ANÁLISE EXPLORATÓRIA E VISUALIZAÇÃO DE PROCESSOS DE CLUSTERIZAÇÃO DE PONTOS GEORREFERENCIADOS: PROPOSTA DE UMA FERRAMENTA GRÁFICA DE ANÁLISE E VISUALIZAÇÃO

Krizhan Wesley Santos da Silva e Arnaldo Rabello de Aguiar Vallim Filho

Apoio:PIBIC Mackenzie

RESUMO

O presente projeto de pesquisa foi desenvolvido com o intuito de desenvolver soluções e ferramentas gráficas envolvendo análises exploratórias e visualização de resultados de um processo de clusterização. O estudo foi desenvolvido para uma aplicação em operações logísticas em que os clusters representam agrupamentos de pontos geográficos que demandam por serviços logísticos. Foi desenvolvido e implementado em uma ferramenta computacional um processo em que os dados de entrada são pontos com sua informação de localização geografia e já clusterizados. Esses pontos passam por um tratamento de transformação de coordenadas e os clusters são apresentados na ferramenta gráfica desenvolvida na linguagem R, com o uso da biblioteca R-*Shiny*., tendo como fundo um mapa da região de estudo. Além disso, a ferramenta permite ainda algumas análises exploratórias dos dados. O resultado do estudo se mostrou satisfatório, tendo-se chegado à implementação de uma ferramenta computacional de visualização que apresenta com clareza os resultados do processo de clusterização em um contexto de operação logística, que era o objetivo da pesquisa.

Palavras-chave: Visualização, clusterização, ferramenta gráfica

ABSTRACT

This project was developed aiming at developing solutions and graphic tools involving exploratory analysis and visualization of results from a clustering process. The study was developed for an application in logistics operations where clusters represent groupings of geographical points that demand logistics services. The process was developed and implemented in a computational tool in which the input data are points with their geographic location information and already clustered. These points undergo a coordinate transformation treatment and the clusters are presented in the graphic tool developed in R language, using the R-Shiny. library, having as background a map of the study region. In addition, the tool also allows some exploratory analysis of the data. The result of the study proved to be satisfactory, having reached the implementation of a computational visualization tool that clearly presents the results of the clustering process in a context of logistic operation, which was the research objective.

Keywords: Visualization, clustering, graphic tool

1. INTRODUÇÃO

Este trabalho é parte de um projeto mais amplo que busca o desenvolvimento de heurísticas e metaheurísticas para solução de problemas de otimização combinatória de diferentes tipos e de ferramentas de visualização dos resultados. No caso específico deste artigo, o que se buscou na pesquisa foi o desenvolvimento de ferramentas para análises exploratórias dos dados dos problemas e de ferramentas gráficas de visualização para apresentação dos resultados obtidos.

A pesquisa se justifica porque sabe-se que a competitividade entre empresas vem aumentando. E para que a melhor decisão seja tomada, é necessário que esteja fundamentada em dados que comprovem que tal escolha representa uma solução adequada ou a melhor delas. Em função do grande volume de informações, o mais apropriado é que seja feito uso da Ciência de Dados, em particular da Análise Exploratória de Dados e da Análise Preditiva.

Esta questão é particularmente importante na área de Logística. Em operações logísticas de grande porte o número de pontos de demanda (pontos a atender) pode chegar aos milhares, e inclusive, o tratamento individual de cada ponto só é realmente feito nos processos operacionais finais quando se tem a definição final do atendimento daquele ponto específico.

Um dos aspectos chave deste processo é a clusterização desses dados, que é uma das tarefas principais de Mineração de Dados (MD). Quando se trata de planejamento da operação, os pontos (clientes a atender) são agrupados em clusters, uma vez que favorecem uma visão bem melhor de tendências e padrões de comportamento do conjunto. A construção de clusters é na verdade, uma etapa fundamental do planejamento logístico.

Mas, além do processo de clusterização em si, há a questão de análise dos dados que geraram os clusters e dos resultados da clusterização. E há ainda, a questão da visualização desses dados e resultados.

Esta pesquisa se concentrou assim, em problemas de clusterização com aplicação em operações logísticas. Este estudo parte, portanto, de resultados já obtidos com algoritmos de clusterização e explora esses resultados e os apresenta graficamente por meio de ferramentas desenvolvidas na pesquisa. É uma pesquisa, portanto, voltada fundamentalmente para o desenvolvimento de uma ferramenta de visualização.

Aliados importantes dessa visualização são as bibliotecas gráficas, que permitem o

desenvolvimento de aplicações web dando suporte para criação de painéis gráficos (*dashboards*) que permitem visualização clara de resultados de análises desenvolvidas em operações logísticas.

Assim, uma vez que a clusterização em operações logísticas já tenha se desenvolvido, tem-se dois pontos importantes a considerar:

. o desenvolvimento de análises e

. o desenvolvimento de ferramentas de visualização dessas análises.

Este projeto de pesquisa não pretende se ocupar da clusterização em si, mas, sim, dessas duas fases posteriores:

. A análise dos resultados da clusterização e

. A Visualização desses resultados.

Assim, dentro desse quadro a pergunta de pesquisa que se colocou neste projeto foi:

> *"Uma vez que a partir de um grande volume de dados de uma operação logística tenham sido gerados clusters, como poderia se desenvolver uma análise exploratória que caracterizasse os resultados da clusterização, e como poderia ainda, ser desenvolvida uma ferramenta gráfica de visualização desses resultados?*

Nas próximas seções apresenta-se como o estudo foi desenvolvido. Inicia-se por uma revisão bibliográfica na seção 2, seguida da metodologia empregada, seção 3. Na sequência, seção 4, são apresentados os resultados e discussão e, finalmente, na seção 5 apresenta-se as considerações finais.

2. REFERENCIAL TEÓRICO

Este capítulo tem como objetivo apresentar uma revisão da literatura de forma a se obter uma visão geral de trabalhos associados ao tema deste projeto. É aquilo que se denomina o "estado da arte". Essas informações permitem que se tenha uma base mais sólida para o desenvolvimento do projeto.

Como este projeto está diretamente associado à Mineração de Dados (MD), inicia-se esta revisão com um breve levantamento do assunto, procurando-se dar uma visão geral deste tema.

Uma obra que que fornece uma visão clara do assunto é o livro de SILVA et al. (2016), em que coloca que a MD pode ser um processo automático ou semiautomático, que envolve análises exploratória e preditiva de uma massa de dados.

A análise exploratória tem por objetivo procurar padrões relevantes sobre o conjunto de dados, e com isto, conseguir gerar conhecimento novo.

A análise exploratória de uma massa de dados possibilita que se compreenda o

comportamento dos dados olhando-se o conjunto como um todo, mas também, eventualmente, há interesse em se analisar individualmente determinadas observações (chamadas de exemplares ou amostras) para se escolher aquelas que são diferenciadas ou aquelas que são mais representativas da massa de dados, fornecendo uma identidade do conjunto.

A análise preditiva, dentre suas várias possibilidades, uma que é importante é que possibilita a identificação de relacionamentos entre variáveis (atributos) de uma massa de dados. Isto pode ser feito analisando-se as observações desses atributos, que podem ser categóricos ou numéricos (SILVA et al., 2016). A análise preditiva compreende aspectos de predição, agrupamento ou associação de dados, e está associada à inferência estatística, heurísticas e *machine learning*, uma área da Inteligência Artificial. A análise exploratória, por outro lado procura criar visões que representem o conjunto de dados, utilizando instrumentos da estatística descritiva.

Em termos de sequência, inicia-se com a análise exploratória que irá mostrar o comportamento dos dados, e em seguida, vem a análise preditiva, que envolve agrupamento ou associação e a predição propriamente, dita, projetando resultados para o futuro.

Em MD, portanto, busca-se fundamentalmente ampliar o conhecimento a respeito de uma massa de dados. Essa “descoberta de conhecimento” (*knowledge discovery*) em bases de dados se desenvolve segundo algumas etapas principais.

Todo o processo se inicia pela Captura dos Dados ou o chamado ETL (*Extraction, Transform and Loading*, que é o processo de extração de dados, sua transformação com fins de análise e sua carga em um base de dados.

Considerando-se que a extração de dados já tenha ocorrido, a sequência da MD envolveria:

Pré-processamento dos Dados

. Onde é feito o tratamento de “ruídos”, (amostras com informações faltando, inconsistências, outliers, etc.);

. Normalização de valores, para que não se trabalhe com diferentes unidades e escalas,

. Organização dos dados em um único repositório;

Análise Exploratória;

Análise Preditiva;

Visualização de Dados e Resultados.

Deve ser observado que a análise exploratória é importante em mais de uma fase, como é o caso do pré-processamento e do pôs processamento, permitindo resultado

mais precisos e com visualização adequada para favorecer a compreensão dos resultados

Os passos anteriores não necessariamente precisam ser executados em uma sequência rígida, podendo ser realizados em ordens ligeiramente diferentes, dependendo da situação (SILVA et al., 2016; DE CASTRO e Ferrari, 2016).

Ao final do processo de MD, os resultados são validados e avaliados e note-se que um aspecto importante do processo ocorre justamente ao final, que é a visualização de resultados, que devem ser representações visuais que facilitem a compreensão (SILVA et al., 2016; DE CASTRO e Ferrari, 2016).

Para que todo esse processo se realize, a MD envolve a implementação de técnicas e algoritmos computacionais, que importam (leem) dados de uma base que esteja em estudo, analisam esses dados e retornam definindo um padrão de comportamento para aqueles dados. Espera-se que esses resultados auxiliem na tomada de decisão.

O projeto deverá envolver também uma análise descritiva de resultados de um processo de análise preditiva, assim convém lembrar alguns dos pontos fundamentais deste ramo da estatística.

A estatística é uma ciência antiga e um de seus ramos é a estatística descritiva que permite o cálculo de medidas que descrevem o comportamento dos dados, envolvendo, por exemplo, posição e variabilidade, como a média e variância ou desvio padrão.

Segundo Bussab e Morettin (2013) uma corrente mais recente, liderada por J. W. Tukey, passou a utilizar muito mais as técnicas gráficas, ao invés de resumos numéricos. O que significa que sumários estatísticos não devam ser desenvolvidos, mas quando se trata de análise exploratória de dados, esta não deve se limitar exclusivamente à apresentação de medidas estatísticas na forma de um número. Gráficos e outros recursos gráficos devem se constituir em parte fundamental da apresentação.

A Estatística Descritiva pode ser entendida como sendo um conjunto de instrumentos para descrever e resumir dados, revelando seus aspectos mais importantes (SILVA et al., 2016).

Os instrumentos podem ser segmentados em medidas estatísticas e ferramentas gráficas. Em termos de medidas as principais são as medidas de posição, de dispersão e as separatrizes. As medidas de posição mostram os dados em relação à sua localização. As medidas de dispersão mostram o nível de variação dos dados. E as separatrizes subdividem o conjunto de dados em partes, não necessariamente iguais Como medidas de posição, do tipo central, tem-se a média, mediana e moda. As

medidas de dispersão mais importantes são a variância e o desvio padrão e as separatrizes são a mediana, quartis e percentis (Devore, 2014 e Montgomery e Runger, 2013).

Moore et al, (2017) trabalha bastante as ferramentas gráficas da estatística descritiva, apresentando diferentes tipos de instrumentos. Dentre eles destaca os gráficos de dispersão, que mostram a evolução de uma variável no tempo, os histogramas, que mostram a distribuição dos valores de uma variável, gráficos de setores ("pizza"), os diagramas de ramos e folhas, semelhante aos histogramas, mas que apresentam a distribuição da variável de forma mais detalhada, e o diagrama em caixa (*boxplot*), que apresenta em uma forma gráfica um resumo de cinco números notáveis de um conjunto de dados: os quartis, a mediana (2º Quartil) e os limites para caracterizar pontos fora de padrão, os chamados *outliers*.
Essas ferramentas gráficas se constituem em parte integrante do processo de visualização de dados, que é um dos focos deste projeto.

A visualização de informações é um campo que busca a construção de representações visuais de dados para facilitar sua compreensão, assim como, apoiar a descoberta de conhecimento contido nesses dados (Nascimento e Ferreira, 2005).
Uma apresentação gráfica tem muitas vantagens, pois, grande quantidade de dados pode ser sintetizada em uma simples visualização. O famoso ditado popular "uma imagem vale mais do que mil palavras" mostra valor de uma imagem. Além disso, o sistema visual humano é treinado para reconhecer padrões (Nascimento e Ferreira, 2005), o que facilita muito o entendimento de resultados.
De forma mais ampla, Tavares e Alexandre (2009) colocam que visualização corresponde à construção de uma imagem visual na mente humana, e isto é mais do que uma representação gráfica de dados ou conceitos. Uma visualização representa assim, uma ferramenta cognitiva; tornando-se um artifício externo para gerar conhecimento, e faz isto, utilizando as capacidades perceptivas e cognitivas humanas. Entende-se assim, que a definição e construção de representações gráficas deve ser um processo muito cuidadoso e muito bem planejado.

A figura 1 extraída de Gil e Barleta (2015) mostra como uma figura pode agregar muitas informações que em caso contrário, estariam descritas em vários parágrafos de texto.

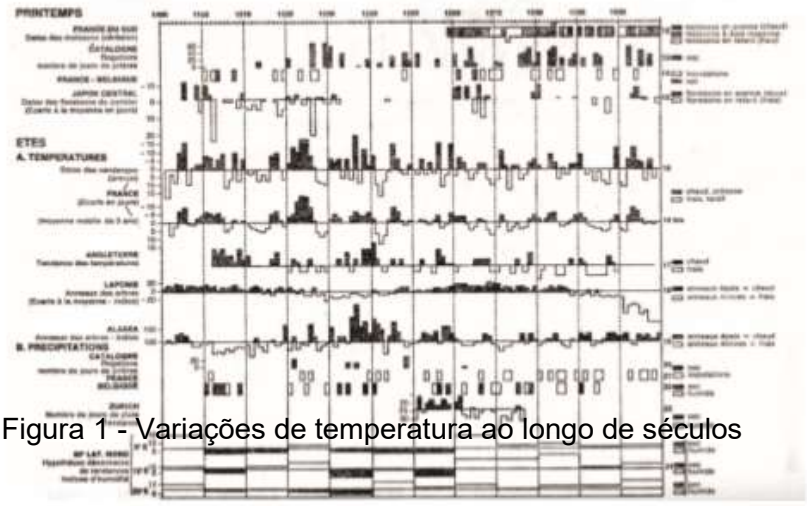

Figura 1 - Variações de temperatura ao longo de séculos

A Figura 2 mostra um modelo de Referência para Visualização apresentado em Freitas et al. (2001) proposto por Card et al. (1999).

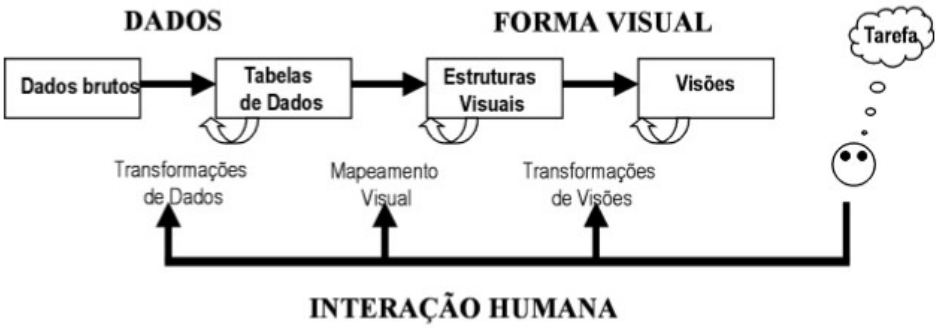

Figura 2 - Modelo de Referência para Visualização

Freitas et al. (2001) mostram inúmeras técnicas de visualização sendo que a maior parte delas está implementada em ferramentas de análise estatística e visualização.

A biblioteca R-Shiny da linguagem R, por exemplo, (R-Studio, 2019) é um pacote do R que permite a criação de aplicativos Web interativos diretamente do R, com uma variedade de recursos gráficos.

Há também bibliotecas disponíveis para a linguagem Python que é muito utilizada em MD. O Matplotlib (Matplotlib, 2019) é uma biblioteca de plotagem 2D do Python que produz representações gráficas em uma variedade de formatos. Pode ser usado em scripts Python facilitando a construção das visualizações.

Há ainda o *Seaborn* (Seaborn, 2019) que é uma biblioteca de visualização Python baseada no matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos

Outra ferramenta que tem sido bem utilizada é o Tableau (Tableau, 2019)
O Tableau (Tableau, 2019) permite análises visuais em tempo real facilitando a exploração de dados. Painéis interativos apoiam o processo investigativo da descoberta de conhecimento.

3. METODOLOGIA

A pesquisa desenvolvida neste projeto foi do tipo “pesquisa aplicada”, e o estudo adotou uma abordagem quantitativa. A sua finalidade foi o desenvolvimento de análises exploratórias e de uma ferramenta de visualização.

De forma geral, a metodologia seguiu as etapas apresentadas abaixo:

. Revisão Bibliográfica;

. Seleção de técnicas e ferramentas para análise e visualização de dados;

. Desenvolvimento do ETL - *extract, transform and load;*

. Estudo dos tipos de coordenadas para representar os pontos dos clusters;

. Desenvolvimento de ferramenta de análises exploratórias;

. Desenvolvimento de ferramenta de visualização dos resultados da clusterização

O primeiro passo para a pesquisa foi fazer uma pesquisa bibliográfica em livros e artigos desenvolvidos anteriormente, para se buscar informações que permitissem criar um embasamento que ajudasse a guiar o projeto como um todo.

Em seguida, foi necessário fazer uma verificação dos métodos e definir quais seriam as técnicas utilizadas para composição da análise exploratória do projeto como um todo. Como demonstrado por Freitas et al. (2001), existem inúmeras técnicas de visualização. Alinhando com as perspectivas e necessidades do projeto, nossa melhor opção para desenvolvimento era utilizar a biblioteca R-Shiny da linguagem R.

O *Shiny* no geral é um sistema para desenvolvimento de aplicações web usando o R, que utiliza um pacote do R(shiny) e um servidor *web* (*shiny server*). Tem o principal objetivo de facilitar a criação de aplicações interativas na web que auxiliam na apresentação dos dados de maneira mais fluida, já que oferece também uma enorme variedade de recursos para visualização em conjunto com a interface gráfica. Uma de suas melhores características é conseguir integrar várias outras bibliotecas, além de suas integrações em web, que em conjunto conseguem enriquecer muito no desenvolvimento da aplicação como um todo.

Outro pacote que vale a pena ser ressaltado é o *Leaflet*. Esta biblioteca que tem origem no javascript tem como finalidade a criação de mapas extremamente interativos e precisos. Interage com incrível facilidade com o Shiny por ser um htmlwidget, o que permite gerar os mapas de forma direta no R.

O terceiro passo consistiu em desenvolver o ETL ou ETT - extrair, transformar e carregar, que é o processo de capturar os dados, tratá-los e deixá-los disponíveis em um repositório para o desenvolvimento das análises que se deseja desenvolver.

Os dados recebidos apresentavam os pontos dos clusters baseados em coordenadas cartesianas. Essas coordenadas precisaram ser transformadas para atender as bibliotecas de visualização. Para executar tais representações, foram utilizadas as coordenadas geodésicas latitude e longitude para representações tridimensionais e o sistema de coordenadas UTM para representar duas dimensões.

Nesse sentido, um dos principais fatores que contribuíram para essa preparação dos dados foi representá-los geograficamente no *Leaflet*.

A latitude é basicamente a distância medida em graus de qualquer ponto na superfície do planeta até a linha do Equador que corta o planeta horizontalmente. Já a longitude, ao invés de levar em consideração o Equador, se referencia com base no /meridiano e Greenwich. Ambas podem ser demonstradas como abaixo:

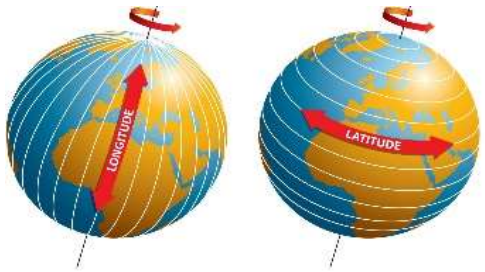

Figura 3 – Direções da latitude e longitude

O sistema de coordenadas UTM (Universal Transversa de Mercator) é um sistema de coordenadas cartesianas bidimensional que utiliza uma projeção cilíndrica do globo terrestre desenvolvido por Gerardus Mercator em que os meridianos, os paralelos e o Equador são representados por linhas retas verticais e horizontais. É divido em 60 fusos sequenciais de 6 graus de amplitude que são conhecidos como zonas UTM, representados abaixo:

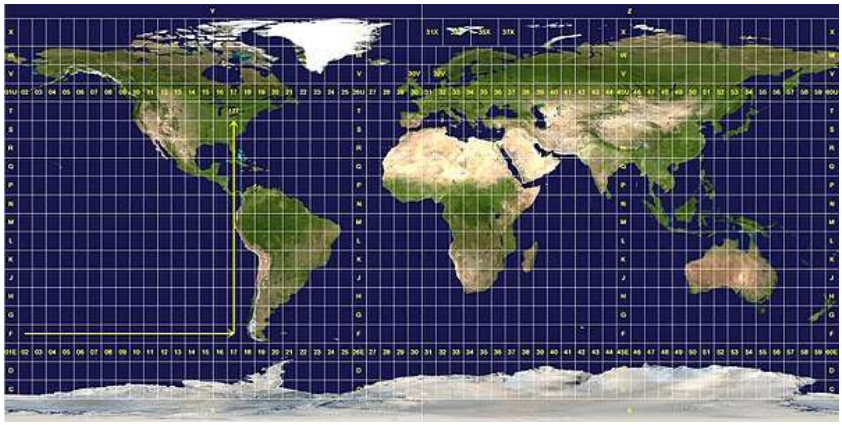

Figura 4 – Representação das divisões de zonas UTM

Pela ampla extensão territorial o solo brasileiro é dividido em 8 das 60 zonas existentes que devem ser consideradas no processo de conversão para que não haja distorção. Durante este processo utilizamos uma função presente no R para fazer essa conversão.

Após preparar os dados, foram utilizadas as técnicas de clusterização desenvolvidas nas etapas anteriores do projeto para que o output possa ser representado através das análises.

Tendo-se todas essas etapas finalizadas, o desenvolvimento final teve duas frentes e objetivos principais:

Desenvolver análise exploratórias correlacionadas com os resultados obtidos através dos algoritmos de clusterização;

Desenvolver um painel gráfico (*dashboard*) de visualização desses resultados.

4. RESULTADO E DISCUSSÃO

Entre os principais resultados obtidos, pode-se subdividi-los da seguinte maneira:

- Bases de dados resultantes dos processos de ETL;
- Análises exploratórias integradas e implementadas em uma aplicação web responsiva e dinâmica desenvolvida em *Shiny*.

A estrutura inicial da base de dados contém 4 colunas. A primeira coluna serve como chave primária e identificador linear, enquanto as colunas 2, 3 e 4 são colunas numéricas que correspondem a longitude, latitude e o peso de cada ponto,

respectivamente. Como os algoritmos são baseados em coordenadas com duas dimensões, o primeiro passo é transformá-las em medidas UTM para que essas medidas possam corresponder aos eixos x e y que serão utilizados no processo de clusterização.

A ferramenta tecnológica utilizada no estudo foi a linguagem R (R Core Team, 2020), e no próprio R há um algoritmo de conversão que têm como parâmetros de entrada a latitude e a longitude, e automaticamente já se consegue obter suas coordenadas UTM levando em consideração a zona que também é identificada conforme amostra abaixo:

Tabela 1 – Resultados das conversões de coordenadas geodésicas para UTM...

long	lat	codeUTM	X	Y
-51.07967	-21.68690	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	491758.8	7601826
-49.64338	-21.23494	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	640780.3	7651247
-46.97954	-22.05866	+proj=utm +zone=23 +south +ellps=WGS84 +datum=WG...	295725.3	7559354
-46.71126	-21.93333	+proj=utm +zone=23 +south +ellps=WGS84 +datum=WG...	323260.8	7573566
-46.59325	-22.48621	+proj=utm +zone=23 +south +ellps=WGS84 +datum=WG...	336094.5	7512483
-49.24047	-22.87993	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	680498.3	7468694
-47.87795	-22.59616	+proj=utm +zone=23 +south +ellps=WGS84 +datum=WG...	204105.3	7498327
-48.96470	-22.47156	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	709415.4	7513554
-47.89890	-23.55319	+proj=utm +zone=23 +south +ellps=WGS84 +datum=WG...	204054.3	7392246
-51.41898	-21.95166	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	456737.9	7572464
-49.06230	-20.52368	+proj=utm +zone=22 +south +ellps=WGS84 +datum=WG...	702043.8	7729368

Ao final do processo tem-se uma tabela com as coordenadas cartesianas e geodésicas de cada ponto, seu peso e uma nova coluna que identifica o cluster que a linha pertence, como na tabela 2.

Tabela 2 – Resultado do algoritmo de clusterização

p	x	y	w	cluster	lat	long
0	491758.81	7601826	96	9	-21.68690	-51.07967
1	640780.29	7651247	104	11	-21.23494	-49.64338
2	295725.34	7559354	96	3	-22.05866	-46.97954
3	23260.77	7573566	96	5	-21.93333	-46.71126
4	336094.47	7512483	96	3	-22.48621	-46.59325
5	680498.30	7468694	97	11	-22.87993	-49.24047
6	204105.35	7498327	101	8	-22.59616	-47.87795
7	709415.36	7513554	101	11	-22.47156	-48.96470
8	204054.32	7392246	104	8	-23.55319	-47.89890
9	456737.94	7572464	97	9	-21.95166	-51.41898
10	702043.84	7729368	95	7	-20.52368	-49.06230

A partir do resultado gerado é possível criar outra tabela para desenvolvimento da análise exploratória. Nesta nova tabela foram contempladas informações estatísticas que

demonstram o equilíbrio dos clusters que ajudam a interpretar se o algoritmo realizou ou não uma “boa clusterização” de acordo com o volume de dados que havia. Através da biblioteca dplyr foram agrupados os dados por cluster e foi utlizada a função *summarise* para construir o restante da tabela e trazer informações como o número de pontos, a média, a mediana , o erro percentual absoluto médio, a amplitude interquartílica, o desvio padrão, os pontos com peso mínimo e máximo, além da cor que o cluster terá em sua representação em um mapa. Subsequentemente pode-se visualizar o resultado de um processo como o citado acima na tabela 2:

Tabela 3 – Resultado com dados estatísticos dos clusters

cluster	N	MEDIA	MEDIANA	MAD	MIDSPREAD	DP	PESO	MIN	MAX	COLOR
1	94	99.84043	100	2.9652	5	3.084136	9385	95	105	blue
2	93	100.11828	100	2.9652	4	2.955648	9311	95	105	orange
3	94	99.78723	100	4.4478	5	3.082634	9380	95	105	green
4	93	100.11828	100	4.4478	6	3.269920	9311	95	105	purple
5	93	100.20430	100	4.4478	5	3.094733	9319	95	105	red
6	93	100.32258	101	2.9652	5	2.953117	9330	95	105	brown
7	94	99.65957	99	4.4478	6	3.342632	9368	95	105	fuchsia
8	93	100.08602	100	2.9652	4	2.958613	9308	95	105	gray
9	93	100.13978	100	4.4478	5	3.055968	9313	95	105	white
10	93	100.05376	100	2.9652	5	3.076438	9305	95	105	pink
11	93	99.87097	100	4.4478	6	3.111081	9288	95	105	yellow
12	93	100.03226	100	4.4478	5	3.204791	9303	95	105	royal blue

Após se obter todos esses resultados era necessário que tudo fosse adaptado para uma maneira gráfica que contribuísse para uma visão mais simples e objetiva.

Sobre a ferramenta gráfica que apresenta todos os resultados, esta foi desenvolvida com base em três estruturas internas à aplicação.

A primeira estrutura tem o intuito de facilitar todo processo citado anteriormente e permite fazer o *upload* da base de dados e definir suas propriedades como, por exemplo, o delimitador entre campos do arquivo. Automaticamente as operações são realizadas por trás da interface para que os demais passos possam continuar com fluidez. Posteriormente a isso, também é possível escolher qual será o algoritmo que será executado e ao lado direito da tela visualizar os resultados, resumidos ou não.

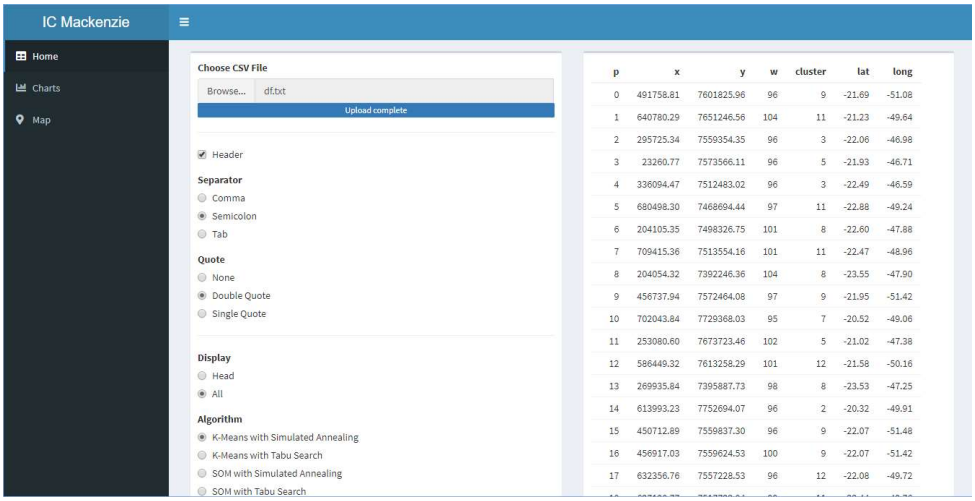

Figura 5 - Representação da primeira aba da aplicação

A segunda parte da ferramenta gráfica foi preparada para já se obter uma análise mais visual a respeito da clusterização. Dentro da aba correspondente, têm-se dois histogramas. O primeiro tem o objetivo de analisar a quantidade de ponto contidos em cada grupo e o segundo para analisar a distribuição dos pesos dentro de cada cluster. É válido ressaltar que pelo fato de serem construídos com a utilização da biblioteca *plotly* do R, eles apresentam propriedades interativas no canto superior direito que auxiliam e facilitam muito na análise, como mostrado em seguida nas figuras XX e YY

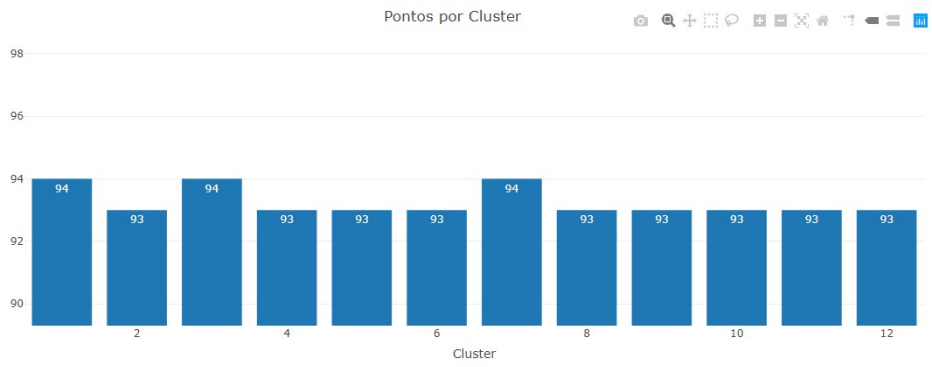

Figura 6 – Gráfico de pontos por cluster presente na aba 2 da aplicação

Figura 7 – Gráfico de peso por cluster presente na aba 2 da aplicação

A terceira e última parte da ferramenta gráfica é apresentada em uma aba que mescla tanto a forma visual, quanto a apresentação de resultados numéricos de maneira mais direta. Utilizando a biblioteca *Leaflet*, descendente do Javascript desenvolvido um mapa que recebe os parâmetros de latitude, longitude e a cor do cluster para representá-los geograficamente. Através deste recurso é possível enxergar com clareza a divisão e a efetividade dos algoritmos, assim como alguns *outliers* e outras informações que são de suma importância para realização da análise de maneira generalizada. A terceira aba tem o formato apresentado na figura 8.

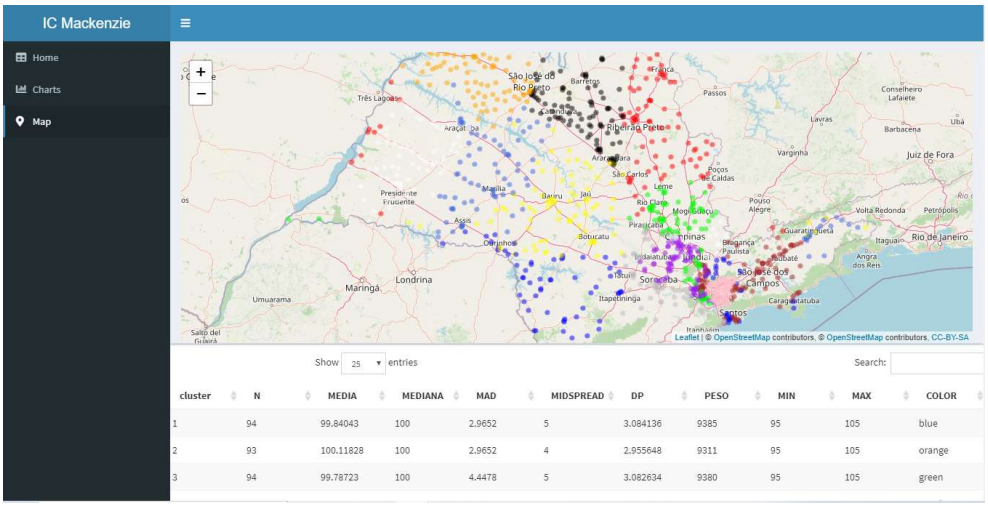

Figura 8 – Representação da aba 3 com mapa e tabela de análise dos clusters

Uma possibilidade é utilizar também os números abaixo do gráfico que possuem filtros interativos e dinâmicos ao final de cada coluna ou também utilizar um filtro geral localizado no canto superior direito da tabela XX.

Tabela 3 – Tabela de análise estatística dos clusters

Show 25 entries Search:

cluster	N	MEDIA	MEDIANA	MAD	MIDSPREAD	DP	PESO	MIN	MAX	COLOR
1	94	99.84043	100	2.9652	5	3.084136	9385	95	105	blue
2	93	100.11828	100	2.9652	4	2.955648	9311	95	105	orange
3	94	99.78723	100	4.4478	5	3.082634	9380	95	105	green
4	93	100.11828	100	4.4478	6	3.269920	9311	95	105	purple
5	93	100.20430	100	4.4478	5	3.094733	9319	95	105	red
6	93	100.32258	101	2.9652	5	2.953117	9330	95	105	brown
7	94	99.65957	99	4.4478	6	3.342632	9368	95	105	fuchsia
8	93	100.08602	100	2.9652	4	2.958613	9308	95	105	gray
9	93	100.13978	100	4.4478	5	3.055968	9313	95	105	white
10	93	100.05376	100	2.9652	5	3.076438	9305	95	105	pink
11	93	99.87097	100	4.4478	6	3.111081	9288	95	105	yellow
12	93	100.03226	100	4.4478	5	3.204791	9303	95	105	royal blue
cluster	N	MEDIA	MEDIANA	MAD	MIDSPREAD	DP	PESO	MIN	MAX	COLOR

Showing 1 to 12 of 12 entries Previous 1 Next

5. CONSIDERAÇÕES FINAIS

Tendo em levado em consideração todos os resultados obtidos ao final dessa pesquisa é válido ressaltar quais aspectos foram de suma importância para que fosse possível alcançar os resultados apresentados.

O processo de tratamento de dados auxilia em problemas futuros que exigem a transformação de coordenadas geodésicas em coordenadas bidimensionais, como o UTM, de maneira fluida e automática.

Ao utilizar bibliotecas e ferramentas que ofereceram recursos de alta qualidade, foi observado que além de ter um painel de rápida visualização e uma interação eficaz com usuário, também notou-se que com o desenvolvimento conseguiu-se abranger o projeto como um todo e concentrá-lo em uma única ferramenta que pudesse agregar valor e transmitir o que foi desenvolvido nas etapas anteriores.

Além disso, foi percebido que a fermenta gráfica pode ser utilizado de insumo para gerar inúmeros outros tipos de análises além da que foi realizada, uma vez que com resultados visuais, torna-se mais intuitivo ter novas visões e com isso ter novos problemas e soluções para sem desenvolvidos e incorporados a ferramenta até que ela atinja um grau de amadurecimento maior.

Assim, em termos de continuidade desta pesquisa, entende-se que poderiam ser desenvolvidos novos recursos para a ferramenta gráfica, particularmente com a apresentação de novos gráficos, assim como, a apresentação dos centróides dos clusters e de algumas estatísticas sobrepostas aos mapas, caminhando mais para uma ferramenta de informações geográficas.

6. REFERÊNCIAS

ALEXANDRE, D. S., TAVARES, J. M. R. S. (2007) Factores da Percepção Visual Humana na Visualização De Dados. CMNE/CILAMCE 13 a 15 de junho, 2007. Porto, Portugal.

BUSSAB, W. O.; MORETTIN, P. A. (2013) Estatística básica. 8 ed. São Paulo: Saraiva.

CARD, S. T., MACKINLAY, J. D., SCHEIDERMAN, B. (1999). Information Visualization in: Readings in Information Visualization, Using vision to think Stuart T. Card, Jock D. Mackinlay, Ben Scheiderman. San Francisco. Morgan Kauffman, pp 1-34

DEVORE, J. L. Probabilidade e estatística para Engenharia e Ciências. São Paulo. CENGAGE Learning. 2011.

FREITAS, C. M. D. S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; Cava R. A. (2001). Introdução à Visualização de Informações. RITA – Revista de Informática Teórica e Aplicada, Instituto de Informática UFRGS, Porto Alegre, RS, v. VIII, n. 2, p. 143-158, outubro, 2001. Disponível em https://lume.ufrgs.br/handle/10183/19398. Acesso em 25 de abril de 2019

GIL, T. L. e BARLETA, L. (2015). Formas Alternativas de Visualização de Dados na Área de História: Algumas Notas de Pesquisa. Revista de História, n. 173, p. 427-455, jul.-dez., 2015. São Paulo.

JÚNIOR, ANTONIO C. DA SILVA. Construindo um dashboard interativo com R, Shiny e Plotly. LinedIn, 2019. Disponível em: < https://www.linkedin.com/pulse/construindo-um-dashboard-interativo-com-r-shiny-e-da-silva-j%C3%BAnior/?originalSubdomain=pt>. Acesso em: 18 de junho de 2020.

MATPLOTLIB (2019) Matplotlib version 3.0.3. Disponível em https://matplotlib.org/. Acesso em 20/03/2019

MONTGOMERY, D. C.; RUNGER, G. C. Estatística Aplicada e Probabilidade para Engenheiros. 5 ed. Rio de Janeiro: Livros Técnicos e Científicos, 2013

MOORE, D. S.; NOTZ, W. I.; FLINGER, M. A. A Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC, 2017.

NASCIMENTO, H. A. D. e FERREIRA, C. B. R. (2005) Visualização de Informações – Uma Abordagem Prática. XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo, RS.

R STUDIO (2019) Shiny from R Studio. Disponível em https://shiny.rstudio.com/. Acesso em 25/03/2019.

SEABORN (2019). Seaborn: Statistical Data Visualization. Disponível em https://seaborn.pydata.org/. Acesso em 29/03/2019

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. (2016). “Introdução à Mineração de Dados: Com Aplicações em R”. 1ª Ed. Elsevier Editora Ltda. R.J. 277 p.

TABLEAU (2019) Tableau Desktop – Descrição do Produto. Disponível em https://www.tableau.com/pt-br. Acesso em 26/03/2019.

Contatos: zhanwesley@gmail.com e aavallim@mackenzie.br