Learning Sampling in Financial Statement Audits using Vector Quantised Variational Autoencoder Neural Networks

Marco Schreyer
University of St.Gallen
St.Gallen, Switzerland
marco.schreyer@unisg.ch

Timur Sattarov
Deutsche Bundesbank
Frankfurt am Main, Germany
timur.sattarov@bundesbank.de

Anita Gierbl
University of St.Gallen
St.Gallen, Switzerland
anita.gierbl@unisg.ch

Bernd Reimer
PricewaterhouseCoopers GmbH WPG
Stuttgart, Germany
reimer.bernd@pwc.com

Damian Borth
University of St.Gallen
St.Gallen, Switzerland
damian.borth@unisg.ch

ABSTRACT

The audit of financial statements is designed to collect reasonable assurance that an issued statement is free from material misstatement ('true and fair presentation'). International audit standards require the assessment of a statements' underlying accounting relevant transactions referred to as 'journal entries' to detect potential misstatements. To efficiently audit the increasing quantities of such journal entries, auditors regularly conduct an 'audit sampling' i.e. a sample-based assessment of a subset of these journal entries. However, the task of audit sampling is often conducted early in the overall audit process, where the auditor might not be aware of all generative factors and their dynamics that resulted in the journal entries in-scope of the audit. To overcome this challenge, we propose the use of a *Vector Quantised-Variational Autoencoder (VQ-VAE)* neural networks to learn a representation of journal entries able to provide a *comprehensive* 'audit sampling' to the auditor. We demonstrate, based on two real-world city payment datasets, that such artificial neural networks are capable of learning a quantised representation of accounting data. We show that the learned quantisation uncovers (i) the latent factors of variation and (ii) can be utilised as a highly representative audit sample in financial statement audits.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; *Unsupervised learning*; *Dimensionality reduction and manifold learning*; • **Information systems** → *Enterprise resource planning*.

KEYWORDS

deep learning, audit, autoencoder neural networks, vector quantisation, audit sampling, computer assisted audit, audit data analytics, accounting information systems

ICAIF '20, October 15–16, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7584-9/20/10…$15.00

ACM Reference Format:
Marco Schreyer, Timur Sattarov, Anita Gierbl, Bernd Reimer, and Damian Borth. 2020. Learning Sampling in Financial Statement Audits using Vector Quantised Variational Autoencoder Neural Networks. In Proceedings of the *ACM International Conference on AI in Finance (ICAIF '20), October 15–16, 2020, New York, NY, USA* Proceedings of the First ACM International Conference on AI in Finance 2020, New York, NY, USA, 8 pages.

1 INTRODUCTION

The trustworthiness of financial statements plays a fundamental role [2] in today's economic decision making by investors. The audit of such statements, conducted by external auditors, is designed to collect reasonable assurance that an issued financial statement is free from material misstatement ('true and fair presentation') [1, 5]. International audit standards require an assessment of the statement's underlying accounting relevant *journal entries* to detect a potential misstatement [3]. Journal entries debit and credit the separate accounting ledgers of a financial statement evident in its balance sheet and profit and loss statement. Nowadays, organizations collect vast quantities of such journal entries in *Accounting Information Systems (AIS)* or more general *Enterprise Resource Planning (ERP)* systems [19]. Figure 1 depicts a hierarchical view of the journal entry recording process in designated database tables of an ERP system.

To efficiently audit the increasing quantities of journal entries, auditors regularly conduct a sample-based assessment referred to as *audit sampling*. Formally, audit sampling is defined as the *'selection and evaluation of less than 100% of the entire population of journal entries'* [6]. While sampling increases the efficiency of a financial audit, it also increases its *sampling risk*. The term sampling risk denotes the likelihood that the auditor's conclusion based on auditing a subset of entries may differ from the conclusion of auditing the entire population [21]. Auditors are required to select a *representative* sample from the population of journal entries [4] to mitigate sampling risks. The selection of such a representative sample is determined by the design of the applied sampling technique. International audit standards require auditors to built their audit approach on either (i) *non-statistical* or (ii) *statistical* sampling techniques. Non-statistical or 'judgemental' sampling denotes the selection of audit samples based on the auditor's experience, inherent risk assessment, and 'professional judgment'. Statistical sampling, in contrast, aims to provide greater sampling objectivity.

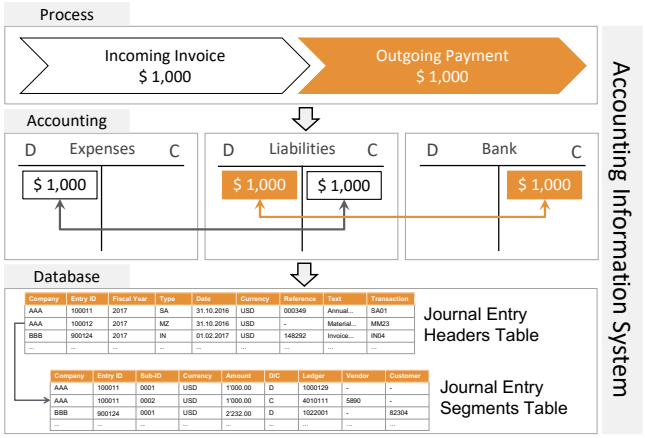

Figure 1: Hierarchical view of an Accounting Information System (AIS) that records distinct layer of abstractions, namely (1) the business process, (2) the accounting and (3) technical journal entry information in designated tables.

It refers to the random selection of audit samples and the determination of probabilities to evaluate the sampling results [4]. To conduct statistical sampling and determine a representative sample size, auditors nowadays use pre-calculated probability tables or sampling functions available in computer-assisted audit software [44], such as ACL[1], IDEA[2], or proprietary applications.

During an annual audit, the task of audit sampling is regularly conducted early in the audit process. Thereby, auditors need to decide on sensitive sampling parameters before conducting the sampling, e.g. materiality levels, confidence intervals, and acceptable risk thresholds [26]. Even though the auditor might be unaware of all latent factors that generated the journal entries in-scope of the audit, e.g. the underlying business processes and workflows. This challenge is in particular evident in the context of new audit engagements in which auditors are mandated to audit an organisation's financial statement for the first time. However, it is also of relevance in mature audits, especially in scenarios where the generative factors of an organisation vary dynamically over time [38], e.g. due to a merger or carve-out that results in organisational process- and workflows-changes.

Driven by the rapid technological advances of artificial intelligence in recent years, techniques based on deep learning [31] have emerged into the field of finance [13, 47], and particular financial statement audits [10, 41–43]. These developments raise the question: Can such techniques also be utilised to learn representative audit samples? And if so, can the samples be drawn in a way that they are interpretable by a human auditor? In this work, we propose the application of *Vector Quantised-Variational Autoencoder (VQ-VAE)* neural networks [46] to address those questions. Inspired by the idea of discrete neural dimensionality reduction, we demonstrate how VQ-VAE neural networks can be trained to learn (i) representative and (ii) human interpretable audit samples. In summary, we present the following contributions:

- We demonstrate that VQ-VAEs can be utilised to learn a low-dimensional representation of journal entries recorded in AIS systems that disentangle the entries latent generative factors of variation;
- We show that the VQ-VAEs training can also be regularised to learn a set of discrete embeddings that quantise the representations generative factors and therefore constitute a representative audit sample;
- We illustrate that the learned quantisation reflects the entries generative accounting processes and provides a starting point for human interpretable audit sampling and downstream substantive audit procedures.

The remainder of this work is structured as follows: In section 2, we provide an overview of related work. Section 3 follows with a description of the VQ-VAE architecture and presents the proposed methodology to learn a representative sample from vast quantities of journal entries. The experimental setup and results are outlined in section 4 and section 5. In section 6, the paper concludes with a summary of the current work and future research directions.

2 RELATED WORK

Due to its high relevance for the financial audit practice, the task of audit sampling triggered a sizable body of research by academia [7, 15] and practitioners [20, 25]. In the realm of this work, we focus our literature review on the two main classes of statistical sampling techniques, namely (i) attribute sampling, and (ii) variables sampling used nowadays [23] in auditing journal entries:

2.1 Attribute Sampling Techniques

The technique of attribute sampling is applied by auditors to estimate the percentage of journal entries that possess a specific attribute or characteristic. The different techniques of attribute sampling encompass the following classes [21]:

- *Random sampling* in which each journal entry of the population has an equal chance of being selected.
- *Systematic* or *sequential sampling* in which the sampling starts by selecting an entry at random and then every n-th journal entry of an ordered sampling frame is selected.
- *Proportional, block* or *stratified sampling* in which the population is sub-divided into homogeneous groups of journal entries to be sampled from [33].
- *Haphazard sampling*, in which no explicit or structured selection strategy is employed by the auditor. The sampling is conducted without a specific reason for including or excluding journal entries [22].

In general, attribute sampling is utilised by auditors to test the effectiveness of internal controls, e.g. the percentage of vendor invoices that are approved in compliance with the organisation's internal controls, e.g. by following a 'four-eye' approval principle.

2.2 Variable Sampling Techniques

The technique of variable sampling is applied by auditors to estimate the amount or value of specific journal entry characteristics. The different techniques of variable sampling encompass the following classes [21]:

[1] https://www.wegalvanize.com
[2] https://idea.caseware.com

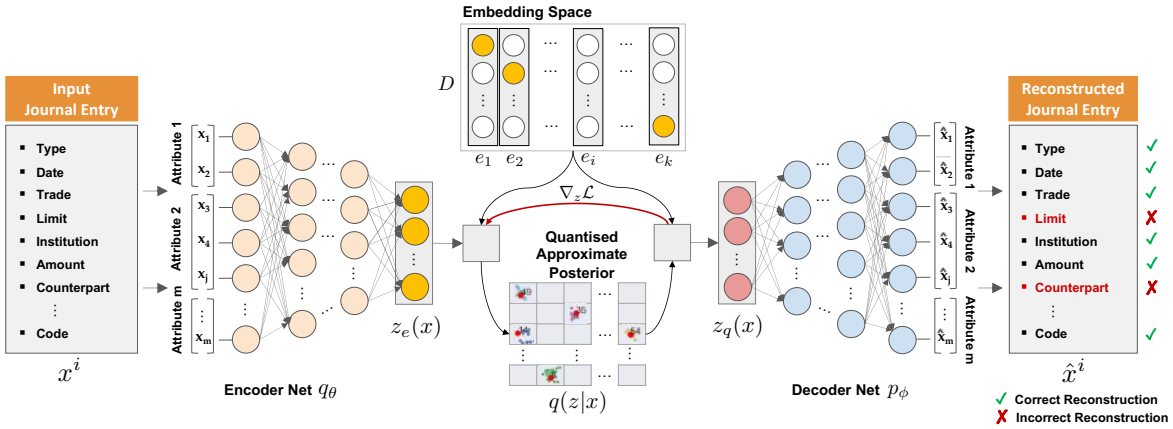

Figure 2: The VQ-VAE architecture [46], applied to learn audit sampling. During the models forward pass, an input journal entry x^i is passed through the encoder network q_θ producing a latent representation z_e. The VQ-VAE quantises z_e using a *codebook* of embedding vectors e_j. Afterwards, the quantised embedding z_q is passed to the decoder network p_ϕ to reconstruct the journal entry $\hat{x}^i$ as faithfully as possible.

- *Difference estimation* calculates the average difference between audited amounts and recorded amounts to estimate the total audited amount of a population [39].
- *Ratio estimation* calculates the ratio of audited amounts to recorded amounts to estimate the total dollar amount of the journal entry population [17].
- *Mean-per-unit estimation* projects the sample average to the total population by multiplying the sample average by the number of items in a journal entry population [35].
- *Monetary-* or *Dollar-unit sampling* considers each monetary unit as a sampling unit of the journal entry population. As a result, entries that record-high posting amounts exhibit a proportionally higher likelihood of being selected [32, 45].

In general, variable sampling is utilised by auditors in the context of substantive testing procedures, e.g. the overstatement of accounts receivables.

To the best of our knowledge, this work presents the first deep-learning inspired approach to learn representative and human interpretable audit samples from real-world accounting data.

3 METHODOLOGY

In this section, we describe the architectural setup of the VQ-VAE model [46] used to learn representative audit sampling. Furthermore, we provide details on the objective function applied to optimise the model parameters.

3.1 Latent Generative Factors

Let X formally be a set of N journal entries $x^1, x^2, ..., x^n$, where each journal entry x^i consists of M accounting specific attributes $x^i_1, x^i_2, ..., x^i_j, ..., x^i_m$. The individual attributes x_j describe the journal entries details, e.g., the entries' fiscal year, posting type, posting date, amount, general-ledger. Following the theoretical assumptions of unsupervised representation learning, [8] we assume that distinct factors of variation generate each entry x^i_j. The different variational factors are not directly observable and correspond to manifolds in a latent space Z. We hypothesise that each generative latent factor $z_i \in Z$ corresponds to a behavioural posting pattern. As a result, it can be uncovered by an unsupervised deep learning algorithm [12, 37].

3.2 VQ-VAE Model

It is often intractable to directly calculate the exact posterior distribution $q(z)$ over the latent generative factors in Z. In [30] Kingma and Welling proposed the Variational Autoencoder (VAE) to learn an approximation of the intractable posterior distribution $q(z|x)$ given the input data X. The VAE's encoder network q_θ learns to parameterise a continuous approximate posterior $q(z|x)$ over the latent factors Z. In parallel, the VAE's decoder network p_ϕ learns to reconstruct the input data X as faithfully as possible with distribution $p(\hat{x}|z)$ over the reconstructions $\hat{X}$. To quantise the approximate posterior $q(z|x)$ and thereby learn a representative audit sample we apply the VQ-VAE model introduced by Van den Oord et al. in [46] and as shown in Fig. 2. In contrast to the VAE, the VQ-VAE model defines a discrete latent embedding space $E \in \mathcal{R}^{KxD}$ where K denotes the size of the space, and D is the dimensionality of each discrete latent embedding vector $e_j \in E$. Due to its discrete nature the embedding space encompasses in total K distinct embedding vectors $e_j \in \mathcal{R}^D$, $j = 1, 2, ..., K$. During the models forward pass an input journal entry x^i is passed through the encoder network q_θ producing a latent representation $z_e(x)$. To obtain its corresponding quantised representation $z_q(x)$, a nearest neighbour lookup is performed, as defined by:

$$z_q(x) = e_k, \text{ where } k = \arg\min_j ||z_e(x) - e_j||_2. \quad (1)$$

where $z_e(x)$ denotes the output of the encoder network and e_j the distinct embedding vectors. This process can be viewed as passing each $z_e(x)$ through a discretisation bottleneck by mapping it onto its nearest embedding in E. Thereby the VQ-VAE autoencoder

quantises the representations $z_e(x)$ using a *codebook* of 1-of-K embedding vectors e_j. Afterwards, the quantised embedding $z_q(x)$ is passed to the decoder. The complete set of parameters of the model are the union of the parameters of the encoder, decoder, and the embedding space. The learned quantised posterior distribution $q(z|x)$ probabilities are defined as 'one-hot', given by:

$$q(z = k|x) = \begin{cases} 1 & \text{for } k = \arg\min_j ||z_e(x) - e_j||_2, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $z_e(x)$ denotes the output of the encoder network and e_j denotes a particular embedding vector.

3.3 VQ-VAE Learning

In the forward pass, to learn a set of quantised embeddings of real-world journal entry data, we compute the model loss $\mathcal{L}$ as defined in Eq. 3. The loss function is comprised of four terms that are optimised in parallel to train the different model components of the VQ-VAE, given by:

$$\begin{aligned} \mathcal{L}(x) = \; & \log p(x|z_q(x)) \; + \; \alpha||sg[z_e(x)] - e||_2^2 \\ & + \; \beta||z_e(x) - sg[e]||_2^2 \; + \; \gamma \log p(x|z_e(x)), \end{aligned} \tag{3}$$

where $z_e(x)$ denotes encoder output, $z_q(x)$ the quantised encoder output and e the set of embedding vectors. The first term denotes the discrete reconstruction loss derived from the reconstruction of the quantised embeddings $z_q(x)$. It encourages the embeddings $z_q(x)$ to be an informative representation of the input [40]. Throughout the training process, the embeddings e receive no gradients from the reconstruction loss. This originates from the straight-through gradient estimation $\nabla_z \mathcal{L}$ of mapping $z_e(x)$ to its nearest neighbor $z_q(x)$. The embeddings are optimized using a vector quantisation technique that applies a *stop-gradient* operator denoted by $sg[\cdot]$. The operator is defined as the identity in the forward pass and has zero partial derivatives in the backward pass [46], denoted by:

$$sg[x] = \begin{cases} x & \text{forward pass ,} \\ 0 & \text{backward pass ,} \end{cases} \tag{4}$$

where x denotes the input vector to the stop-gradient operator.

Both the second and third loss term use stop-gradients. The second term denotes the *embedding loss*, that moves the embeddings e towards the encoder outputs $z_e(x)$. Due to the non-differentiability of the embedding assignment, the embedding space is dimensionless. It can grow arbitrarily if the embeddings e do not train as fast as the encoder parameters θ. Therefore, the third term of Eq. 3 denotes the *commitment loss* which guarantees that the encoder output $z_e(x)$ commits to one of the embeddings. To encourage the encoder output $z_e(x)$ to remain an informative representation and not 'collapsing' towards one of the embeddings we enhanced $\mathcal{L}$ by a fourth loss term as recently introduced by Fortuin et al. in [16]. The added term, denotes the reconstruction loss derived from the encoder outputs $z_e(x)$. The gradient $\nabla_z \mathcal{L}$ of the backward pass is approximated similar to the straight-through estimator presented in [9]. Thereby, the gradients of the decoder input $z_q(x)$ are copied back to the encoder output $z_e(x)$ as shown in Fig. 2. Since the output representation of the encoder and the input to decoder share the same D dimensional space, the gradients contain useful information on how to update the encoder parameters θ to increase the model likelihood. The VQ-VAE can be interpreted a special case of the VAE in which the model's likelihood $\log p(x)$ can be maximised by the optimisation of the *Expectation Lower Bound (ELBO)* [30]. Maximising the ELBO increases the representativeness of the learned quantised representations $z_q(x)$ and therefore mitigates the sampling risk with progressing model training.

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup and model training. Due to the high confidentiality of journal entry data, we evaluate the proposed methodology based on two public available real-world datasets to allow for reproducibility of our results.

4.1 Datasets and Data Preparation

To evaluate the audit sampling capability of the VQ-VAE architecture, we use two publicly available datasets of real-world financial payment data that exhibit high similarity to real-world accounting data. The datasets are referred to as *dataset A* and *dataset B* in the following. Dataset A corresponds to the City of Philadelphia payment data of the fiscal year 2017 [3]. It represents the city's nearly $4.2 billion in payments obtained from almost 60 city offices, departments, boards and committees. Dataset B corresponds to vendor payments of the City of Chicago ranging from 1996 to 2020 [4]. The data is collected from the city's 'Vendor, Contract and Payment Search' and encompasses the procurement of goods and services. The majority of attributes recorded in both datasets (similar to real-world ERP data) correspond to categorical (discrete) variables, e.g. posting date, department, vendor name, document type. We pre-process the original payment line-item attributes to (i) remove of semantically redundant attributes and (ii) obtain a binary ('one-hot' encoded) representation of each payment. The following descriptive statistics summarise both datasets upon successful data pre-processing:

- **Dataset A:** The 'City of Philadelphia' payments encompass a total of $n = 238,894$ payments comprised of 10 categorical and one numerical attribute. The encoding resulted in a total of $8,565$ one-hot encoded dimensions for each of the city's vendor payment record $x^i \in \mathcal{R}^{8,565}$.
- **Dataset B:** The 'City of Chicago' payments encompass a total of $n = 72,814$ payments comprised of 7 categorical and one numerical attribute. The encoding resulted in a total of $2,354$ one-hot encoded dimensions for each of the city's vendor payment record $x^i \in \mathcal{R}^{2,354}$.

4.2 VQ-VAE Training

Our architectural setup follows the VQ-VAE architecture [46] as shown in Fig. 2, comprised of an encoder- and decoder-network as well as an additional embedding layer that are trained in parallel. The encoder network q_θ uses Leaky Rectified Linear Unit (LReLU) activation functions [48] except in the 'bottleneck' layer where no

[3] The dataset is publicly available via: https://www.phila.gov/2019-03-29-philadelphias-initial-release-of-city-payments-data/.
[4] The dataset is publicly available via: https://data.cityofchicago.org/Administration-Finance/Payments/s4vu-giwb/.

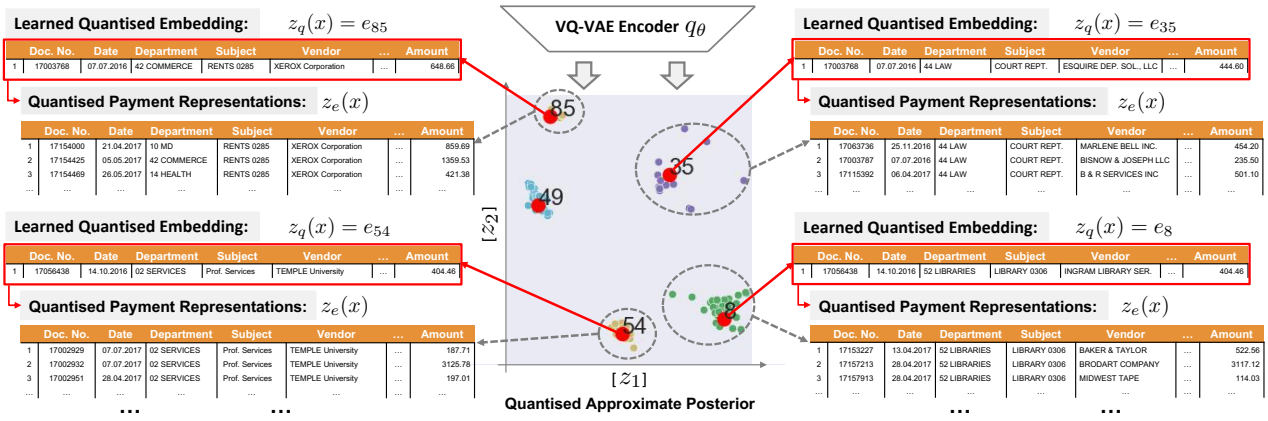

Figure 3: Exemplary VQ-VAE vector quantisation of city payments and corresponding audit samples represented by the models learned embeddings e_k, for $k = \arg\min_j ||z_e(x) - e_j||_2$. For each entry x^i VQ-VAE infers a low-dimensional representation z_e in the latent space Z. The distinct representations z_e are quantised z_q by the embeddings e_k. As a result, the quantisations z_q constitute a set of representative audit samples of the original entry population X.

non-linearity is applied. The decoder network p_θ use LReLUs in all layers except the output layers where sigmoid activations are used. Table 1 depicts the architectural details of the networks which are implemented using PyTorch [36].

Table 1: Number of neurons per layer ℓ of the encoder q_θ and decoder p_ϕ networks that comprise the VQ-VAE architecture [46] used in our experiments.

Net	Dataset	$\ell = 1$	2	3	4	...	10	11
$q_\theta(z\|x)$	A	5,096	2,048	1,024	512	...	4	2
$p_\phi(\hat{x}\|z)$	A	2	4	8	16	...	2,048	5,096
$q_\theta(z\|x)$	B	2048	1024	512	256	...	4	2
$p_\phi(\hat{x}\|z)$	B	2	4	8	16	...	1024	2048

In accordance with [48], we set the scaling factor of the LReLUs to $\alpha = 0.4$. We initialize the parameters of the encoder decoder networks as described in [18]. The embeddings e are initialized by sampling from a uniform prior distribution $e_j \sim \mathcal{U}(-1, 1)$. To allow for interpretation and visual inspection by human auditors we sample each discrete latent embedding vector $e_j \in \mathcal{R}^2$. We evaluate distinct codebook sizes of $K \in \{2^4, 2^5, 2^6, 2^7\}$ embeddings to learn different degrees of payment quantisation. The models are trained with batch wise SGD for a max. of 4,000 training epochs, a mini-batch size of $m = 128$ journal entries, and early stopping once the loss converges. We use Adam optimization [29] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in the optimization of the network parameters. To determine the $z_q(x)$ and $z_e(x)$ reconstruction losses (first and fourth term of Eq. 3) we use a Mean-Squared-Error (MSE) loss, as defined by:

$$\mathcal{L}_{MSE}(x) = ||x - p_\phi(q_\theta(x))||_2^2, \tag{5}$$

where x denotes an encoded journal entry, q_θ the encoder, and p_ϕ the decoder network. Kaiser et al. in [27] observed that training the embeddings, as done on the second term of Eq. 3, can be stabilized. This is achieved by maintaining an Exponential Moving Average (EMA) over (1) the embedding vectors e_j and (2) the count π_j of nearest embeddings $z_q(x)$ mapped onto the individual embedding vectors. Thereby, the count per embedding is defined as $\pi_j = \sum_{i=1}^{N} \mathbb{1}[z_q(x^i) = e_j]$ where $\mathbb{1}$ denotes the indicator function. The EMA count c_j of each embedding is then updated per mini-batch m, as defined by [40]:

$$c_j^{m+1} = \eta\, c_j^m + (1 - \eta)\, \pi_j, \tag{6}$$

with updating each embedding e_j respectively, as follows:

$$e_j^{m+1} = \eta\, e_j^m + (1 - \eta) \sum_{i=1}^{N} \frac{\pi_j\, z_e(x^i)}{c_j^m}, \tag{7}$$

where η denotes the EMA decay parameter. We used this enhancement in all our experiments and set $\eta = 0.95$. Throughout the training, we are also interested in the average number of bits used by a particular model to quantise the latent factors of variation. This is measured by codebook perplexity of the each model, as defined by:

$$\mathcal{P}_{erp}(\pi) = 2^{-\sum_{j=1}^{K} p(\pi_j) \log_2 p(\pi_j)}, \tag{8}$$

the likelihood of a quantised representation $z_q(x)$ of being assigned to a particular embedding e_j. Furthermore, we determine the average purity of all payments ω_j quantised by a particular embedding e_j, as defined by:

$$\mathcal{P}_{urity}(\omega, \pi) = \frac{1}{K} \sum_{j=1}^{K} \frac{|\omega_j|}{\pi_j}, \tag{9}$$

where $\omega_j = \{x^i \in X | z_q(x^i) = e_j\}$. Figure 3 depicts an exemplary VQ-VAE vector quantisation of the city payments and corresponding audit samples represented by the learned embeddings e.

5 EXPERIMENTAL RESULTS

In this section, we assess the latent quantisation learned from real-world city payments quantitatively and qualitatively. Also, we examine the semantic disentanglement of the payments attributes in the latent dimensions in terms of interpretability by a human auditor.

5.1 Quantitative Evaluation

We are interested in the degree of representativeness of the learned quantised embeddings when training VQ-VAE models with varying codebook sizes. The quantitative results obtained for both datasets are shown in Tab. 2. It can be observed that both, the average codebook usage ($\mathcal{P}_{erp}$) and quantisation purity ($\mathcal{P}_{urity}$) increases with larger codebook size. Hence, a more fine-grained quantisation of the latent generative factors is learned. Also, the reconstruction loss derived from the quantised embeddings ($\mathcal{L}^{z_q}_{MSE}$) and the reconstruction loss derived from all embeddings ($\mathcal{L}^{z_e}_{MSE}$) converge with increased codebook size K. This observation corresponds to our initial hypothesis that real-world accounting data is generated by a limited set of latent factors that can be uncovered by an unsupervised deep learning model. The quantitative results indicate that the VQ-VAE provides the ability to learn embeddings that quantise such latent generative factors and therefore constitute a representative audit sample.

Table 2: Reconstruction losses, codebook perplexity, and cluster purity obtained for different codebook size K on both city payment datasets (variances originate from parameter initialization using five random seeds).

Data	K	$\mathcal{L}^{z_q}_{MSE}$	$\mathcal{L}^{z_e}_{MSE}$	$\mathcal{P}_{erp}$	$\mathcal{P}_{urity}$
A	2^3	0.577 ± 0.13	0.453 ± 0.24	5.394 ± 3.81	0.864 ± 0.09
A	2^4	0.454 ± 0.06	0.312 ± 0.34	12.668 ± 0.62	0.845 ± 0.01
A	2^5	0.417 ± 0.10	0.281 ± 0.33	22.024 ± 1.29	0.853 ± 0.02
A	2^6	0.382 ± 0.05	0.232 ± 0.13	37.677 ± 1.72	0.872 ± 0.01
A	2^7	0.345 ± 0.17	0.208 ± 0.19	59.755 ± 2.83	0.890 ± 0.01
B	2^3	1.675 ± 0.03	1.535 ± 0.04	6.082 ± 0.17	0.440 ± 0.03
B	2^4	1.622 ± 0.06	1.424 ± 0.10	9.788 ± 1.12	0.416 ± 0.01
B	2^5	1.587 ± 0.07	1.351 ± 0.15	17.941 ± 2.87	0.377 ± 0.06
B	2^6	1.467 ± 0.02	1.121 ± 0.04	31.217 ± 5.04	0.373 ± 0.01
B	2^7	1.407 ± 0.05	1.071 ± 0.07	42.171 ± 9.39	0.321 ± 0.02

5.2 Qualitative Evaluation

We are also interested in the semantics of the latent generative factors represented by the learned embeddings. Figures 4 and 5 (left) show a quantisation learned by two VQ-VAE models trained with codebook sizes $K = 2^7$ (dataset A) and $K = 2^6$ (dataset B). For each learned embedding, we conduct a qualitative review of the corresponding payment quantisations. The review of the corresponding quantised payments for models with $K = 2^7$ result in the following observations:

- **Dataset A:** Among others, the learned embeddings quantise the payments according to (1) fleet management e_{50} ('auto parts'), (2) legal services e_{52} ('appointed attorneys'), (3) office materials and supplies e_{51} ('Staples Business Advantage'), (4) professional services e_{121} ('consultancy services'), as well as (5) material and supply e_{127} ('fuel and gasoline') payments;
- **Dataset B:** Among others, the learned embeddings quantise the payments according to (1) transportation services e_2 ('public transport maintanance'), (2) family assistance services e_{54} ('homeless financial support'), (3) aviation maintenance e_{17} ('cleaning and fuel'), (4) IT services e_{11} ('software'), (5) water management e_{48} ('pipe supply'), as well as (5) library services e_{60} ('library facilities') payments.

The qualitative results indicate that the embeddings learned by the VQ-VAE semantically quantises the payments generative factors. The representativeness of the embeddings is underpinned by the high quantisation $\mathcal{P}_{urity}$ obtainable with increased codebook size.

Table 3: Disentanglement metrics and scores obtained for both city payment datasets using different codebook sizes K (variances originate from parameter initialization using five random seeds).

Data	K	β-VAE [24]	*Fac*-VAE [28]	MIG [11]	DCI [14]
A	2^3	0.160 ± 0.02	0.110 ± 0.06	0.025 ± 0.02	0.039 ± 0.01
A	2^4	0.166 ± 0.01	0.108 ± 0.06	0.029 ± 0.01	0.038 ± 0.01
A	2^5	0.166 ± 0.03	0.119 ± 0.01	0.031 ± 0.02	0.046 ± 0.01
A	2^6	0.182 ± 0.02	0.134 ± 0.01	0.068 ± 0.08	0.061 ± 0.04
A	2^7	0.193 ± 0.01	0.149 ± 0.03	0.081 ± 0.06	0.139 ± 0.07
B	2^3	0.244 ± 0.01	0.142 ± 0.02	0.051 ± 0.03	0.690 ± 0.03
B	2^4	0.240 ± 0.01	0.145 ± 0.03	0.057 ± 0.03	0.703 ± 0.02
B	2^5	0.277 ± 0.04	0.144 ± 0.02	0.053 ± 0.04	0.709 ± 0.01
B	2^6	0.290 ± 0.02	0.144 ± 0.01	0.067 ± 0.01	0.715 ± 0.01
B	2^7	0.324 ± 0.01	0.146 ± 0.01	0.080 ± 0.03	0.717 ± 0.01

5.3 Disentanglement Evaluation

Finally, we are interested to which extend the individual journal entry attribute characteristics are disentangled in the distinct latent dimensions z_i. A high disentanglement of the journal entry attributes increases the interpretability of the learned quantised embeddings by a human auditor. While there is no formally accepted notion of disentanglement yet, the intuition is that a disentangled representation separates the different informative factors of variation of a dataset [8]. We evaluate four disentanglement metrics, commonly used in unsupervised representation learning [34]:

- The *β-VAE* metric proposed in [24] measures the disentanglement as the accuracy of a linear classifier that predicts the index of the fixed journal entry attribute x_j. We sample batches of 16 representations, trained the classifier on 1,000 batches, and evaluate on 500 batches.
- The *Factor-VAE* proposed in [28] measures the disentanglement as the accuracy of a majority vote classifier to predict the index of the fixed journal entry attribute x_j. We sample batches of 16 representations, trained the classifier on 1,000 batches, and evaluate on 500 batches.
- The *Mutual Information Gap (MIG)* as proposed in [11] measures for each journal entry attribute x_j the two dimensions in $z_e(x)$ that have the highest mutual information with x_j. We bin each dimension in Z into 20 bins and calculated the average mutual information over 1,000 samples.
- The *Disentanglement Metric (DCI)* proposed in [14] computes the entropy of the distribution obtained by normalizing the importance of each dimension in $z_e(x)$ for predicting the

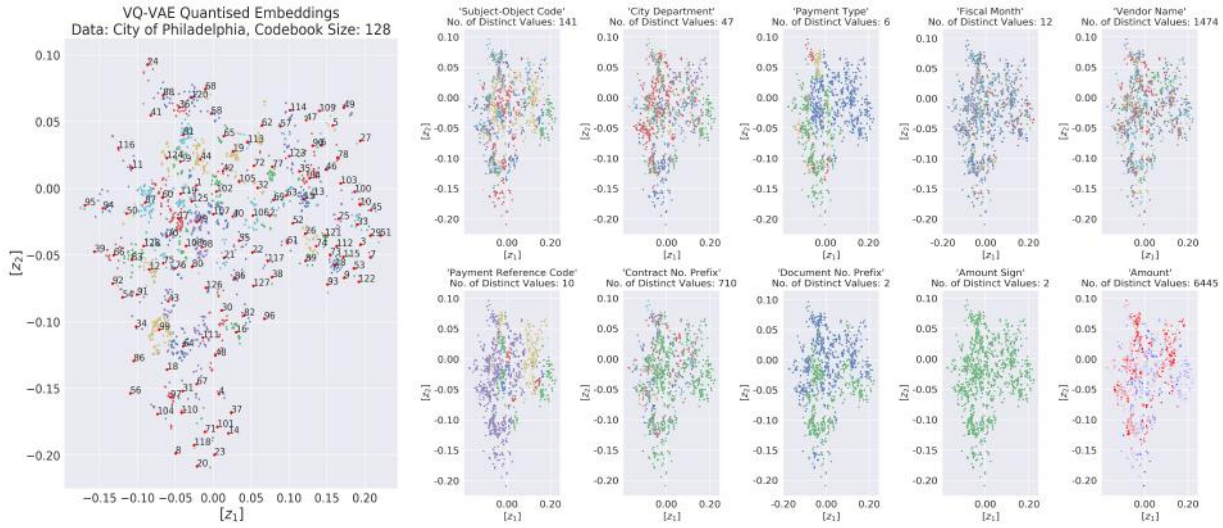

Figure 4: Exemplary quantised latent representations z_e (colored by quantisation) and embeddings e (numbered red circles) in $\mathcal{R}^2$ learned by the VQ-VAE of the 238,894 'City of Philadelphia' vendor payments (dataset A) using a codebook size of $K = 2^7$ (128) embeddings (left). Learned disentanglement of the distinct payment attributes x_j (e.g., 'subject-object code', 'payment type', 'fiscal month') by the learned representations z_e (colored by attribute value) in the latent dimensions z_i (middle-right).

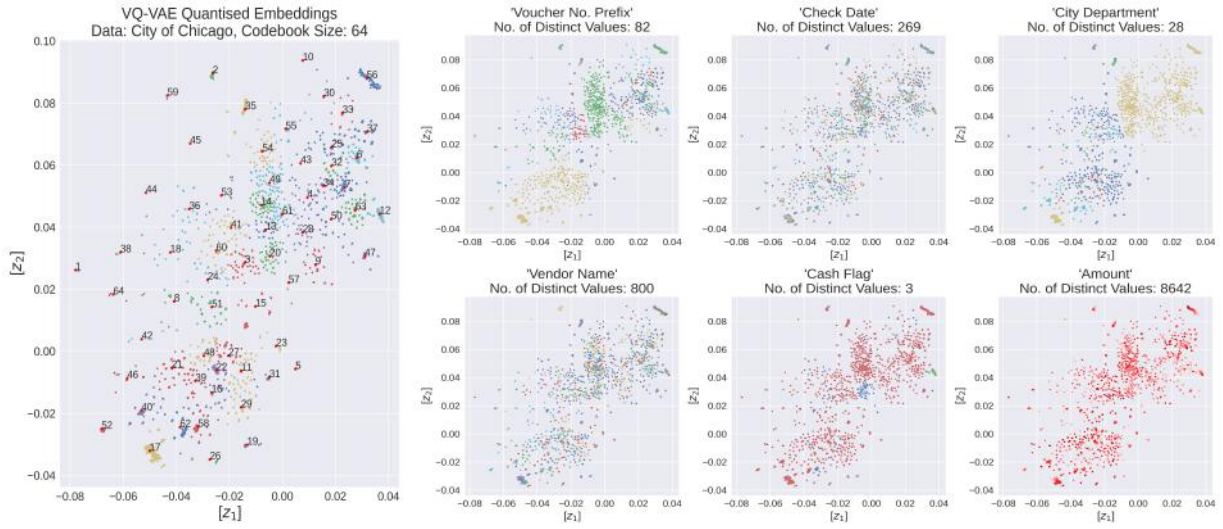

Figure 5: Exemplary quantised latent representations z_e (colored by quantisation) and embeddings e (numbered red circles) in $\mathcal{R}^2$ learned by the VQ-VAE of the 72,814 'City of Chicago' vendor payments (dataset B) using a codebook size of $K = 2^6$ (64) embeddings (left). Learned disentanglement of the distinct payment attributes x_j (e.g., 'city department', 'vendor name', 'amount') by the learned representations z_e (colored by attribute value) in the latent dimensions z_i (middle-right).

value of an attribute x_j. We use a decision tree, trained the classifier on 1,000 batches, and evaluate on 500 batches.

The disentanglement scores obtained for the distinct metrics are presented in table 3. It can be observed that increasing the codebook size K yield an increased disentanglement of the journal entry attributes in the latent dimension of Z. Figure 4 and 5 (middle-right) illustrate the learned disentanglement of the distinct payment attributes, e.g. city department, payment type, fiscal month. It can be observed that the learned quantised posterior disentangles the payment attributes to allow for a highly explainable audit sampling.

6 SUMMARY

In this work, we proposed a deep learning inspired approach to conduct audit sampling in the context of financial statement audits. We showed that Vector Quantised-Variational Autoencoder (VQ-VAE) neural networks could be trained to learn a quantised

representation of financial payment data recorded in ERP systems. We demonstrated, based on two real-world datasets of city payments, that such a learned quantisation corresponds to the latent generative factors in both datasets. Our experimental results provide initial evidence that VQ-VAE's can be utilised in the context of representative audit sampling and therefore offer the ability to reduce sampling risks. Furthermore, the learned representation allows for human interpretable discrete sampling. We hope that this technique will enhance the toolbox of auditors in the near future to sample journal entries from large-scale financial accounting data. Given the tremendous amount of journal entries recorded by organisations, deep-learning-based sampling techniques can provide a starting point for a variety of further downstream audit tasks.

ACKNOWLEDGEMENTS

We thank the members of the statistics department at Deutsche Bundesbank for their valuable review and remarks. Opinions expressed in this work are solely those of the authors and do not necessarily reflect the view of the Deutsche Bundesbank nor PricewaterhouseCoopers (PwC) International Ltd. and its network firms.

REFERENCES

[1] *Consideration of Fraud in a Financial Statement Audit, AU Section 316.* American Institute of Certified Public Accountants (AICPA), 2002.
[2] *International Accounting Standard (ISA) 1 - Presentation of Financial Statements.* International Federation of Accountants (IFAC), 2007.
[3] *Practice Aid for Testing Journal Entries and Other Adjustments Pursuant to AU Section 316.* Center for Audit Quality (CAQ), 2008.
[4] *International Standard on Auditing (ISA) 530, Audit sampling and Other Means of Testing.* International Federation of Accountants (IFAC), 2009.
[5] *International Standards on Auditing 240, The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements.* International Federation of Accountants (IFAC), 2009.
[6] *Audit Sampling, AU-C Section 530.* American Institute of Certified Public Accountants (AICPA), 2012.
[7] A. D. Akresh, J. K. Loebbecke, and W. R. Scott. Audit approaches and techniques. *Research Opportunities in Auditing: The Second Decade*, 13, 1988.
[8] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
[9] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
[10] I. Bhattacharya and E. Roos Lindgreen. A semi-supervised machine learning approach to detect anomalies in big accounting data. In *Proceedings of the European Conference on Information Systems (ECIS)*, 2020.
[11] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.
[12] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.
[13] M. L. De Prado. *Advances in financial machine learning.* John Wiley & Sons, 2018.
[14] C. Eastwood and C. K. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.
[15] R. J. Elder, A. D. Akresh, S. M. Glover, J. L. Higgs, and J. Liljegren. Audit sampling research: A synthesis and implications for future research. *Auditing: A Journal of Practice & Theory*, 32(1):99–129, 2013.
[16] V. Fortuin, M. Hüser, F. Locatello, H. Strathmann, and G. Rätsch. Som-vae: Interpretable discrete representation learning on time series. *arXiv preprint arXiv:1806.02199*, 2018.
[17] S. J. Garstka and P. A. Ohlson. Ratio estimation in accounting populations with probabilities of sample selection proportional to size of book values. *Journal of Accounting Research*, pages 23–59, 1979.
[18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.
[19] S. V. Grabski, S. A. Leech, and P. J. Schmidt. A review of erp research: A future agenda for accounting information systems. *Journal of information systems*, 25(1):37–78, 2011.
[20] D. M. Guy, D. R. Carmichael, and R. Whittington. *Practitioner's guide to audit sampling.* Wiley, 1998.
[21] D. M. Guy, D. R. Carmichael, and R. Whittington. *Audit sampling: An introduction (5th edition).* John Wiley & Sons Inc, 2002.
[22] T. W. Hall, A. W. Higson, B. J. Pierce, K. H. Price, and C. J. Skousen. Haphazard sampling: selection biases and the estimation consequences of these biases. *Current Issues in Auditing*, 7(2):P16–P22, 2013.
[23] T. W. Hall, J. E. Hunton, and B. J. Pierce. Sampling practices of auditors in public accounting, industry, and government. *Accounting Horizons*, 16(2):125–136, 2002.
[24] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.
[25] N. B. Hitzig. Audit sampling: A survey of current practice. *The CPA Journal*, 65(7):54, 1995.
[26] G. Jokovich. Statistical sampling in auditing. *International Journal of Accounting and Financial Management*, 16:892–898, 2013.
[27] Ł. Kaiser, A. Roy, A. Vaswani, N. Parmar, S. Bengio, J. Uszkoreit, and N. Shazeer. Fast decoding in sequence models using discrete latent variables. *arXiv preprint arXiv:1803.03382*, 2018.
[28] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.
[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[30] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
[31] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.
[32] D. A. Leslie, A. D. Teitlebaum, and R. J. Anderson. *Dollar-unit sampling: a practical guide for auditors.* Copp Clark Pitman; Belmont, Calif.: distributed by Fearon-Pitman, 1979.
[33] Y. Liu, M. Batcher, and F. Scheuren. Efficient sampling design in audit data. *Journal of Data Science*, 3:213–222, 2005.
[34] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.
[35] J. Neter and J. K. Loebbecke. *Behavior of major statistical estimators in sampling accounting populations: An Empirical Study.* American Institute of Certified Public Accountants (AICPA), 1975.
[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.
[37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
[38] M. Reichert and B. Weber. Ad hoc changes of process instances. In *Enabling Flexibility in Process-Aware Information Systems*, pages 153–217. Springer, 2012.
[39] D. M. Roberts. *Statistical Auditing.* American Institute of Certified Public Accountants (AICPA), 1978.
[40] A. Roy, A. Vaswani, N. Parmar, and A. Neelakantan. Towards a better understanding of vector quantized autoencoders. 2018.
[41] M. Schreyer, T. Sattarov, B. Reimer, and D. Borth. Adversarial learning of deepfakes in accounting. *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy, Vancouver, BC, Canada*, 2019.
[42] M. Schreyer, T. Sattarov, C. Schulze, B. Reimer, and D. Borth. Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks. *2nd KDD Workshop on Anomaly Detection in Finance, Anchorage, Alaska, USA*, 2019.
[43] M. Schultz and M. Tropmann-Frick. Autoencoder neural networks versus external auditors: Detecting unusual journal entries in financial statement audits. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.
[44] D. A. Schwartz. Computerized audit sampling. *The CPA Journal*, 68(11):46, 1998.
[45] K. W. Stringer. Practical aspects of statistical sampling in auditing. In *Proceedings of the Business and Economic Statistics Section*, pages 405–411. American Statistical Association, 1963.
[46] A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.
[47] M. Wiese, R. Knobloch, R. Korn, and P. Kretschmer. Quant gans: deep generation of financial time series. *Quantitative Finance*, pages 1–22, 2020.
[48] B. Xu, N. Wang, T. Chen, and M. Li. Empirical Evaluation of Rectified Activations in Convolution Network. pages 1–5, 2015.