Labor Accidents in Brazil: a Descriptive Analysis

D. F. Giacomelli[1], M. C. Naldi[2], E. R. Faria[1]

[1] Universidade Federal de Uberlândia, Brazil
danielafg@ufu.br,elaine@ufu.br
[2] Universidade Federal de São Carlos, Brazil
naldi@ufscar.br

Abstract. Labor accidents cause several misfortunes, such as inconvenience to the injured ones, loss of labor productivity, and public spending on aid and accident compensation. This work aims to search and characterize groups of labor accidents, granting interpretability to the obtained results, to extract information that can be relevant to public managers. The method proposed in this work consists of the following steps: data pre-processing; the application of two hierarchical clustering algorithms, HDBSCAN * and COBWEB; the evaluation of results using the Simplified Silhouette. The research demonstrated the susceptibility of male workers, focused on ages between 18 and 34 years old, with labor accidents that caused injuries on the fingers, by handling machines and equipment or manual tools, followed by those activities such as fishing. Considering clusters majorly composed by female victims, those related to work in cellulose, paper, and related products stand out. Moreover, fingers are the most affected part, featured for incidents caused by the handling of chemical, biological, or hand tools.

CCS Concepts: • **Applied computing**;

Keywords: labor accidents, clustering, data mining, machine learning

1. INTRODUÇÃO

No ínicio de 2018, segundo levantamento realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o número de trabalhadores formais totalizou 33,3 milhões de pessoas [IBGE 2018]. No entanto, um elemento que gera preocupação é o fato de que o Brasil ocupa o 4º lugar no *ranking* mundial de acidentes de trabalho catalogados. Entre os anos de 2015 e 2018, foram registradas mais de 2 milhões de acidentes de trabalho e gastou-se mais de R$50 bilhões em auxílios e indenizações acidentárias [BRASIL 2018].

O registro de um acidente de trabalho é feito por meio de um documento denominado Comunicação de Acidente de Trabalho (CAT). A análise precisa do conteúdo das CAT's pode oferecer, aos gestores públicos: i) evidências acerca da natureza dos acidentes de trabalho mais frequentes; ii) o estabelecimento de potenciais relações entre região geográfica e os tipos de acidentes; iii) a viabilidade de conduzir políticas públicas em locais diferentes que apresentam o mesmo delineamento; iv) o direcionamento de pesquisas e procedimentos que ambicionem a prevenção de acidentes, de doenças ocupacionais, e o controle dos gastos com benefícios acidentários.

O Ministério Público do Trabalho (MPT), em parceria com a Organização Internacional do Trabalho (OIT), lançou o Observatório Digital de Saúde e Segurança do Trabalho [de Trabalho Decente MPT OIT 2017]. Trata-se de uma plataforma digital que fornece visões sobre as CAT's registradas, gastos previdenciários, mortes acidentárias, perfil das vítimas, dentre outras informações. Ademais, está disponível publicamente a base de dados denominada CATWEB, composta por todos as CAT's registradas entre os anos de 2012 a 2017. O Observatório digital apresenta as informações das CAT's

de forma visual, combinando apenas dois atributos por vez. Contudo, não há como relacionar, de modo automático, os diversos atributos da base de dados CATWEB em busca de grupos de acidentes com alta ocorrência e, por isso, este é foco deste trabalho

O objetivo desta pesquisa consiste em reconhecer e caracterizar grupos nos dados da CATWEB, a fim de conferir integibilidade aos resultados obtidos e extrair informações que subsidiem a tomada de decisão dos gestores públicos. O método proposto neste trabalho é composto das seguintes etapas: pré-processamento dos dados; criação de subconjuntos da base original; seleção dos melhores atributos para realizar a tarefa de agrupamento; aplicação de dois algoritmos de agrupamento hierárquicos, HDBSCAN* e COBWEB; avaliação dos resultados usando a Silhueta Simplificada e emprego da ferramenta PowerBI, para visualizar gráficos que possibilitem avaliar a composição dos grupos encontrados. Adicionalmente, uma medida de dissimilaridade adaptada para a aplicação foi também proposta.

2. TRABALHOS RELACIONADOS

Trabalhos têm explorado o uso de técnicas de mineração de dados em acidentes de trabalho, como [Bartolomeu 2002], que propôs um modelo para investigação dos dados sobre acidentes de trabalho e doenças ocupacionais utilizando técnicas como distribuição de frequência, teste de hipótese e correlação de variáveis. A base de dados utilizada era composta apenas pelos acidentes notificados ao INSS no estado de Santa Catarina em 2000. Conseguiram identificar correlações, padrões, informações implícitas e regras que caracterizam tendências em um curto espaço de tempo, mas destacam o alto índice de campos em branco das CAT's, indicando uma dificuldade dos usuários para seu preenchimento.

O estudo de [Porto and Júnior 2006] utilizou o algoritmo k-médias para identificar grupos em relação aos acidentes do trabalho ocorridos em um hospital. O objetivo da pesquisa foi encontrar, mensurar e descrever as causas e respectivos custos incorridos para empresa. Verificou-se que os setores de maior frequência de acidentes de trabalho são: internação, centro cirúrgico, higiene e limpeza, o cargo mais susceptível é o de auxiliar de enfermagem e a causa majoritária é agulha. Os autores afirmam que, por meio da contabilidade de custos e da análise de agrupamentos, é possível identificar as causas e respectivos custos incorridos pela empresa, o que pode permitir uma gestão mais efetiva.

Os trabalhos de [Brito 2019] e [Rodrigues 2019] utilizaram técnicas de exploração visual na base CATWEB, a mesma base utilizada neste trabalho. No trabalho de [Brito 2019], os *layouts* temporais foram projetados para revelar a evolução dos acidentes ao longo do tempo, como comportamento anômalo e sazonalidade, enquanto que os *layouts* geográficos possibilitam uma navegação entre as localidades, explorando o contexto e comparando as particularidades de cada localidade. O trabalho de [Rodrigues 2019] valeu-se de projeções multidimensionais e *layouts* hierárquicos para compreender questões como a influência de cada atributo na caracterização dos grupos visualizados, a identificação de correlações entre os atributos envolvidos, o comportamento de tendências, entre outras tarefas.

Nossa proposta se diferencia dos trabalhos de [Bartolomeu 2002], [Brito 2019] e [Rodrigues 2019] em relação à técnica de extração de conhecimento, feita por meio do agrupamento de dados. Em relação ao trabalho de [Porto and Júnior 2006], o diferencial consiste em não restringir o acidente de trabalho ao tipo de atividade laboral exercida, explorando toda a diversidade de acidentes de trabalho, nas diferentes ocupações e áreas econômicas. Além disso, os algoritmos de agrupamento eleitos são do tipo hierárquico, enquanto [Porto and Júnior 2006] usou um algoritmo particional.

3. MATERIAIS E MÉTODOS

O presente trabalho propõe o uso de técnicas de aprendizado não-supervisionado para a extração de conhecimento dos dados referentes aos acidentes de trabalho. Dessa forma, as relações entre os dados será estabelecida pelas suas próprias características, sem nenhum conhecimento externo. Isso se justifica pelo fato de a base de dados não possui classes conhecidas (*ground truth*).

3.1 Base de dados - CATWEB

A CATWEB é composta por 3.879.755 instâncias, que consistem em todas as Comunicações de Acidentes de Trabalho registradas entre os anos de 2012 e 2017. A mencionada base possui um total de 18 atributos, dos quais 15 são categóricos e 3 são numéricos. Trata-se de uma base com grande volume de dados com valores ausentes, que contempla a heterogeneidade das regiões do país e possui uma ampla diversidade quanto aos tipos, causas e perfis dos acidentes. Seus atributos são: Indicador de Acidente em Feriado, Agente Causador, Ano do Acidente, Classe de Atividade Econômica (CNAE), Data do Acidente, Dia da Semana, Emitente, Hora do Acidente, Idade, Indicador de Óbito, Município, Unidade Federativa, Natureza da Lesão, Classificação Brasileira de Ocupações (CBO), Parte do Corpo Atingida, Sexo, Tipo de Acidente e Local do Acidente.

3.2 Método proposto

A Figura 1 ilustra as etapas do método proposto para a condução deste trabalho, detalhadas a seguir.

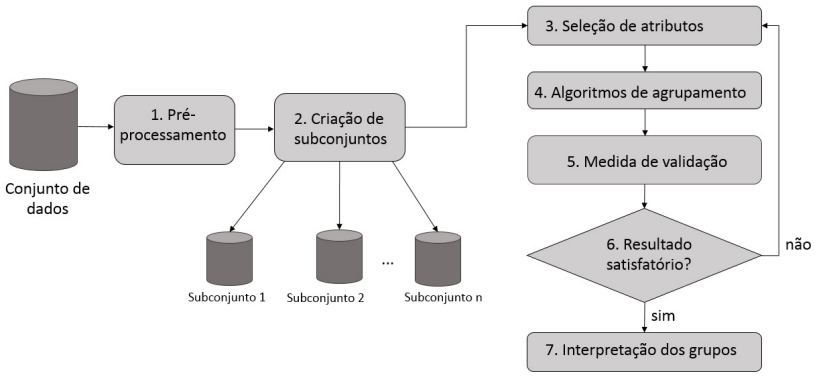

Fig. 1. Etapas do método proposto

1ª Etapa - Pré-processamento: Consiste em 3 passos: sumarização de atributos, conversão de atributos e tratamento de valores ausentes.

—A sumarização objetiva diminuir a quantidade de possíveis valores para os atributos categóricos. Os atributos "Classe Nacional de Atividade Econômica (CNAE)", "Município" e "Agente causador" passaram por esse processo, aumentando sua granularidade.

—O atributo "Município" apresentava 5.285 valores distintos e, após ser convertido para "Mesorregião", a quantidade de valores distintos diminuiu para 165. Já o atributo "Agente causador" foi convertido, de acordo com tabela de referência disponibilizada pelos especialistas do MPT. A sumarização reduziu 302 valores distintos para apenas 21.

—O atributo CNAE é representado por uma estrutura hierárquica, dividida em: seção, divisão, grupo e classe. Na base original, este atributo estava representado pela classe, que consiste na menor granularidade, contendo 668 valores distintos. A classe CNAE foi convertida para o nível seção, que possui apenas 21 valores diferentes.

—O atributo "Idade", originalmente numérico, foi convertido para categórico. A seguinte divisão em faixas etárias foi adotada para o pré-processamento: Jovem Adulto ($18 \leq idade < 35$), Adulto ($35 \leq idade < 55$) e Idoso ($idade \geq 55$).

—Todas as instâncias da base que possuíam valores ausentes foram removidas, de forma que mais de 80% da base foi preservada. A presença desses valores causa um impacto negativo para os algoritmos de aprendizado de máquina.

2ª Etapa - Criação de subconjuntos: Devido ao grande volume de dados, fez-se necessária a divisão da base de dados em subjconjuntos para facilitar a análise dos dados. O primeiro critério adotado para dividir a base de dados foi o ano do acidente. Desta forma, foram criados 6 subconjuntos de dados, um para cada ano da base (2012 a 2017). O segundo critério adotado para dividir a base de dados foi a mesorregião, totalizando 165 subconjuntos.

3ª Etapa - Seleção de atributos: Foi realizado um estudo experimental a fim de verificar o impacto da variação do conjunto de atributos no desempenho do algoritmo de agrupamento, segundo uma medida de validação. Após a discussão com especialista (analista do ministério do trabalho) e o estudo experimental, optou-se por manter os seguintes atributos: Agente Causador, CNAE, Sexo, Parte do Corpo Atingida, Idade, Tipo de acidente e Local do Acidente.

4ª Etapa - Algoritmo de agrupamento: Este trabalho propõe o uso de dois algoritmos de agrupamento, o HDBSCAN* [Campello et al. 2015] e o COBWEB [Fisher 1987]. O HDBSCAN* é um algoritmo de agrupamento hierárquico baseado em densidade, que trabalha com as distâncias entre os objetos. O COBWEB, por sua vez, também é hierárquico mas baseia-se em probabilidades condicionais e é capaz de trabalhar tanto com atributos numéricos quanto categóricos. Como a quantidade de grupos é desconhecida *a priori*, a ideia de usar algoritmos hierárquicos é interessante, pois dessa forma seria possível avaliar várias possibilidades, com grupos aninhados, e buscar pela melhor solução avaliada por um medida ou critério de avaliação.

5ª Etapa - Medida de Validação: A medida de validação adotada para verificar a qualidade do agrupamento obtido foi a Silhueta Simplificada [Hruschka et al. 2004]. Esse índice combina características de coesão no grupo e separação entre os grupos. Para o cálculo da silhueta é preciso computar a distância entre dois objetos. Essa distância foi calculada conforme descrito na seção 3.3.

6ª Etapa - Verificação dos resultados: Para apoiar a etapa de verificação de resultados, foi utilizada a ferramenta PowerBI [1]. Essa ferramenta foi desenvolvida pela Microsoft e possibilita análise de dados por meio da criação de gráficos e painéis. O PowerBI foi utilizado para explorar os grupos encontrados pelos algoritmos de agrupamento. Os gráficos permitem visualizar questões como: a quantidade de objetos por grupo, homogeneidade dos atributos dentro de um grupo e comparar as características de diferentes grupos. No caso dos resultados do agrupamento não serem satisfatórios, uma nova seleção de atributos é feita e todas as etapas seguintes são repetidas.

7ª Etapa - Interpretação dos grupos: Após obter agrupamentos que apresentem uma boa avaliação, os maiores grupos serão analisados, buscando-se por fontes externas que possam colaborar com a interpetação dos resultados obtidos.

3.3 Adaptações no cálculo da distância

O HDBSCAN* é um algoritmo de agrupamento do tipo relacional, que necessita apenas das dissimilaridades entre os objetos. Considerando que a base CATWeb é formada tanta por atributos categóricos quanto numéricos, uma medida de distância adaptada ao problema foi criada. Essa medida foi usada tanto pelo HDBSCAN* quando no cálculo da silhueta, para validação do agrupamento.

As fórmulas utilizadas para realizar o cálculo da distância entre dois atributos estão expostas na Tabela I. Um atributo específico da base de dados, o dia da semana, requereu uma maneira diferenciada para o cálculo da distância. Nesse caso foi utilizada uma lista circular e seu valor corresponde à menor distância entre dois dias da semana, não importando a ordem dos fatores.

Por fim, para encontrar a distância total entre dois objetos, é preciso compor a distância quanto a todos os atributos. Neste estudo, adotou-se a média aritmética para tal composição. A distância entre dois objetos quaisquer pertence ao intervalo $[0, 1]$.

[1] https://powerbi.microsoft.com/pt-br/

Tipo de atributo	Fórmula
Categórico	$d = \begin{cases} 1, & m \neq n \\ 0, & m = n \end{cases}$
Numérico	$d = \frac{x_a - y_a}{max_a - min_a}$
Caso especial - Dia da semana	$d = \frac{min[\lvert x_{diasemana} - y_{diasemana} \rvert, \lvert y_{diasemana} - x_{diasemana} \rvert]}{3}$

Table I. Medida proposta para o cálculo da distância

4. EXPERIMENTOS E ANÁLISE DE RESULTADOS

Quatro cenários experimentais distintos foram considerados e seus resultados são apresentados nas próximas seções. Em seguida, uma comparação entra os algoritmos de agrupamento utilizados é feita.

4.1 Cenário experimental 1:

Objetiva caracterizar os maiores grupos encontrados pelos algoritmos de agrupamento, estabelecendo semelhanças e diferenças entre eles, e verificar se as características se mantêm ao longo dos anos.

Os resultados do HDBSCAN* atingiram o valor de Silhueta Simplificada igual a 1 e o percentual de *outliers* foi sempre em torno de 50%. Os resultados do COBWEB apresentaram valores de Silhueta Simplificada entre 0,32 e 0,38 e, uma vez que o algoritmo não trata *outliers*, todas as instâncias foram inseridas em algum grupo. Observou-se que os grupos do HDBSCAN* são formados apenas por instâncias iguais. Analisando os agrupamentos obtidos pelos dois algoritmos utilizados e comparando os anos de 2012 a 2017, foi possível observar que o maior grupo de cada ano preservava as seguintes características: **Sexo:** Masculino; **Agente causador:** "Máquinas e equipamentos" ou "Ferramentas Manuais"; **CNAE:** 3 (Pesca e Aquicultura); **Parte do corpo atingida:** Dedo; **Tipo de acidente:** Típico; **Local do acidente:** Empregadora; **Faixa etária:** "Jovem Adulto" ou "Adulto".

Como o HDBSCAN* agrupou apenas instâncias idênticas, optou-se por analisar também os demais atributos do maior grupo de cada ano, a fim de obter mais informações. Ao analisar os demais atributos do maior grupo de cada ano, utilizando o PowerBI, o seguinte comportamento foi observado:

—**Indicador de acidente em feriado:** Mais de 98% das CAT's não ocorreram em feriado;
—**Dia da semana:** Acidentes são mais frequentes durante a semana, "Terça-feira" possui mais casos;
—**Emitente:** Observou-se que mais de 90% das CAT's foram registradas pelo empregador;
—**Indicador de óbito:** O maior grupo de cada ano que não possui nenhum caso de óbito;
—**Natureza da lesão:** As lesões mais frequentes foram "Corte, laceração, ferida contusa, punctura", "Fratura" e "Contusão, esmagamento";
—**CBO:** a ocupação foi o atributo que apresentou maior diversidade de valores dentro dos grupos. No entanto, a ocupação com maior registro de CAT's foi "Alimentador de linha de produção", seguida por "Açogueiro", "Operador de máquinas fixas em geral" e "Mecânico de manutenção de máquinas";

O COBWEB diferencia-se quanto ao dia da semana em que mais ocorre acidentes, sendo "Segunda-feira" o dia com mais casos registrados. E quanto à ocupação, destaca "Alimentador de linha de produção", "Operador de máquinas fixas em geral" e "Mecânico de manutenção de máquinas".

É importante destacar que os algoritmos baseados em densidade, como o HDBSCAN*, buscam por regiões de alta densidade, que estejam rodeadas por regiões de baixa densidade [Semaan 2013]. Verificou-se que a amostra de dados era composta por um grande número de instâncias idênticas, fator que compromete o desempenho desse tipo de algoritmo, uma vez que essas instâncias configuram regiões de alta densidade e ficam sobrepostas. No entanto, constatar o grande número de acidentes iguais é uma questão que pode ser importante para o Ministério Público do Trabalho.

Ademais, inferiu-se que as características dos maiores grupo de acidentes se mantém ao longo dos anos, nos agrupamentos produzidos por ambos algoritmos de agrupamento. Portanto, concluiu-se que os trabalhadores do sexo masculino, com idade entre 18 e 55 anos (jovem adulto e adulto), que exercem atividades de Pesca e Aquicultura, frequentemente se acidentam durante o exercício de suas atividades e têm o dedo como a parte do corpo lesionada por máquinas e equipamentos.

4.2 Cenário experimental 2:

No cenário 1, observou-se que a base possui um alto número de instâncias idênticas, fator que pode comprometer o desempenho de algoritmos baseados em densidade, pois as instâncias idênticas são vistas como regiões de alta densidade. O experimento 2 foi realizado para avaliar o comportamento do HDBSCAN* após a eliminação das instâncias repetidas da base. A amostra era composta por 53.780 instâncias, das quais 40.646 eram idênticas umas às outras. Após eliminar as instâncias idênticas, foi obtida uma amostra de 13.134 instâncias distintas. Contudo, observou-se que o algoritmo não foi capaz de encontrar nenhum grupo, todas as instâncias foram classificadas como *outliers*.

Disposto a avaliar e justificar tal comportamento do algoritmo, realizou-se uma análise sobre o funcionamento da medida de distância empregada. A medida de distância proposta produz valores no intervalo entre 0, quando duas instâncias são exatamente iguais, e 1, para instâncias diferentes em todos os atributos. As instâncias da amostra de dados utilizadas neste experimento apresentaram valores sempre iguais a $0,7142$ (20,38% da amostra), $0,8571$ (39,79% da amostra) ou a 1 (39,83% da amostra). Enfatizando que, os 7 atributos selecionados neste estudo são categóricos, os valores de distância observados indicam que as instâncias apresentam dois, um ou nenhum atributo(s) em comum, respectivamente. Dessa forma, percebe-se que a amostra contém acidentes de trabalho muito diferentes (dados os atributos selecionados) e o algoritmo não foi capaz de agrupá-las.

A fim de gerar mais valores possíveis de distância entre instâncias, foi realizado um experimento considerando a idade não como atributo categórico, mas como numérico. O novo experimento resultou em valores de distância mais variados e o algoritmo foi capaz de realizar o agrupamento. O HDBSCAN* segmentou a base em 1954 grupos e alcançou um coeficiente de Silhueta Simplificada igual a $0,57$. Todavia, constatou-se que dentro dos grupos todas as instâncias eram iguais entre si, exceto em relação ao atributo idade. Concluiu-se então que a forma como se realizou o cálculo da distância entre atributos categóricos foi rígido e uma maior variedade de valores é necessária.

4.3 Cenário experimental 3:

O objetivo do Experimento 3 foi caracterizar os maiores grupos encontrados na base CATWEB ao dividí-la por Mesorregião, buscando comparar duas Mesorregiões distintas, a fim de estabelecer semelhanças e diferenças entre elas. Ao dividir a base pela mesorregião foram obtidos 165 subconjuntos. A mesorregião com a maior quantidade de CAT's registradas é a Região Metropolitana de São Paulo. Analisando a qualidade do agrupamento, observou-se que 49,18% da amostra foi considerada *outlier* e a medida de validação alcançou seu valor máximo, ao utilizar o HDBSCAN*. O COBWEB, por sua vez, apresentou o coeficiente de Silhueta Simplificada igual a 0,35. A mesorregião do Triângulo Mineiro é a 12ª com a maior quantidade de CAT's registradas. Destaca-se que no agrupamento obtido pelo HDBSCAN*, 49,42% das amostras foram consideradas *outliers* e a Silhueta Simplicada foi igual a um. Ao passo que, o COBWEB alcançou o valor de Silhueta Simplificada igual a 0,32.

Outro grupo que se sobressai é dos acidentados que tiveram o sistema nervoso como parte do corpo atingida, predominantemente trabalhadores do sexo masculino ao exercer atividades de extração de minerais não-metálicos (CNAE 8). Por fim, um grupo composto por vítimas do sexo feminino, é verificado nas duas mesorregiões. Esse grupo é composto por trabalhadoras que atuam na fabricação de celulose, papel e produtos de papel (CNAE 17), que tiveram o dedo como parte do corpo atingida. Diferenciando-se apenas quanto ao agente causador.

4.4 Cenário experimental 4:

O objetivo do Experimento 4 foi analisar os grupos de cada ano, buscando aqueles que não tenham o dedo como a parte do corpo atingida (mais frequente), e assim verificar se as características ainda se mantêm ao longo dos anos.

A conclusão obtida com o Experimento 4 foi que mesmo considerando outra parte do corpo atingida pelo acidente de trabalho, as carcaterísticas dos grupos ainda se assemelham ao longo da maioria dos anos. Dessa forma, os trabalhadores do sexo masculino, que exercem atividades de extração de minerais não-metálicos, representam um grupo que frequentemente se acidentam e lesionam o sistema nervoso, tendo como consequência a perda auditiva.

4.5 Comparação de resultados - HDBSCAN* x COBWEB

Os experimentos 1, 3 e 4 realizados com o HDBSCAN* apresentaram valores máximos de Silhueta Simplificada, enquanto o algoritmo COBWEB apresentou valores entre $0,32$ e $0,38$. Em relação ao tamanho dos grupos, constatou-se que os grupos definidos pelo COBWEB são maiores (em número de instâncias) que os grupos gerados pelo HDBSCAN* e a quantidade de grupos definidos pelo COBWEB é inferior. Isso pode ser explicado por dois fatores: o COBWEB não agrupou apenas instâncias iguais e, também, não rotula nenhuma instância como *outlier*. A Tabela II ilustra um comparativo.

Table II. Comparação de agrupamentos - HDBSCAN* x COBWEB

	Região Metropolitana de São Paulo		Triângulo Mineiro	
	HDBSCAN*	COBWEB	HDBSCAN*	COBWEB
Silhueta Simplificada	1	0,35	1	0,32
Número de grupos	1952	1124	1851	1133
Maior grupo	754 instâncias	1374 instâncias	792 instâncias	1548 instâncias
Menor grupo	4 instâncias	4 instâncias	4 instâncias	4 instâncias

Finalmente, as características dos maiores grupos destacados nos experimentos realizados com o HDBSCAN* se mantêm nos grupos destacados nos experimentos que empregaram o COBWEB. Levando-se a conclusão de que tais grupos realmente se destacam na base de dados CATWEB. O alto número de grupos do conjunto reflete a grande variedade acidentes de trabalho e, por isso, considerar os grupos majoritários é importante.

5. CONCLUSÕES

Este trabalho apresenta como contribuições a disponibilização de um método para condução de experimentos envolvendo agrupamento de dados; uma ferramenta capaz de realizar o pré-processamento dos dados da base CATWEB; uma medida de cálculo de distância adaptada para calcular a distância entre dois objetos da base CATWEB (considerando os seus atributos numéricos e categóricos), e a caracterização dos maiores grupos de acidentes de trabalho. Neste trabalho concluímos, por meio dos experimentos realizados sobre as Comunicações de Acidentes de Trabalho registradas entre os anos de 2012 e 2017, que:

—Trabalhadores do sexo masculino, com idade entre 18 e 34 anos, que exercem atividades de Pesca e Aquicultura, estão suscetíveis a acidentes durante o exercício de suas atividades e têm o dedo como a parte do corpo lesionada por máquinas e equipamentos ou ferramentas manuais.

—Trabalhadores do sexo masculino, que exercem atividades de extrações de minerais não metálicos, estão suscetíveis a acidentes causados por um agente biológico, danificando o sistema nervoso e apresentando como principal consequência a perda ou diminuição da audição;

—Trabalhadoras do sexo feminino, que atuam fabricação de celulose, papel e produtos de papel, estão suscetíveis a acidentes que têm o dedo como a parte do corpo atingida. Nesses casos, os agentes causadores mais frequentes são agente químico, agente biológico ou ferramentas manuais;

—Em geral, as características observadas nos grandes grupos de acidentes de trabalho são as mesmas em todos os anos analisados, nos agrupamentos produzidos pelos dois diferentes algoritmos empregados nesta pesquisa;

—O agrupamento produzido pelo COBWEB resulta em um número menor de grupos e grupos com um maior número de instâncias, quando comparado ao agrupamento resultante do HDBSCAN*.

—As características observadas nos grandes grupos de acidentes de trabalho quando a base de dados é subdividida pelo ano do acidente, se repetem nos grandes grupos de acidente quando a base é subdividida por mesorregião, para as duas mesorregiões analisadas nesta pesquisa;

Como trabalhos futuros, recomenda-se: o uso de outros algoritmos de agrupamento de dados; a aplicação de um método diferente do proposto neste trabalho para a conduzir a extração de conhecimento; a utilização de outros tipos de pré-processamento; e o uso de outras medidas de validação para verificar a qualidade do agrupamento obtido.

6. AGRADECIMENTOS

Agradecemos às agências FAPEMIG, pela bolsa concedida, e FAPESP (Grant 2019/09817-6), e ao Ministério Público do Trabalho, pelo fornecimento dos dados.

REFERENCES

Bartolomeu, T. A. *Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento*. Ph.D. thesis, Florianópolis, SC, 2002. Programa de Pós-Graduação em Engenharia de Produção.

BRASIL. Anuário estatístico da previdência social, 2018. Acesso em: 09/06/2019.

Brito, L. L. *A strategy for temporal visual analysis of labor accident data*. M.S. thesis, Universidade Federal de Uberlândia, 2019. Programa de Pós-Graduação em Computação.

Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10 (1): 1–51, 2015.

de Trabalho Decente MPT OIT, S. Observatório digital de saúde e segurança no trabalho (MPT-OIT), 2017. Dados acessados em: 10/03/2018. Disponível online no seguinte endereço http://observatoriosst.mpt.mp.br.

Fisher, D. H. Knowledge acquisition via incremental conceptual clustering. *Machine learning* 2 (2): 139–172, 1987.

Hruschka, E. R., de Castro, L. N., and Campello, R. J. Evolutionary algorithms for clustering gene-expression data. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, pp. 403–406, 2004.

IBGE. Desemprego recua em dezembro, mas taxa media do ano e a maior desde 2012, 2018. Acesso em: 01/08/2018.

Porto, S. S. S. and Júnior, I. J. d. N. Acidentes de trabalho no hospital anchieta: Uma análise exploratória de suas características a partir da análise de agrupamentos (clusters). In *Anais do Congresso Brasileiro de Custos-ABC*, 2006.

Rodrigues, M. P. *A strategy for visual structural data analysis of labor accident data*. M.S. thesis, Universidade Federal de Uberlândia, 2019. Programa de Pós-Graduação em Computação.

Semaan, G. S. *Algoritmos para o Problema de Agrupamento Automático*. Ph.D. thesis, Tese de Doutorado, Instituto de Computação, Universidade Federal Fluminense, 2013.