# Improving automatic data extraction from financial statements with clustering analysis

V. V. Ferraz[1], G. Olivato[2], I. R. Magollo[2], M. C. Naldi[2]

[1] Serasa Experian
victor.ferraz@br.experian.com
[2] Universidade Federal de São Carlos, Brazil
{gabrielolivato,igor.magollo}@estudante.ufscar.br,naldi@ufscar.br

**Abstract.** The financial statement analysis is a fundamental part of the credit risk attribution process, producing documents that are valuable sources of information about companies' economic and financial wealth. Large volumes of that type of document demand automatic data extraction, and locators drive the tools for that task. However, due to the lack of regulation, there is not a standard layout for such documents, which originates a variety of document structures. Such variety burdens the feature extraction tools, reducing their performance. Clustering analysis overcomes such burden by finding the best document clusters, allowing the development of fine-tuned locators for each cluster based on their main characteristics, which is the main objective of this work. We applied state-of-the-art clustering techniques, RNG-HDBSCAN*, FOSC and MustaCHE, over financial statements documents to assess their clusters and main structures, separate outliers, and analyze their main features. The result allows the specialists to define proper locators for each cluster, increasing the performance of the data extraction tools.

CCS Concepts: • **Applied computing**; • **Computing methodologies** → **Artificial intelligence**;

Keywords: data science, clustering, feature extraction

## 1. INTRODUÇÃO

Desde a década passada, segundo a Lei das Sociedades por Ações (11.638/07 e 11.941/09), as empresas devem elaborar suas demonstrações contábeis conforme as normas contábeis brasileiras de elaboração convergentes aos padrões internacionais [Brasil 2009; 2007; 1976]. Tais normas determinam as peças contábeis, assim como orientam o processo de classificação e ordenação de contas. Porém, as normas não definem uma estrutura (*layout*) de relatório a ser adotado pelas empresas [Comitê de Pronunciamentos Contábeis 2011b; Banco Central do Brasil 2010], permitindo que contadores e desenvolvedores de sistemas contábeis estruturem as demonstrações conforme suas necessidades.

Fonte valiosa de informações sobre a situação econômico-financeira das empresas, a análise das demonstrações contábeis é parte fundamental dos processos de atribuição do risco de crédito [Assaf Neto 2020]. A fim de agilizar tal análise, foram propostas soluções tecnológicas para extração automática de dados, através de reconhecimento óptico de caracteres (*OCR*) e localizadores de chaves e valores no texto, como descrições de contas e saldos. Entretanto, tais soluções são eficientes quando os documentos não apresentam alta variabilidade estrutural, possibilitando aplicar o localizador mais apropriado. Em particular, devido à falta de informação estrutural precisa, a Serasa Experian adotou um localizador "genérico" sobre 85% do volume de documentos, o que limitou a eficiência da extração. O ideal seria possuir localizadores adequados para cada um dos *layouts* dos documentos. Entretanto, não se sabe à *priori* quantos *layouts* encontram-se no conjunto de documentos, nem mesmo se existem documentos com *layouts* semelhantes. Tal informação ajudaria muito o desenvolvimento de localiza-

dores apropriados para cada conjunto de documento com *layouts* parecidos. O agrupamento de dados consiste em dividir os dados de forma que elementos correlacionados estejam em um mesmo grupo. Adicionalmente é possível detectar elementos que não se encaixam em nenhum grupo, chamados de externos *outliers*. O agrupamento de documentos com *layouts* correlacionados permite uma análise dos grupos obtidos, de forma que seja possível gera localizadores específicos para suas características.

Esse trabalho possuiu como principal objetivo o agrupamento de demonstrações contábeis para fins da extração automática de descrição de contas e saldos contábeis. Baseia-se na premissa de que documentos semelhantes possuam estruturas com características muito próximas e, por isso, viabilizam a melhoria do processo de detecção e extração dos dados. Para isso, usamos um conjunto de técnicas estado da arte para análise de grupos, o que ainda não foi feito em trabalhos de aplicações anteriores. RNG-HDBSCAN* [Araujo Neto et al. 2019], capaz de obter inúmeras hierarquias de HDBSCAN* em uma única execução, o arcabouço FOSC [Campello et al. 2013] e a ferramenta de visualização MustaCHE [Araujo Neto et al. 2018], que permitem a obtenção e análise de partições multiníveis para cada hierarquia obtida, resultando em uma análise robusta da estrutura dos dados e independente de parâmetros comuns para outros algoritmos, como a pré-definição do número de grupos, possíveis inicializações ou mesmo um valor mínimo de densidade. Por meio da análise apresentada nesse trabalho, é possível selecionar os grupos de maior interesse entre a alta variedade de modelos de documentos contábeis e implementar localizadores específicos para extração de informação precisa de cada grupo, aumentando a qualidade e eficiência do processo de extração automática. Adicionalmente, a técnica aplicada permite a detecção de documentos externos (*outliers*).

## 2. TRABALHOS RELACIONADOS

O aumento do volume de dados gerados levou à necessidade de classificação, organização e extração de conhecimento dos mais diversos tipos de documentos. [Moura 2004] analisou empiricamente uma série de ferramentas de mineração para melhorar a automação do processo de identificação, seleção e classificação de conteúdo relevante para a Agência de Informação Embrapa. Em sua proposta estabeleceu-se o uso de ferramentas para obtenção de grupos que auxiliem na classificação do conteúdo, posteriormente validados por "uma parceria com um especialista do domínio escolhido" [Moura 2004].

[Madeira 2015] relatou resultados favoráveis ao aplicar $k$-médias [Jain 2010] para identificar possíveis empresas alvo de fiscalização tributária no município do Rio de Janeiro, através de dados extraídos das Notas Fiscais de Serviços Eletrônicas (NFS-e). Mais recentemente [Snow 2018], dados extraídos de formulários enviados por empresas inglesas ao Companies House, órgão oficial de registro das empresas britânicas, foram agrupados e analisados para gerar informação relevante para o processo de definição do código *SIC* (*Standard Industrial Classification*), sistema de classificação das empresas por ramo de atividade. O modelo conseguiu identificar erros de classificação e novas tendências de atividades empresariais através da vetorização do conteúdo textual referente à descrição da atividade informada pelas próprias empresas, adicionada da incorporação bidimensional dos dados para visualização intuitiva e agrupamento hierárquico dos dados baseado em densidade, com uso do HDBSCAN*.

## 3. METODOLOGIA

### 3.1 Caracterização e Processamento de Demonstrativos Contábeis

Dentre os principais demonstrativos obrigatórios que compõem um relatório financeiro, estão o balanço patrimonial e a demonstração do resultado do exercício. Essas peças contábeis provêm as informações necessárias para a maioria das análises financeiras [Assaf Neto 2020] e, portanto, são o principal alvo do processo de extração automática. Com base em suas posições e composições, foram definidas 12 características dos relatórios financeiros que remetem à sua estrutura, identificadas nos documentos financeiros com uso de aplicação de extração automática.

Para efetuar a extração foi implementado um localizador que, após o processo de *OCR* sobre os textos originais, processou um conjunto de regras de decisão, elaboradas em conjunto com os especialistas em risco de crédito, para definir os valores dessas características. O localizador utiliza objetos que a aplicação oferece sobre o documento (páginas, linhas de texto, palavras etc.), cada qual com uma série de métodos e propriedades como, por exemplo, no caso de uma palavra, é possível saber: a distância das margens e do topo em *pixels*, o número da página, se ela é uma palavra-chave, dentre outras possibilidades. Nesse trabalho, as características com seus respectivos pesos são apresentadas na Tabela I. O peso determina a relevância de uma característica para a análise dos agrupamentos, segundo os especialistas em risco de crédito da Serasa Experian.

| Nº. | Peso | Tipo | Descrição |
|---|---|---|---|
| 1 | 16% | Categórica | Posição do ativo em relação ao passivo |
| 2 | 7% | Booleana | Indica se o grupo de ativos de longo prazo e permanentes existe |
| 3 | 9% | Booleana | Indica se o grupo de passivos de longo prazo existe |
| 4 | 1% | Booleana | Indica se a demonstração do resultado do exercício existe |
| 5 | 9% | Categórica | Indica o tipo: balanço ou balancete |
| 6 | 12% | Booleana | Indica se os saldos das contas estão desdobrados |
| 7 | 10% | Percentual | Proporção de números em relação ao total de termos |
| 8 | 12% | Inteiro | Quantidade de colunas com saldos e/ou movimentação de contas |
| 9 | 8% | Inteiro | Quantidade de páginas do ativo |
| 10 | 8% | Inteiro | Quantidade de páginas do passivo |
| 11 | 8% | Inteiro | Quantidade de páginas do documento |
| 12 | 0% | Booleana | Indica se o modelo já é padronizado SPED |

Table I: Características estruturais dos relatórios financeiros.

De acordo com [Assaf Neto 2020], o balanço patrimonial é o "elemento de partida indispensável para o conhecimento da situação econômica e financeira de uma empresa" e seu conceito ("balanço") tem origem no equilíbrio entre as partes que o compõem: ativo, passivo e patrimônio líquido[1]. A forte relação conceitual entre ativo e passivo reflete em contas com descrições muito semelhantes, frequentemente observadas nos documentos financeiros. Essa relação dificulta a parametrização do método de extração, uma vez que as orientações são passadas através de palavras-chave, coordenadas cardeais e colaterais referentes aos dados. Por exemplo, uma conta pode apresentar o termo "circulante" ao seu norte (acima no texto da mesma página), que nomeia tanto o grupo de ativos, quanto de passivos de curto prazo. Nessas situações, é necessário identificar o grupo correto para que a conta receba a classificação apropriada. Posto isso, definiu-se a característica 1, a mais relevante a ser identificada no documento financeiro: a posição do ativo em relação ao passivo, com três situações: ativo à esquerda e passivo à direita, lado a lado na mesma página; ativo acima do passivo, na mesma página; e ativo e passivo em páginas diferentes.

Conforme definido pelo [Comitê de Pronunciamentos Contábeis 2011b] a empresa deve apresentar ativos e passivos circulantes e não circulantes como grupos separados no balanço patrimonial. Isso reflete na ocorrência de contas com descrições idênticas e classificações diferentes, comumente observadas nas demonstrações contábeis, conflitando a parametrização através de palavras-chave e coordenadas. Por exemplo, um passivo pode apresentar uma conta com a descrição "contas a pagar" tanto no grupo circulante, quanto no não circulante. Esse fator é definido nas características 2 e 3.

A demonstração do resultado do exercício (*DRE*) contém o cálculo do resultado (lucro ou prejuízo) auferido pela empresa em um período. Ele é composto pelas contas de receitas, despesas e custos em uma sequência esquematizada [Assaf Neto 2020]. A *DRE* apresenta conteúdo bem distinto do restante de um documento, o que minimiza conflitos na extração automática. Uma parcela pequena de relatórios financeiros são incompletos, contendo somente o balanço patrimonial. Sendo assim, a característica 4 indica a presença da *DRE* no documento.

[1]O passivo e o patrimônio líquido são comumente apresentados em conjunto e denominados simplesmente como passivo.

O [Comitê de Pronunciamentos Contábeis 2011a] estabelece o conteúdo mínimo das demonstrações contábeis intermediárias, ou seja, aquelas que são elaboradas no decorrer do exercício e, portanto, se referem à parte do ano fiscal. Essas demonstrações, denominadas balancetes, são elaboradas para fins gerencias e de acompanhamento do desempenho da empresa e, por isso, costumam apresentar alto grau de detalhamento dos dados. Isso aumenta significativamente a complexidade do processo de extração e, portanto, balanços e balancetes são indicados pela característica 5.

Relatórios financeiros mais complexos, que demandam maior quantidade de regras específicas na parametrização do localizador de características, geralmente apresentam:

(1) Demonstrações contábeis comparativas entre períodos distintos;
(2) Demonstração da movimentação dos saldos (saldo inicial, créditos, débitos e saldo final);
(3) Desdobramento dos saldos das contas;
(4) Alta proporção de termos numéricos em relação ao total de termos na página inicial do balanço;
(5) Elevada quantidade de páginas como um todo, e/ou especificamente nos grupos ativo e/ou passivo.

Tais características foram definidas de 6 a 11. Em particular, as características 6, 7 e 8 consideraram a probabilidade de alta correlação e tendência a médias superiores para o conjunto de documentos do tipo balancete. A característica 12 identificou quais documentos já apresentavam o único modelo padronizado de relatório financeiro considerado até então: o modelo SPED (Sistema Público de Escrituração Digital), com o único objetivo de excluir esses casos da amostra (uma vez que são padronizados e a extração automática é facilitada). Adicionalmente também foram excluídos da amostra os vetores de características duplicados e documentos para os quais não foram extraídas todas as características. Por fim, o conjunto final amostrado possui 1.492 vetores de características de documentos financeiros.

### 3.2 HDBSCAN*

O algoritmo HDBSCAN* [Campello et al. 2013] possui diversas vantagens comparado a outros algoritmos particionais e hierárquicos tradicionais. Combina os aspectos de agrupamento baseado em densidade e hierárquico, onde suas hierarquias são construídas pelas densidades dos grupos, dos quais podem ser extraídos os mais proeminentes. Seus resultados podem ser visualizados através de um dendrograma, uma árvore de agrupamento simplificada e outras técnicas de visualização que não necessitam de nenhum parâmetro crítico como entrada. O algoritmo HDBSCAN* recebe apenas um parâmetro de entrada, $m_{pts}$, que pode ser entendido como um fator de suavização de densidade não paramétrico realizado pelo algoritmo. HDBSCAN* também é flexível na análise de seus resultados, deixando o usuário analisar diretamente a hierarquia e a árvore de agrupamentos diretamente ou até realizar cortes na árvore para obter um particionamento dos dados equivalente ao DBSCAN.

Dado um conjunto de dados $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ com $n$ objetos e um valor de suavização $m_{pts}$, as seguintes definições são utilizadas pelo HDBSCAN*:

—*Core Distance*: a distância de núcleo de um objeto $\mathbf{x}_p \in \mathbf{X}$ para $m_{pts}$, $d_{core}(\mathbf{x}_p)$ é a distância de $\mathbf{x}_p$ até o seu $m_{pts}$-ésimo vizinho mais próximo (incluindo $\mathbf{x}_p$), ou seja, o raio mínimo $\epsilon$ em que $\mathbf{x}_p$ é considerado um objeto denso (*core object*).
—*Mutual Reachability Distance*: a distância de alcance mútuo entre dois objetos $\mathbf{x}_p$, $\mathbf{x}_q \in \mathbf{X}$ para $m_{pts}$ é definida como $d_{m_{reach}}(\mathbf{x}_p, \mathbf{x}_q) = max(d_{core}(\mathbf{x}_p), d_{core}(\mathbf{x}_q), d(\mathbf{x}_p, \mathbf{x}_q))$ e pode ser interpretada como um raio mínimo $\epsilon$ tal que ambos os objetos são densos e estão dentro da $\epsilon$-vizinhança um do outro.
—*Mutual Reachability Graph*: o grafo de alcance mútuo de um conjunto de dados $\mathbf{X}$ para $m_{pts}$ é um grafo completo e ponderado, $G_{m_{reach}}$, em que os objetos são os vértices e o peso de cada aresta entre cada par de vértices é dados por sua *mutual reachability distance*.

Os passos principais do HDBSCAN* estão descritos no Algoritmo 1 [Campello et al. 2013]. Com seu resultado é possível realizar análise de agrupamentos, detecção de ruído e visualização dos dados.

**Algorithm 1:** HDBSCAN*

**Data:** $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, $m_{pts}$
**Result:** Hierarquia HDBSCAN*

(1) Dado um conjunto de dados $\mathbf{X}$, calcular as *core distances* de todos os seus objetos.
(2) Calcular a árvore de abrangência mínima *MST*, sobre o *mutual reachability graph*, $G_{m_{reach}}$.
(3) Estender a *MST* com laços a cada vértice com peso iguais ao seu *core distance*, obtendo $MST_{ext}$.
(4) Extrair a hierarquia HDBSCAN* como um dendrograma da $MST_{ext}$.
   (a) Todos os objetos são definidos com o mesmo rótulo (a raíz da árvore de grupos).
   (b) Iterativamente remover arestas da $MST_{ext}$ em ordem decrescente de pesos.
      i. Arestas com o mesmo peso são removidas simultaneamente.
      ii. Após a remoção de uma aresta, os rótulo do agrupamento são designados aos dois componentes conexos que contém um vértice da aresta removida. Um novo rótulo é adicionado se o componente tem pelo menos uma aresta restante,. Caso contrário, os objetos no componente são rotulados como externos (*outliers*).

HDBSCAN* necessita do número de objetos mínimo, chamado de $m_{pts}$, para uma região ser considerada "densa", e sua mais recente versão, o RNG-HDBSCAN* [Araujo Neto et al. 2019], calcula de forma eficiente todas as hierarquias para diferentes valores em um intervalo $[min, max]$ de $m_{pts}$ definido. Outra vantagem da utilização do HDBSCAN* no contexto empresarial é ser relacional, ou seja, o algoritmo necessita apenas das relações (similaridades) entre os objetos a serem agrupados, mas não dos objetos propriamente ditos. Portanto, algoritmos relacionais podem garantir a confidencialidade dos dados. Adicionalmente, é possível analisar visualmente todas as hierarquias geradas por meio do MustaCHE [Araujo Neto et al. 2018], uma poderosa ferramenta de visualização que permite a análise de múltiplas hierarquias de agrupamentos baseados em densidade gerados com o HDBSCAN*. Através de gráficos interativos, ele permite a comparação simultânea dos resultados para diversos valores de $m_{pts}$. O MustaCHE utiliza o índice de acordo hierárquico (*HAI*) [Johnson et al. 2013] para relacionar os agrupamentos de acordo com suas semelhanças, podendo fundir os agrupamentos obtidos utilizando o próprio HDBSCAN*, o que dá origem a um meta-agrupamento.

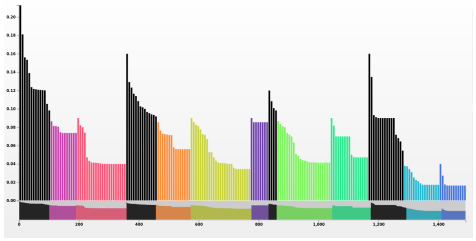
## 4. ANÁLISE DOS RESULTADOS

A execução do RNG-HDBSCAN*, seguida da análise dos resultados por meio do MustaCHE, três partições geradas com diferentes valores de $m_{pts}$ foram escolhidas porque possuíam menos de 10 grupos e maior homogeneidade dos valores de alcance mútuo entre documentos do mesmo grupo. A Tabela II resume os resultados e a Figura 1 mostra os gráficos de alcance mútuo para valores de $m_{pts}$.

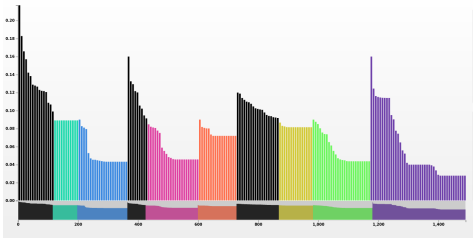
| $m_{pts}$ | 58 | 81 | 95 |
|---|---|---|---|
| Quantidade de grupos | 9 | 7 | 6 |
| Percentual de externos | 23 | 21 | 16 |

Table II: Quantidade de grupos e percentual de externos *outliers* por $m_{pts}$.

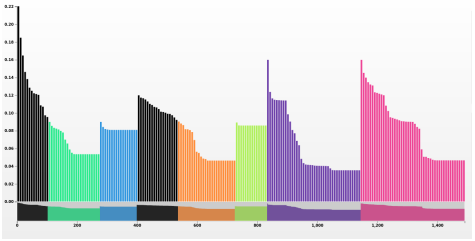
Os gráficos de alcance (Figura 1) mostram que grupos homogêneos são mais estáveis, pois há uma baixa variação no seu nível de densidade mútua. Os "vales" são as regiões mais densas de cada grupo, e no geral, eles contém os documentos mais semelhantes com o restante de seu grupo. Ao aumentar o parâmetro $m_{pts}$ a quantidade de grupos diminui, assim como os externos (*outliers*), de cor preta.



(a) $m_{pts} = 58$.



(b) $m_{pts} = 81$.



(c) $m_{pts} = 95$.

Fig. 1: Visualização dos grupos nos gráficos de alcance mútuo gerados a partir de diferentes valores de $m_{pts}$.

Os resultados foram analisados pelos especialistas em risco de crédito da Serasa Experian. A posição do ativo em relação ao passivo foi a característica determinante na definição dos resultados, posto que em todos os grupos formados, todos os documentos apresentaram valores iguais. Documentos com ativo e passivo em páginas diferentes foram predominantes, em média 42% dos casos agrupados para cada $m_{pts}$, formando grupos exclusivos. Com 95 e 81 $m_{pts}$, foram formados 4 grupos de documentos nessa situação, distintos entre si pela presença de contas do passivo de longo prazo e saldos desdobrados e, ainda, pela quantidade de colunas com saldos de contas e/ou pela quantidade de páginas do ativo e do passivo. Com 58 $m_{pts}$, foram formados 5 grupos de documentos com ativo e passivo em páginas diferentes e, além dos padrões identificados acima, notou-se que um dos grupos continha somente balanços e outro somente balancetes. Também notou-se influência dessas características na distinção dos agrupamentos de documentos com ativo à esquerda e passivo à direita, lado a lado na mesma página, e com ativo acima do passivo na mesma página.

A tabela III traz um quadro com o resumo das características predominantes em cada grupo formado com 58 $m_{pts}$, dado o detalhamento e grau de distinção obtido entre os documentos agrupados com esse valor. Os grupos foram identificados com cores de acordo com o gráfico da figura 1 e a coluna "%" refere-se à proporção de objetos em relação à amostra. Nota-se a predominância de agrupamentos com documentos com ativo e passivo em páginas diferentes e as demais características evidenciam os padrões de distinção descritos anteriormente. A coluna 5 exibe o percentual de documentos do tipo balanço e as demais colunas trazem o valores médios de cada uma das características, observados para o respectivo agrupamento.

| Característica | % | Posição do ativo e do passivo | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Azul Claro | 8 | lado a lado, mesma página | 100% | 100% | 71% | 100% | 0% | 31% | 2.00 | 1.00 | 1.00 | 7.79 |
| Azul | 5 | lado a lado, mesma página | 100% | 100% | 82% | 100% | 0% | 25% | 1.00 | 1.27 | 1.27 | 5.36 |
| Lilás | 6 | ativo acima, mesma página | 20% | 0% | 70% | 100% | 0% | 30% | 1.20 | 1.00 | 1.10 | 8.50 |
| Rosa | 11 | ativo acima, mesma página | 100% | 100% | 90% | 100% | 0% | 31% | 2.10 | 1.00 | 1.20 | 9.60 |
| Roxo | 4 | páginas diferentes | 100% | 100% | 71% | 0% | 100% | 53% | 3.86 | 3.71 | 3.71 | 14.29 |
| Verde musgo | 14 | páginas diferentes | 92% | 100% | 67% | 100% | 100% | 31% | 1.67 | 2.79 | 2.13 | 11.13 |
| Laranja | 8 | páginas diferentes | 64% | 0% | 79% | 100% | 100% | 33% | 1.50 | 2.00 | 1.50 | 8.64 |
| Verde Claro | 12 | páginas diferentes | 91% | 100% | 91% | 100% | 0% | 35% | 2.23 | 1.09 | 1.14 | 21.09 |
| Verde | 9 | páginas diferentes | 47% | 0% | 100% | 100% | 0% | 34% | 1.40 | 1.00 | 1.33 | 6.80 |
| **Média** | - | - | **82%** | **72%** | **81%** | **95%** | **33%** | **33%** | **1.83** | **1.59** | **1.50** | **10.94** |

Table III: Análise para 58 $m_{pts}$.

Além do levantamento estatístico dos vetores de características dos documentos agrupados, também foram resgatados os arquivos originais com os relatórios financeiros em uma quantidade equivalente a 12% da amostra. Esses foram examinados quanto à similaridade visual e complexidade técnica pelos especialistas em análise de risco de crédito de empresas. Para esse processo, os relatórios foram organizados conforme os resultados do agrupamento e examinados comparativamente em conjunto.

O exame dos arquivos originais também apresentou conclusões favoráveis aos resultados do agrupamento, para todos os valores de $m_{pts}$. Em média, 88% dos documentos de um mesmo grupo apresentaram alto grau de similaridade visual e técnica. Os especialistas apontaram poucas exceções (cerca de 12%), para as quais foram identificadas inconsistências nos valores de determinadas características, dadas situações não previstas pela árvore de decisões do *script* de localização.

Para validar os agrupamentos no processo de extração automática de dados, os especialistas selecionaram o documento mais representativo de 4 agrupamentos com características distintas. A qualidade da extração automática dos dados do ativo e do passivo desses documentos foi auferida através do processo atual de extração e, então, foram apurados os erros e dados não extraídos passíveis de tratamento, considerado o conhecimento prévio de características de acordo com o agrupamento. Tal análise permitiu estimar o potencial de melhoria da qualidade da extração automática dos dados desses documentos, demonstrado na tabela abaixo. A coluna Rep. média exibe a representatividade média do grupo no qual o documento foi classificado, considerando os 3 valores de $m_{pts}$.

| | mpts | | | Rep. média | Qualidade da extração | |
|---|---|---|---|---|---|---|
| | **58** | **81** | **95** | | **Genérica** | **Com agrupamento** |
| **Clusters** | Azul Claro | Roxo | Roxo | 17% | 69% | 80% |
| | Rosa | Azul | Rosa | 15% | 67% | 77% |
| | Verde Musgo | Verde Claro | Laranja | 13% | 45% | 69% |
| | Verde Claro | Rosa | Verde | 12% | 43% | 60% |

Table IV: Potencial de melhoria da qualidade da extração automática.

O agrupamento demonstrou potencial de melhoria ao processo de extração automática de dados, com a qualidade média variando de 56% para 71% nos documentos analisados. Considerando ainda que atualmente apenas 15% dos documentos têm *layout* conhecido (SPED), destaca-se que os agrupamentos analisados representam, em média, 57% do volume processado, o que indica potencial para aumento significativo da faixa de casos cobertos por métodos padronizados de extração automática.

## 5. CONCLUSÃO

Esta pesquisa objetivou avaliar o potencial de melhoria do processo de extração automática de dados de demonstrações contábeis de empresas através da análise dos agrupamentos gerados com HDBSCAN* sobre vetores de características extraídas desses documentos. A ferramenta MustaCHE apoiou o processo de seleção dos 3 valores de $m_{pts}$, cujos resultados foram analisados quanto à significância dos agrupamentos formados e à aderência ao processo de extração automática de dados.

De acordo com os resultados apurados com a análise estatística dos vetores de características dos documentos agrupados e com os apontamentos à partir do exame dos arquivos originais realizado pelos especialistas em análise de risco de crédito, o HDBSCAN* agrupou os documentos de forma coerente, em grupos distintos entre si, com os 3 valores de $m_{pts}$ selecionados. Dada a quantidade maior de agrupamentos formados com 58 $m_{pts}$, notou-se um grau superior de detalhamento e distinção entre os grupos, o que viabilizou a identificação de complexidades de extração automática diferentes e a percepção de maior aproveitamento conceitual das características e suas respectivas relevâncias.

A atribuição de pesos às características contribuiu para a formação de agrupamentos aderentes ao processo de extração automática de dados de demonstrações contábeis e, consequentemente, poten-

cial de melhoria através da implementação de localizadores mais robustos e ajustados conforme as características estruturais mais relevantes dos documentos agrupados.

Vale ressaltar que o exame dos relatórios financeiros, realizado pelos especialistas em análise de risco de crédito, ratificou alta assertividade do localizador por *script* em aproximadamente 88% dos casos. Além disso, identificou-se oportunidades de melhoria nesse processo de extração das características, fator fundamental para implementação do modelo de classificação dos documentos financeiros.

Como próxima etapa, complementar a esse trabalho, serão implementados localizadores customizados com base nas características predominantes nos documentos de cada agrupamento, com o objetivo de auferir a melhoria do processo de extração automática de dados. Além disso, para pesquisas futuras, sugere-se o refinamento das características relativas ao balanço patrimonial, em paralelo ao aprimoramento do localizador por *script*, além de maior exploração de características exclusivas da demonstração de resultado do exercício (DRE), a fim de avaliar o potencial de resultados de agrupamentos baseados em densidade gerados isoladamente para essas peças contábeis.

## Agradecimentos

Agradecimentos especiais à FAPESP (Grant 2019/09817-6), à Serasa Experian e à UFSCar, por incentivarem, apoiarem e disponibilizarem os recursos necessários para a realização desse trabalho. Parcerias como essa destacam a importância do engajamento entre o setor privado e a academia.

## REFERENCES

Araujo Neto, A. C., Nascimento, M. A., Sander, J., and Campello, R. J. G. B. Mustache: A multiple clustering hierarchies explorer. *Proc. VLDB Endow.* 11 (12): 2058–2061, Aug., 2018.

Araujo Neto, A. C., Sander, J., Campello, R., and Nascimento, M. Efficient computation and visualization of multiple density-based clustering hierarchies. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

Assaf Neto, A. *Estrutura e análise de balanços: um enfoque econômico-financeiro.* Atlas, 2020.

Banco Central do Brasil. *Diagnóstico da convergência às Normas Internacionais: IAS 1 - Presentation of financial statements.* Banco Central do Brasil, Brasília/DF, 2010.

Brasil. Lei nº 6.404, de 15 de dezembro de 1976. *Diário Oficial da União*, 1976.

Brasil. Lei nº 11.638, de 28 de dezembro de 2007. *Diário Oficial da União*, 2007.

Brasil. Lei nº 11.941, de 27 de maio de 2009. *Diário Oficial da União*, 2009.

Campello, R. J. G. B., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. In *PAKDD (2)*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Eds.). Lecture Notes in Computer Science, vol. 7819. Springer, pp. 160–172, 2013.

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery* 27 (3): 344–371, Nov, 2013.

Comitê de Pronunciamentos Contábeis. *Pronunciamento Técnico CPC 21 (R1): Demonstração intermediária: Correlação às Normas Internacionais de Contabilidade - IAS 34 (IASB - BV 2011).* Comitê de Pronunciamentos Contábeis, Brasília/DF, 2011a.

Comitê de Pronunciamentos Contábeis. *Pronunciamento Técnico CPC 26 (R1): Apresentação das demonstrações contábeis: Correlação às Normas Internacionais de Contabilidade - IAS 1 (IASB - BV 2011).* Comitê de Pronunciamentos Contábeis, Brasília/DF, 2011b.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31 (8): 651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

Johnson, D., Xiong, C., Gao, J., and Corso, J. Comprehensive cross-hierarchy cluster agreement evaluation, 2013.

Madeira, R. d. O. C. *Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais.* M.S. thesis, Fundação Getúlio Vargas, Rio de Janeiro/RJ, 2015.

Moura, M. F. Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos. Tech. rep., Embrapa Informática Agropecuária. Dez., 2004.

Snow, M. Unsupervised document clustering with cluster topic identification. Tech. rep., Office for National Statistics. Abr., 2018.