effektutvärdering inom företagandeutbildning. Helsingfors 2011.

240. HEIKKI RANNIKKO: Early Development of New Technology-Based Firms. A Longitudinal Analysis of New Technology-Based Firms' Development from Population Level and Firm Level Perspectives. Helsinki 2012.

241. PAULINA JUNNI: Knowledge Transfer in Acquisitions: A Socio-Cultural Perspective. Helsinki 2012.

242. HENRIKA FRANCK: Ethics in Strategic Management: An Inquiry into Otherness of a Strategy Process. Helsinki 2012.

243. SEPPO LAUKKANEN: Making Sense of Ambidexterity. A Process View of the Renewing Effects of Innovation Activities in a Multinational Enterprise. Helsinki 2012.

244. MARKUS WARTIOVAARA: Values and Freedom: An Inquiry into the Rise and Fall of Billionaire Wealth. Helsinki 2012.

245. SAINT KUTTU: Essays on Volatility and Time Varying Conditional Jumps in Thinly Traded African Financial Markets**.** Helsinki 2012.

246. ROSA MARIA BALLARADINI: Intellectual Property Protection for Computer Programs. Developments, Challenges, and Pressures for Change. Helsinki 2012.

247. VIOLETTA KHOREVA: Gender Inequality, Gender Pay Gap, and Pay Inequity. Perceptions and Reactions in Finnish Society and Workplaces. Helsinki 2012.

248. VIRPI SORSA: Discourse and the Social Practice of Strategy. Of Interaction, Texts, and Power Effects. Helsinki 2012.

249. XING LIU: Empirical Research on Spatial and Time Series Properties of Agricultural Commodity Prices. Helsinki 2012.

250. ROLANDO MARIO TOMASINI PONCE: Informal Learning Framework for Secondment: Logistics Lessons from Disaster Relief Operations. Helsinki 2012.

251. LINDA SCHOLLENBERG: Essays on the Economics of Environmental and Sustainability Labelling. Helsinki 2013.

252. NADER SHAHZAD VIRK: Explanations for Finnish Stock Returns with Fundamental and Anomalous Risks. Helsinki 2013.

253. TAMARA GALKINA: Entrepreneurial Networking: Intended and Unintended Processes. Helsinki 2013.

254. JOHANNA ARANTOLA-HATTAB: Family as a Customer Experiencing Co-Created Service Value. Helsinki 2013.