UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

ENG07053 - TRABALHO DE DIPLOMAÇÃO EM ENGENHARIA QUÍMICA

Aplicação de Análise Multivariada em dados de Agrotóxicos na Água para Consumo Humano do Brasil de 2014 a 2018

Autor: Alexandre Rodrigues Conill Gomes

Orientador: Prof. Dr. Jorge Otávio Trierweiler

Co-orientadores: Msc. Lucas Ranzan, Msc. Luciano Barros Zini

Porto Alegre, Junho de 2019

Sumário

Sumário ii
Agradecimentos iii
Resumo iv
Lista de Figuras v
Lista de Tabelas vi
1 Introdução 1
2 Objetivo e delimitação do trabalho 2
3 Revisão Bibliográfica 3
3.1 Vigilância da Qualidade da Água para Consumo Humano no Brasil 3
3.1.1 Vigilância 4
3.1.2 Controle 6
3.1.3 SISAGUA 6
3.2 Análise Multivariada 8
3.2.1 PCA 9
3.2.2 Clusterização 10
3.3 A aplicação de análise multivariada em resultados de análises de qualidade de água para consumo humano 12
3.4 Knowledge Discovery in Databases (KDD) 13
3.5 Agrotóxicos 15
4 Materiais e Métodos 17
5 Resultados e Discussão 19
5.1 Identificação de erros no banco de dados 19
5.1.1 Dados não alimentados no SISAGUA 20
5.1.2 Dados Iguais ao VMP 21
5.1.3 Dados Iguais ao LD 22
5.1.4 Dados Iguais ao LQ 23
5.1.5 Critério de frequência 24
5.2 Aplicação da análise multivariada aos dados do SISAGUA 26
5.2.1 Tratamento dos dados 26
5.2.2 Aplicação do PCA 27
5.2.3 Clusterização 29
5.2.3.1 Análise de Silhueta 29
5.2.3.2 K-means 30
6 Conclusões e Trabalhos Futuros 33
7 Referências 35

Agradecimentos

Agradeço primeiramente à minha família, que, desde sempre, me incentivou a estudar e me proporcionou condições para realizar uma graduação. Esta não é a realidade da maioria dos brasileiros e sou muito grato pela oportunidade.

Agradeço aos amigos, namorada e companheiros de aula e trabalho, que tiveram papel marcante na minha formação pessoal, profissional e acadêmica. Saibam que sou inspirado por, aprendi, e sigo aprendendo muito com cada um de vocês.

Agradeço ao Msc. Luciano Barros Zini, co-orientador deste trabalho por toda a expertise no tema. Ao Msc. Lucas Ranzan pelos conhecimentos em programação e métodos de análise empregados. Ao professor Dr. Jorge Otávio Trierweiler pelos *insights* e contribuições ao trabalho.

Resumo

O Brasil é o maior consumidor de agrotóxicos do mundo. Este cenário gera preocupação quanto aos riscos à saúde provenientes da aplicação de agrotóxicos nas lavouras do país. O programa nacional de vigilância da qualidade da água para consumo humano (VIGIAGUA) é responsável pelo monitoramento da contaminação de agrotóxicos na água para consumo humano. Neste trabalho foi analisado o banco de dados com resultados de análises de controle de qualidade da água, disponibilizado no sistema de informação do programa de vigilância da qualidade da água para consumo humano (SISAGUA), para os 27 parâmetros de agrotóxicos previstos na portaria nacional de potabilidade. Foi aplicado método de análise multivariada que consistiu em aplicação de PCA e posteriormente clusterização K-means com os componentes principais formados. A dimensionalidade do banco de dados foi reduzida de 27 para 3 componentes principais, sendo atribuídos significados físicos a eles. Plotando os 3 PCs entre si, foi identificada que a maior variação dos dados corresponde à presença de Glifosato e Mancozebe, dois dos agrotóxicos mais vendidos no país. Foi detectada também correlação entre a presença de glifosato e de tebuconazol, dois agrotóxicos que tem efeito fungicida, e por isso não são detectados juntos. Anteriormente à aplicação do método foi necessário elaborar uma metodologia rigorosa para tratamento de dados com erro de *input*, expondo a fragilidade do SISAGUA. O trabalho busca ser ponto de partida e exemplo do potencial de aplicação de análise multivariada a dados de qualidade da água, podendo a mesma metodologia ser aplicada para regiões específicas ou para contaminantes diferentes.

Lista de Figuras

Figura 3-1: Fluxograma da Entrada de Dados no SISAGUA 7

Figura 3-2: (a) Deslocamento do eixo para centralização das variáveis; **(b)** Rotação do eixo em busca da maior variância explicada. 10

Figura 3-3: Processo Iterativo da Clusterização K-means 12

Figura 3-4: Exemplo - Análise de Silhueta com três clusters 11

Figura 3-5: Exemplo - Análise de Silhueta com quatro clusters 12

Figura 3-6: Etapas do KDD 14

Figura 4-1: Disposição das colunas do banco de dados original 17

Figura 4-2: Disposição das colunas do banco de dados Multivariado 18

Figura 5-1: Variância explicada para cada componente principal 27

Figura 5-2: Análise de Silhueta com K = 3 clusters 29

Figura 5-3: Análise de Silhueta com K = 4 clusters 30

Figura 5-4: Bi-plot dos Componentes principais com representação dos clusters identificados pelo K-means 31

Lista de Tabelas

Tabela 3-1: Distribuição dos Parâmetros do Padrão de Potabilidade Nacional em grupos 4

Tabela 3-2: Número mínimo mensal de amostras analisadas para os parâmetros cloro residual livre, turbidez, coliformes totais/*Escherichia coli*, segundo faixa populacional do município 5

Tabela 3-3 - Número mínimo mensal de amostras analisadas para o parâmetro fluoreto, segundo faixa populacional do município 5

Tabela 3-4: Informações sobre os agrotóxicos previstos no anexo XX da portaria de consolidação MS nº 5 2017 15

Tabela 5-1: Distribuição acumulada da alimentação de dados de agrotóxicos no sistema por região do Brasil de 2014 a 2018 20

Tabela 5-2: Distribuição acumulada do número de resultados de agrotóxicos iguais ao VMP de 2014 a 2018 21

Tabela 5-3: Distribuição acumulada do número de resultados de agrotóxicos iguais ao LD de 2014 a 2018 22

Tabela 5-4: Distribuição acumulada do número de resultados de agrotóxicos iguais ao LQ de 2014 a 2018 24

Tabela 5-5: Distribuição acumulada do número de resultados iguais ao VMP 25

Tabela 5-6: Progressão do tratamento de dados 27

Tabela 5-7: Loadings de cada componente principal 28

Lista de Abreviações

KDD – Knowledge Discovery in Databases

LD – Limite de Detecção

LQ – Limite de Quantificação

MS – Ministério da Saúde

PCA – Principal Component Analysis

PC – Componente Principal

SAA – Sistema de Abastecimento de Água

SAC – Solução Alternativa Coletiva

SAI – Solução Alternativa Individual

SISAGUA – Sistema de Informação do Programa Nacional de Vigilância da Qualidade da Água para Consumo Humano

VIGIAGUA – Programa Nacional de Vigilância da Qualidade da Água para Consumo Humano

VMP – Valor Máximo Permitido

1 Introdução

É sabido que a água pode ser fonte de doenças causadas por bactérias, vírus e protozoários, mas também pode trazer consigo contaminantes indesejados prejudiciais à saúde. A ONU estabeleceu como uma das metas da sua Agenda 2030 para o desenvolvimento sustentável o acesso universal à água segura e potável para todos (WHO, 2015).

O Brasil é, desde 2008, o maior consumidor de agrotóxicos do mundo (LOPES; ALBUQUERQUE, 2018), com uma venda total de 539.944 toneladas no ano de 2017 (IBAMA, 2017). Há legislações que regulamentam o registro, produção, comercialização e aplicação de agrotóxicos no país, porém a grande fragilidade está na fiscalização de medidas que garantam o cumprimento da legislação (LOPES; ALBUQUERQUE, 2018). Esse cenário gera preocupação quanto aos riscos à saúde associados a esta aplicação desenfreada de agrotóxicos, que eventualmente podem chegar até a água que bebemos.

O Programa de Vigilância da Qualidade da Água para Consumo Humano (VIGIAGUA) é responsável pelo monitoramento da qualidade da água sendo abastecida para os brasileiros. Parte das ações exercidas pelos profissionais envolvidos neste monitoramento está a coleta de amostras de água para posterior análise, seguindo uma diretriz prevista no anexo XX da Portaria de Consolidação MS nº 5/2017, que leva em conta principalmente a população de cada município do país. Outro braço de ações é realizado pelas próprias empresas de abastecimento, estabelecendo um controle de qualidade que é submetido à fiscalização.

Estas regras são aplicáveis a todas as formas de abastecimento de água do Brasil, o que faz com que o monitoramento seja fonte de muitos dados ao longo dos anos. A avaliação dos resultados serve como balizador para análises de série histórica e enquadramento da água como potável ou não, baseado no padrão de potabilidade previsto em lei. Toda a informação é alimentada no Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (SISAGUA), o que forma um grande banco de dados com os resultados das análises.

A análise multivariada do banco de dados do SISAGUA, visando compreender ou aprender a correlação entre os parâmetros analisados, é um campo com oportunidades de estudo ainda não exploradas no Brasil. Mundialmente, há alguns estudos, como PRAUS (2007) que aplicam métodos tradicionais de análise multivariada como PCA e clusterização a resultados de análise de água. Neste estudo serão relatados os principais métodos e resultados encontrados, e realizada a aplicação de métodos de análise multivariada com resultados do SISAGUA a nível nacional.

2 Objetivo e delimitação do trabalho

Este trabalho possui os seguintes objetivos primários:

- Realizar revisão bibliográfica sobre a aplicação de análise multivariada a resultados de análises de água para consumo humano;
- Aplicar métodos de análise multivariada não supervisionada ao banco de dados brasileiro do SISAGUA.

Além dos objetivos primários, os seguintes objetivos secundários:

- Buscar interpretações dos resultados provenientes da análise multivariada;
- Realizar análise quantitativa dos problemas encontrados na base de dados do SISAGUA.

O trabalho não busca quantificar as ocorrências de valores acima do Valor Máximo Permitido, enquadradas como não potáveis, mas sim obter correlações entre a ocorrência de mais de um contaminante, de maneira a descobrir que tipo de informação pode ser obtida a partir dos dados. Devido à relevância do tema, os contaminantes focos do trabalho foram escolhidos como sendo os agrotóxicos.

3 Revisão Bibliográfica

Neste capítulo será discutido o referencial teórico do trabalho, que serve como base para posterior aplicação de método computacional e discussão dos resultados. Os assuntos abordados estão divididos em ordem tal que facilite o entendimento do leitor.

O objetivo da seção 3.1 é descrever o funcionamento da vigilância da qualidade da água para consumo humano no Brasil, de maneira a contextualizar a origem dos dados posteriormente discutidos e analisados. A seção 3.2 introduz o conceito de análise multivariada e aborda as ferramentas empregadas no trabalho. Por fim, a seção 3.3 busca descrever o estado da arte da análise multivariada de análises de água para consumo humano, descrevendo quais estratégias foram empregadas por outros autores e expondo seus resultados.

3.1 Vigilância da Qualidade da Água para Consumo Humano no Brasil

Em 1997, a Organização Mundial da Saúde (OMS) caracterizou a vigilância da qualidade da água como a "Contínua e circunspecta avaliação da saúde pública e análise de segurança e aceitabilidade de formas de abastecimento de água para consumo humano" (WHO, 1997). Para Praus (2007), o monitoramento contínuo é essencial para que o abastecimento de água seja saudável e de alta qualidade, atendendo a todos os requisitos da legislação vigente.

No Brasil, o Programa Nacional de Vigilância da Qualidade da Água foi instituído pelo Ministério da Saúde (MS) em 2005, como instrumento de implementação das ações de vigilância. Acrescentando à definição da OMS, o VIGIAGUA é definido pelo MS como "O conjunto de ações adotadas continuamente pelas autoridades de saúde pública para garantir à população o acesso à água em quantidade suficiente e qualidade compatível com o padrão de potabilidade, estabelecido na legislação vigente, como parte integrante das ações de promoção da saúde e prevenção dos agravos transmitidos pela água."

O Anexo XX da Portaria de consolidação MS nº 5/2017 determina as diretrizes e atribuições do VIGIAGUA, elencando responsabilidades dos representantes do programa no Ministério da Saúde, nos Estados e Distrito Federal, além de nos Municípios. A mesma Portaria determina o Padrão de Potabilidade da água no Brasil, estipulando um valor máximo permitido (VMP) para 89 Parâmetros de qualidade da água. Dessa maneira, a água que não atenda o padrão de potabilidade, ou que ofereça risco à saúde, será classificada como não potável. A distribuição entre os grupos de parâmetros de qualidade do anexo XX da referida Portaria está detalhada na **Tabela 3-1.**

***Tabela 3-1**: Distribuição dos Parâmetros do Padrão de Potabilidade Nacional em grupos*

GRUPO	QUANTIDADE DE PARÂMETROS
INORGÂNICAS	15
ORGÂNICAS	15
AGROTÓXICOS	27
DESINFETANTES E SUBPRODUTOS DA DESINFECÇÃO	7
CIANOTOXINAS	2
RADIOATIVIDADE	2
ORGANOLÉPTICOS	21
TOTAL	**89**

Fonte: Anexo XX da Portaria de consolidação MS nº 5/2017

O padrão de potabilidade deve ser observado em todo o território nacional, o que não impede que estados e municípios definam critérios adicionais. É o caso do Rio Grande do Sul, que, por meio da Portaria SES/RS 320 de 2014, acrescentou mais 46 agrotóxicos aos 27 já presentes no padrão de potabilidade nacional. No RS, os parâmetros acrescentados pela Portaria estadual têm o mesmo valor legal dos originais previstos em Portaria nacional, e com isso as formas de abastecimento estão sujeitas ao monitoramento de 135 ingredientes ativos.

O Anexo XX da Portaria classifica as diferentes formas de abastecimento em três tipos, de acordo com as suas características de canalização e população abastecida. São elas: Sistema de Abastecimento de Água (SAA), Solução Alternativa Coletiva (SAC) e Solução Alternativa Individual (SAI). Os Sistemas de Abastecimento de Água (SAA) são um conjunto de obras civis, englobando a captação, produção, rede de distribuição, até as ligações prediais. Este tipo de abastecimento é comumente presente em cidades e aglomerações urbanas, estando sujeito ao controle de qualidade por parte da empresa distribuidora e à vigilância pelo setor saúde.

Já na zona rural, ou em regiões mais afastadas, onde a instalação de um SAA é inviável, faz-se necessária a implementação de soluções alternativas, podendo ser coletivas (SAC) ou individuais (SAI). Apesar de contemplarem também localidades com canalizações que levam água até as residências, a característica que distingue as soluções alternativas do SAA é a ausência de rede de distribuição. As SAI abastecem apenas uma família e seus agregados, e estão sujeitas à vigilância do setor saúde. Já as SAC fornecem água a mais de uma família, e com isso estão sujeitas ao controle de qualidade por parte do responsável pelo abastecimento, que pode ser uma empresa ou o próprio município, e também à vigilância.

3.1.1 Vigilância

São as ações realizadas pela autoridade de saúde pública, notadamente as secretarias municipais de saúde, secretarias estaduais de saúde e o Ministério da Saúde. Englobam o cadastro ou mapeamento das diferentes formas de abastecimento, as inspeções sanitárias das formas de abastecimento, e a coleta periódica de amostras de água para análise laboratorial.

A quantidade de amostras é determinada pela Diretriz Nacional do Plano de Amostragem (BRASIL, 2016), que determina um quantitativo mínimo de análises a serem realizadas para cada município, de acordo com a sua população. A relação de análises para os quatro parâmetros básicos com o quantitativo mínimo de análises de acordo com a população pode ser consultada nas **Tabela 3-2** e **Tabela 3-3**. Os municípios são obrigados a coletar amostras e encaminhá-las para análise dos 4 parâmetros básicos: coliformes totais/*Escherichia coli*, cloro residual livre (ou outro residual desinfetante), turbidez e fluoreto. Os três primeiros possuem frequência de coleta igual, enquanto que a exigência para o fluoreto é menor.

***Tabela 3-2**: Número mínimo mensal de amostras analisadas para os parâmetros cloro residual livre, turbidez, coliformes totais/Escherichia coli, segundo faixa populacional do município*

PARÂMETROS	**POPULAÇÃO (HAB.)**					
Cloro residual livre Turbidez Coliformes totais *Escherichia coli*	0 a 5.000	5.001 a 10.000	10.001 a 50.000	50.001 a 200.000	200.001 a 500.000	Superior a 500.001
	6	**9**	**8 + 1 PARA CADA 7,5 MIL HAB.**	**10 + 1 PARA CADA 10 MIL HAB.**	**20 + 1 PARA CADA 20 MIL HAB.**	**35 + 1 PARA CADA 50 MIL HAB.**

Fonte: BRASIL (2016)

***Tabela 3-3** - Número mínimo mensal de amostras analisadas para o parâmetro fluoreto, segundo faixa populacional do município*

PARÂMETROS	**POPULAÇÃO (HAB.)**					
Fluoreto	0 a 50.000	50.001 a 100.000	100.001 a 200.000	200.001 a 500.000	500.001 a 1.000.000	Superior a 1.000.001
	5	**7**	**9**	**13**	**18**	**27**

Fonte: BRASIL (2016)

O parâmetro de residual desinfetante, comumente cloro residual livre, permite avaliar a eficiência da desinfecção. Deve haver ao menos 0,2 mg/l de cloro residual livre em todos os pontos da rede de distribuição para evitar recontaminação da água ou proliferação microbiológica após o tratamento. É recomendado manter a concentração até 2,0 mg/l, pois acima disso já é possível sentir cheiro e gosto do cloro, o que pode dificultar a aceitação do consumidor, mas não apresenta risco à saúde. Valores acima de 5,0 mg/l configuram a água como não potável.

A turbidez é o parâmetro que quantifica a transparência da água devido à presença de partículas em suspensão. Está relacionada com a eficiência da etapa de filtração do tratamento, e tem como VMP 5 uT (Unidades de Turbidez). A filtração eficiente também é barreira para a remoção de protozoários da água, de maneira que a turbidez também pode ser um indicador do risco de contaminação por este tipo de microrganismo.

Os Parâmetros microbiológicos, Coliformes Totais e *Escherichia coli*, servem como indicação da presença de contaminação bacteriológica na água, e avaliam a eficiência da desinfecção. O grupo dos coliformes totais contempla bactérias que naturalmente vivem no solo, na água e nas plantas, mas não necessariamente apresentam risco à saúde. Já a *E. coli* é tido como um preciso indicador da contaminação da água por material fecal, sendo

a sua presença um indício da ocorrência de microrganismos patogênicos. A presença de *E. coli* é proibida em qualquer ponto da rede de distribuição.

O fluoreto é adicionado à água obrigatoriamente em todos os SAA brasileiros. A análise de fluoreto por parte da vigilância é necessária para avaliar o controle operacional do tratamento. O valor máximo permitido para fluoreto é 1,5 mg/L. A presença de Fluoreto na água para consumo humano é importante pois reduz a ocorrência de cárie dentária na população. O excesso de Fluoreto, por outro lado, pode causar a fluorose dentária em crianças em fase de formação dentária (BRASIL, 2016).

Outros parâmetros, como os agrotóxicos, devem ter plano de amostragem elaborado pelos estados em conjunto com os municípios. O plano de amostragem é elaborado pelo setor saúde, mas considera informações provenientes da secretaria da agricultura, ao considerar a sazonalidade da aplicação dos agrotóxicos e o tipo de cultura onde é aplicado. Dessa maneira, são definidos quando e onde serão realizadas coletas de água em busca da identificação de agrotóxicos.

3.1.2 Controle

São as ações de controle de qualidade realizadas pelos próprios prestadores de serviço que tratam e fornecem a água. Suas diretrizes estão previstas no Anexo XX da Portaria de consolidação MS nº 5/2017. Como uma das ações, as empresas responsáveis pelo abastecimento realizam análises mensais, trimestrais e semestrais dos parâmetros previstos no padrão de potabilidade (**Tabela 3-1**) e enviam os resultados para a autoridade de saúde pública.

O monitoramento é realizado na água bruta, na saída do tratamento e no sistema de distribuição. Os requisitos mínimos para água bruta são aqueles previstos para água doce de classe I, da resolução CONAMA 357 de 2005, válida para captação superficial, e da resolução CONAMA 396 de 2008, para subterrânea. Apenas os parâmetros em comum dos CONAMAS com o padrão de potabilidade nacional ficam registrados no SISAGUA. Fica dispensada a análise no sistema de distribuição se a substância não for detectada na saída do tratamento ou na água bruta.

Os agrotóxicos são parâmetros analisados semestralmente, e, ao coletar amostras, o prestador de serviços deve levar em conta avaliação dos seus usos na bacia hidrográfica do manancial de contribuição, bem como a sazonalidade das culturas em que são aplicados. No RS, por exemplo, os parâmetros adicionados pela Portaria SES/RS 320 de 2014 estão sujeitos ao controle semestral, porém não serão contemplados neste trabalho por não estarem disponíveis no SISAGUA.

3.1.3 SISAGUA

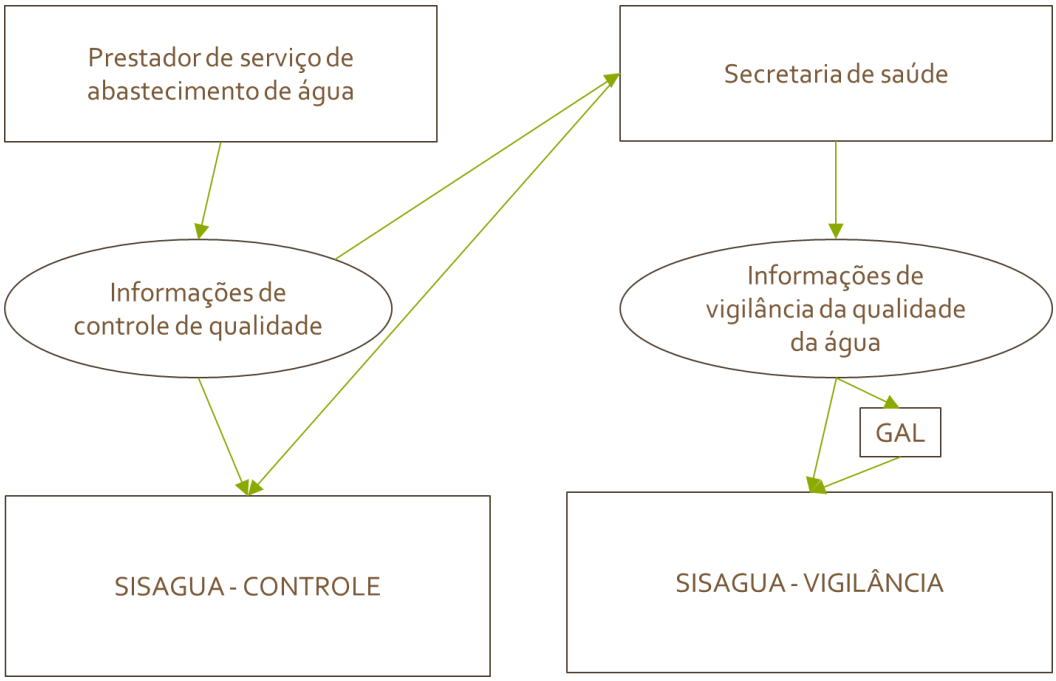
A grande quantidade de dados de Vigilância e de Controle, que são constantemente gerados conforme exposto nas subseções 3.1.1 e 3.1.2, torna necessária a existência de um sistema dedicado para armazenar as informações. Esta seção descreve o funcionamento

do Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (SISAGUA), que é a fonte de dados oficial do Ministério da Saúde.

O SISAGUA é um recurso utilizado no Brasil para cadastro das formas de abastecimento de água e dos dados de monitoramento da qualidade da água previstos na legislação. Disponibilizado na internet pelo Ministério da Saúde, busca auxiliar no gerenciamento e avaliação dos riscos à saúde associados ao abastecimento de água para consumo humano no país. É descrito como "a fonte de dados oficial para obtenção dos resultados dos indicadores institucionais do Vigiagua" (OLIVEIRA et al., 2019). O SISAGUA está na sua quarta versão, que foi implementada em 2014.

O SISAGUA é gerenciado pelo MS, e os dados são alimentados no sistema pelos prestadores de serviço, pelas Secretarias de Saúde dos Estados e Municípios. Os três principais módulos de inserção de informações são denominados Cadastro, Controle e Vigilância. Um fluxograma da entrada de dados no sistema pode ser visto na **Figura 3-1**: Fluxograma da Entrada de Dados no SISAGUA

***Figura 3-1**: Fluxograma da Entrada de Dados no SISAGUA*



Fonte: Adaptado de (OLIVEIRA et al., 2019)

As informações de Cadastro referem-se às características de cada forma de abastecimento, como a classificação de acordo com a legislação (SAA, SAC ou SAI), informações sobre a captação e sobre as etapas do tratamento, e também a população abastecida. Para o caso de SAA e SAC, as informações são fornecidas pela empresa de abastecimento, podendo ser inseridas no sistema pela própria empresa ou repassadas às Secretarias de Saúde para inserção. No caso de SAI, é de responsabilidade inteiramente das Secretarias de Saúde o mapeamento da população e o cadastro no sistema.

Os dados de Cadastro têm importância estratégica pois servem de base para a alimentação de dados dos outros módulos - Todos os dados de Controle e Vigilância são atribuídos a alguma forma de abastecimento que deve estar obrigatoriamente cadastrada.

As informações de cadastro devem ser revalidadas sempre na virada do ano, com o objetivo de identificar alterações no tratamento ou população abastecida.

Os dados de Controle contemplam os parâmetros analisados pelos prestadores de serviço de abastecimento, com frequência mensal, trimestral e semestral, seguindo o recomendado pelo anexo XX da portaria de consolidação n 5 2017. As informações são enviadas para as secretarias de saúde, mas há a opção de alimentação pelas próprias empresas de abastecimento.

Já os dados de vigilância dependem inteiramente do setor saúde. São os resultados de análises de monitoramento da qualidade da água realizadas pelos municípios, conforme previsto na Diretriz nacional do Plano de amostragem.

Todos os dados são alimentados no sistema de forma inteiramente manual, o que pode ser fonte de erros de digitação. Recentemente foram implementadas duas soluções para reduzir essa provável fonte de erro, em formato de *webservice*. A primeira é um sistema que permite a carga automática dos dados de controle por parte das empresas de abastecimento. A segunda é a integração com o Gerenciador de Ambiente Laboratorial (GAL), sistema usado pelos laboratórios de saúde pública que realizam as análises de vigilância, o que permite que estes resultados sejam carregados no sistema facilmente, dependendo apenas de validação por parte da Secretaria de Saúde – Apertar um botão no SISAGUA. É importante ressaltar que nem todos os estados e empresas de saneamento possuem esta integração, o que torna bastante frágil o processo de alimentação dos dados.

3.2 Análise Multivariada

A análise multivariada é tradicionalmente descrita como o estudo estatístico de experimentos em que múltiplas medições são realizadas em cada unidade experimental, e nas quais a relação entre as medições são relevantes e importantes para o entendimento do experimento (OLKIN; SAMPSON, 2001). O espectro de aplicações da análise multivariada é muito amplo, mas sempre busca a identificação de grupos através de características experimentais. Artigos recentes aplicaram o conceito para diferenciar tipos de vazios cósmicos (COUSINOU et al., 2019), e para diferenciar chocolates quanto a sua origem (LE GRESLEY; PERON, 2019), por exemplo.

Tipicamente, o número de variáveis necessárias para representar todas as propriedades de um conjunto de dados é menor do que a verdadeira dimensão do conjunto. Este número reduzido de variáveis representativas é chamado de dimensionalidade intrínseca. Técnicas de redução de dimensão são usadas para encontrar e explorar conjuntos de dados com pequena dimensionalidade intrínseca, de maneira a representá-los com menos variáveis (VAN DER MAATEN, 2009).

Para o caso específico de resultados experimentais de múltiplas características químicas de uma mesma amostra, a análise multivariada pode ser chamada de quimiometria. A quimiometria surgiu no final dos anos 60, principalmente devido a disponibilidade de resultados provenientes de equipamentos que registravam informações com muitas variáveis (GELADI, 2003).

Um conjunto de dados multivariado pode ser denotado como *X.* Os resultados no conjunto de dados são dispostos em uma matriz com elementos do formato $x_{i,j}$, em que o contador i representa o número de observações, e j o número de variáveis do conjunto de dados. A **Equação 1** representa um conjunto de dados multivariado genérico.

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} \\ \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} \end{bmatrix} \tag{1}$$

Um teste estatístico para a identificação de *outliers* em um conjunto de dados multivariado é o teste T^2 de Hotelling. O teste realiza j testes t-Student simultaneamente, levando em conta correlações que as variáveis possam possuir entre si. (TAVARES; RAMOS, 2016)

3.2.1 PCA

A análise de componentes principais, do inglês *Principal Component Analysis* (PCA), é uma das mais antigas e mais utilizadas técnicas exploratórias de análise multivariada. É uma análise não supervisionada pois não leva em conta a natureza dos dados analisados, realizando apenas uma busca por similaridades estatísticas. É uma técnica de redução de dimensionalidade que cria uma representação comprimida das informações na forma de novas variáveis ortogonais chamadas de componentes principais (PCs), que nada mais são do que combinações lineares das variáveis originais (ABDI; WILLIAMS, 2010).

O primeiro passo do PCA é a normalização das variáveis. Essa etapa é importante pois as variáveis podem possuir escalas diferentes, ou ainda representam grandezas físicas distintas. Caso não seja realizada a normalização, o resultado do PCA será afetado, pois a maior variância será atribuída ao componente que tiver a maior escala (MANNING-DAHAN, 2017). O conjunto de dados normalizado pode ser denominado *Y*.

Após a normalização, o próximo passo é a criação da matriz de covariância do conjunto de dados normalizado. Ela pode ser obtida através da multiplicação da matriz Y pela sua transposição. A matriz de covariância pode ser denotada *S* e está representada na **Equação 2**.

$$S = Y^T Y = \begin{bmatrix} var(x_1) & covar(x_1, x_2) & \cdots & covar(x_1, x_i) \\ covar(x_1, x_2) & var(x_2) & \cdots & covar(x_2, x_i) \\ \vdots & \vdots & \ddots & \vdots \\ covar(x_1, x_i) & covar(x_2, x_i) & \cdots & var(x_i) \end{bmatrix} \tag{2}$$

Com a matriz de covariância, é realizada a decomposição em autovalores, obtendo uma matriz Z, de dimensão 1 *x i,* composta pelos autovetores de S, e uma matriz diagonal λ, de dimensão *i x i,* com os autovalores na diagonal principal. A relação entre Z, S e λ é descrita pela **Equação 3**.

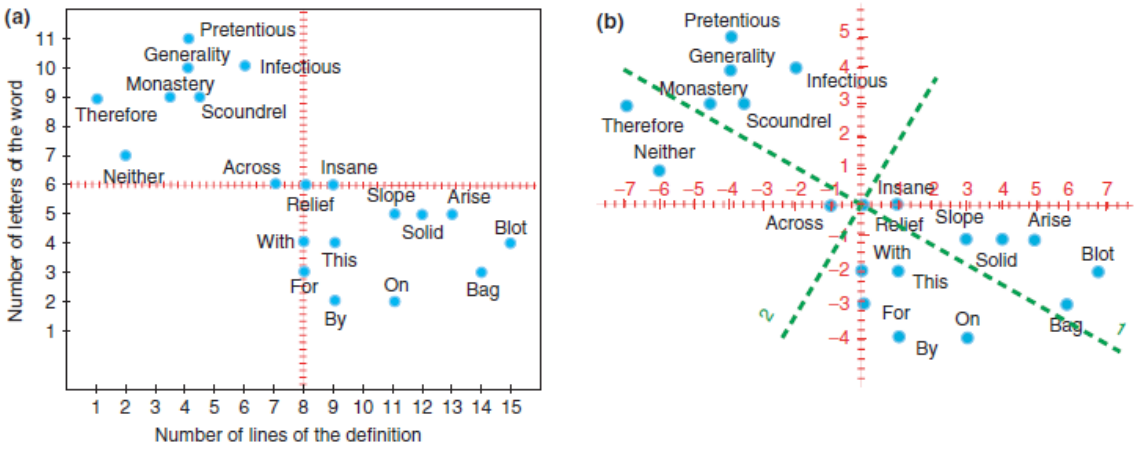
$$S = Z\lambda Z^{-1} \tag{3}$$

Por fim, a matriz Y é multiplicada pela matriz Z, para obter um novo conjunto de dados, chamado T, de mesma dimensão de X, em que as variáveis são os componentes principais. Para o caso de redução de dimensionalidade, pode-se truncar as colunas da matriz Z, de maneira a reduzir a dimensionalidade do conjunto de dados. A matriz Z também é chamada de *loadings*. Por fim, ocorre a multiplicação do banco de dados pela matriz de *loadings*, concretizando a transformação linear, operação representada pela **Equação 4**.

$$T = YZ \tag{4}$$

Para o caso mais simples possível, em que as amostras estejam descritas por apenas duas variáveis em um plano cartesiano, a criação dos componentes principais pode ser facilmente interpretada com a representação geométrica ilustrada na **Figura 3-2**. Na **Figura 3-2 (a)** ocorre a centralização dos pontos em relação aos eixos ortogonais. Já na **Figura 3-2 (b)** ocorre a rotação dos eixos em busca do ângulo em que haja maior variância explicada. Para o caso de haver mais variáveis, e ao invés de um plano bidimensional existir um plano multivariado, as etapas geométricas são as mesmas, porém é impossível de representá-las em desenho.

Figura 3-2: (a) *Deslocamento do eixo para centralização das variáveis;* ***(b)*** *Rotação do eixo em busca da maior variância explicada.*



Fonte: ABDI; WILLIAMS, 2010

3.2.2 Clusterização

Clusterização é uma técnica de agrupamento de dados em grupos, ou *clusters*, de maneira que os objetos dentro de um *cluster* sejam muito semelhantes entre si, mas ao mesmo tempo sejam diferentes dos objetos pertencentes a outros *clusters* (JIAWEI HAN, MICHELINE KAMBER, 2011). A clusterização pode ser interpretada como um modo de aprendizado não supervisionado, pois o agrupamento não está baseado em classes atribuídas aos dados, mas sim nos próprios valores associados a eles. (HAND, 2008)

3.2.2.1 Análise de Silhueta

A qualidade dos *clusters* pode ser determinada ao avaliar se as diferenças entre pontos de dados dentro de um *cluster* são pequenas frente às diferenças entre pontos em *clusters* diferentes. ROUSSEEUW (1987) propôs uma técnica visual que auxilia a compreensão, e que pode servir para identificar quantos *clusters* "naturais" há em um determinado conjunto de dados.

A técnica é chamada de Análise de Silhueta, e gera gráficos que auxiliam a interpretação do usuário sobre a quantidade de *clusters* naturais nos dados. Cada área marcada no gráfico corresponde a um *cluster*. A altura de cada segmento no eixo vertical é reflexo da quantidade de pontos pertencentes ao *cluster*, onde cada incremento corresponde a um ponto. O eixo horizontal representa o *score* de silhueta de cada ponto. O *score* é a diferença entre a distância média do ponto a todos os outros pontos do *cluster* mais próximo e a distância média do ponto e todos os outros pontos dentro do seu próprio *cluster*. Este valor é então dividido pelo maior valor entre ambos, fazendo com que a amplitude do *score* esteja entre -1 e 1. Quanto mais próximo de 1, maior é a semelhança do ponto com seu *cluster*, e maior é sua dessemelhança com o *cluster* vizinho. Valores próximos de zero indicam que o ponto se encontra no limiar entre dois *clusters* e valores negativos representam pontos que estão mais próximos aos pontos do *cluster* vizinho do que ao seu próprio *cluster*. De maneira geral, o número ideal de *clusters* para um conjunto de dados é aquele onde as alturas dos segmentos são semelhantes e onde a média dos *scores* seja a mais próxima de 1, mas cabe também ao usuário avaliar os resultados.

Um exemplo de aplicação é descrito nas **Figura 3-3** e **Figura 3-4**. Percebe-se que a opção por 4 *clusters* é mais adequada. Desta maneira, a altura das seções hachuradas é mais homogênea, e também o valor médio dos *scores* é maior, indicando maior semelhança das amostras ao seu *cluster*.

A reta vertical tracejada representa o *score* médio da análise, que é o resultado numérico da análise. A homogeneidade dos *clusters* pode ser analisada comparando-se a largura das regiões marcadas com a média, buscando sempre a situação em que ambas estejam parecidas, para obter *clusters* homogêneos.

***Figura 3-3**: Exemplo - Análise de Silhueta com três clusters*

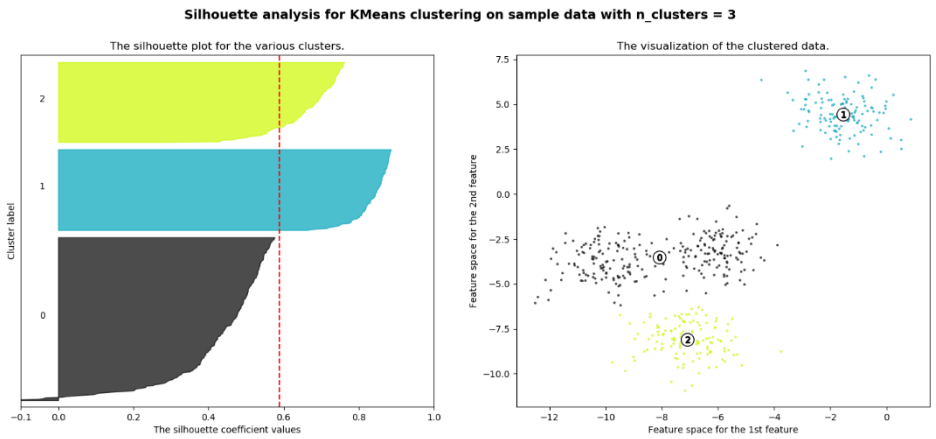
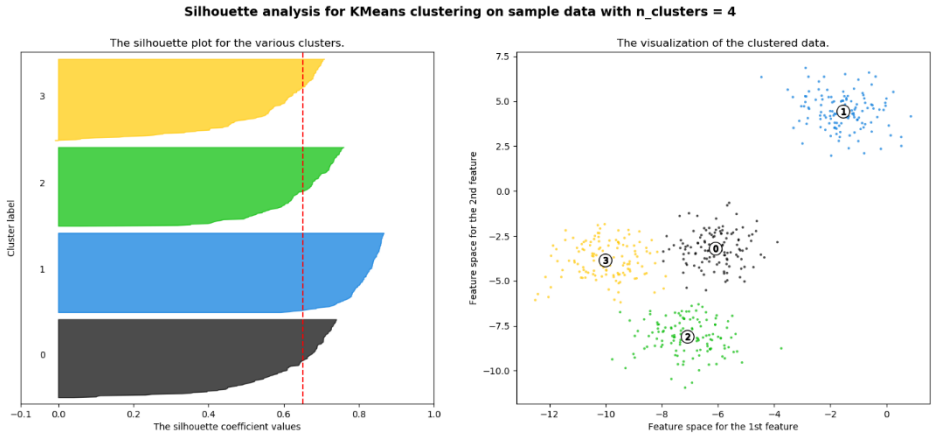


***Figura 3-4**: Exemplo - Análise de Silhueta com quatro clusters*

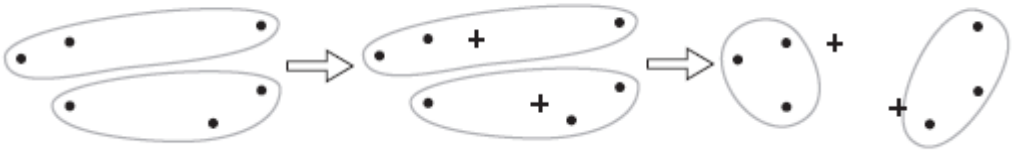


Fonte: PEDREGOSA et al. (2011)

3.2.2.2 K-means

O algoritmo de clusterização K-means é um dos mais utilizados para identificar grupos em conjuntos de dados. Este tipo de clusterização realiza um processo iterativo, em que duas operações são feitas em cada iteração. São estipulados grupos entre os pontos de dados, e estes grupos tem o seu centroide calculado. Com base na posição dos centroides, os pontos são redistribuídos entre os grupos de acordo com a sua proximidade de cada centroide (DARLINGTON; GHANEM; TO, 1993). Novos grupos são formados e assim o processo se repete. Uma representação gráfica do processo iterativo pode ser observada na **Figura 3-3.**

***Figura 3-5:** Processo Iterativo da Clusterização K-means*



Fonte: (DARLINGTON; GHANEM; TO, 1993)

3.3 A aplicação de análise multivariada em resultados de análises de qualidade de água para consumo humano

PRAUS (2007) analisou dados de qualidade da água para consumo humano de uma cidade da República Tcheca através de PCA. Foram coletadas 126 amostras de água dos sistemas de abastecimento presentes na cidade e para cada amostra foram analisados 18 parâmetros físico-químicos e microbiológicos. Após aplicação de PCA, foi possível reduzir a

dimensionalidade do conjunto de dados para 6 componentes principais, que explicavam aproximadamente 83% da variância dos dados. Significados físicos foram atribuídos aos 6 componentes principais através de uma análise dos seus *loadings* (Matriz Z, referida na subseção 3.2.1). Ao primeiro PC foi atribuída a característica "salinidade", e ao segundo "acidez", por exemplo.

No mesmo trabalho, PRAUS (2007) aplicou ainda clusterização aos resultados do PCA, e conseguiu identificar 3 *clusters*, quando plotou-se PC1 *versus* PC2. O autor atribuiu significado aos 3 *clusters*, afirmando que cada *cluster* correspondia a uma das estações de tratamento da cidade. O autor conclui o artigo constatando que técnicas de análise multivariada podem ser muito úteis quando não se possui conhecimento prévio sobre os dados de qualidade da água.

CHOWDHURY, CHAMPAGNE, MCLELLAN (2008) analisaram a formação de trihalometanos, que são subprodutos da desinfecção da água, em sistemas de abastecimento de água na província de Ontário, no Canadá. Ao contrário de estudos anteriores, que analisavam a relevância de determinados fatores em amostras sintéticas de água, ou usavam amostras pontuais de estações de tratamento de água específicas, este foi o primeiro estudo que buscou analisar a formação de trihalometanos em um espectro amplo de estações de tratamento de água.

O estudo de CHOWDHURY, CHAMPAGNE, MCLELLAN (2008) analisou 162 sistemas de abastecimento de água de Ontário, com dados de 2000 a 2004. Foi elaborada matriz de gráfico de dispersão entre cada variável *(bi-plots)*, e com isso foram identificados os parâmetros que mais colaboravam com a formação de THM. Adicionalmente, foi utilizada a metodologia PCA, e significados físicos foram atribuídos aos componentes principais através dos *loadings*. O trabalho tem o objetivo de ser um ponto de partida para a criação de um modelo preditivo da formação de trihalometanos.

EMENIKE; TENEBE; JARVIS (2018) realizaram pesquisa a partir de 63 amostras coletadas em 21 diferentes localidades. Foi avaliada a concentração média de fluor nas amostras e avaliada a correlação entre a presença de fluor e de outros resultados analíticos da água. Foi encontrada correlação entre outros parâmetros, como a presença de sólidos dissolvidos, condutividade elétrica, sulfato e magnésio. O trabalho utilizou os *loadings* do PCA para analisar a correlação entre a ocorrência de fluoreto elevado e outros parâmetros.

3.4 Knowledge Discovery in Databases (KDD)

Segundo Fayyad (1996), o método tradicional de transformação de dados em conhecimento está baseado em análise manual e interpretação. O que acontece hoje em dia é que os bancos de dados estão crescendo de duas maneiras: o número *N* de registros ou ocorrências, e o número *d* de campos, ou atributos por registro, o que dificulta a análise através do método tradicional.

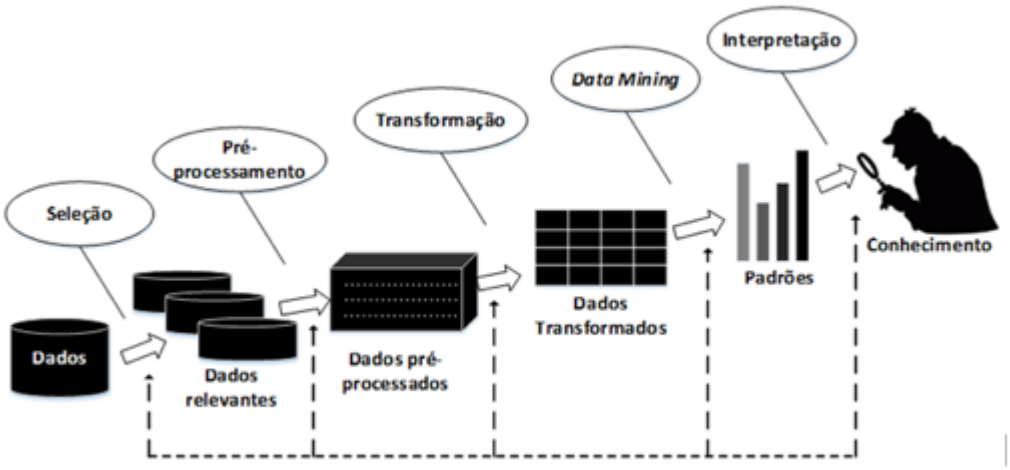
O KDD (Descoberta de Conhecimento em Banco de Dados), por outro lado, é uma metodologia que busca a transformação de dados brutos em informações relevantes e compreensíveis (PRASS, 2012). Para Fayyad (1996) o KDD é um processo não trivial já que alguma técnica de busca ou inferência é envolvida, ou seja, não é apenas um processo de computação direta. Os padrões descobertos devem ser válidos com certa confiança, novos,

potencialmente úteis (trazer algum benefício) e compreensíveis (se não imediatamente então depois de interpretação ou pós-processamento).

O Data Mining é uma das etapas do KDD. Na prática, grande parte dos esforços podem ser direcionados à devida formulação do problema – fazer a pergunta certa – do que apenas em otimizações do algoritmo utilizado (FAYYAD, 1996).

O processo KDD pode ser classificado em nove etapas, aplicadas para chegar de um estágio em que se possui apenas dados, a um cenário em que se obteve conhecimento, conforme FAYYAD (1996) e PRASS (2012). As etapas do processo estão representadas na **Figura 3-6**.

***Figura 3-6**: Etapas do KDD*



Fonte: Adaptado de PRASS (2012)

A primeira etapa do KDD é a seleção. Trata-se de estudar e compreender o domínio da aplicação, e envolve definir o objetivo da análise. Após, ocorre a definição da base de dados com a qual será realizado o trabalho. Nessa etapa, além de simplesmente escolher um banco de dados para a análise, é possível fracionar um banco de dados existente e obter um novo banco contendo apenas os dados mais relevantes com base em algum critério de seleção. A terceira etapa é o pré-processamento e limpeza dos dados. São aplicadas técnicas de remoção de ruído e discrepâncias, os chamados *outliers*. Também está incluída nesta etapa a decisão sobre a abordagem tomada para campos de dados sem valores. A quarta etapa é a transformação, em que o conjunto de dados é preparado para a aplicação da posterior da mineração de dados. A quinta é decidir, através do objetivo definido na primeira etapa, qual método de mineração de dados será utilizado. A sexta é uma análise exploratória e escolha de hipóteses para o modelo. A sétima é o processo de mineração de dados, em que o algoritmo é aplicado e são encontrados padrões. A oitava se trata da interpretação dos padrões encontrados, levando em conta todas as etapas anteriores do processo. A nona e última etapa é a tomada de ação com base nos resultados obtidos.

As etapas do KDD são executadas em ordem, mas podem ser revistas a qualquer momento (FAYYAD, 1996). Na prática as considerações feitas em etapas preliminares podem ser alteradas com base nos resultados posteriores, em uma espécie de retroalimentação.

3.5 Agrotóxicos

A Lei Federal nº 7.802, de 1989, introduziu o termo agrotóxico no Brasil. Segundo a lei, os agrotóxicos são compostos de substâncias químicas destinadas ao controle, destruição ou prevenção, direta ou indiretamente, de agentes patogênicos nocivos a plantas e animais.

Os agrotóxicos podem ser classificados de acordo com o objetivo de sua aplicação, além de suas características físico-químicas (PLIMMER, 2001 CORDATA, 2009). Os agrotóxicos são classificados em herbicidas, inseticidas, fungicidas, nematicidas, acaricidas e raticida. Os principais grupos físico-químicos dos agrotóxicos listados são organoclorados, organofosforados e carbamados. A lista completa com as características dos agrotóxicos e seus metabólitos listados no padrão de potabilidade está disponível na Tabela 3-4. Nota-se que há 38 substâncias listadas, e não 27, pois diversos parâmetros se referem à soma de concentração de um agrotóxico e seu(s) metabólito(s).

A venda de agrotóxicos no Brasil é registrada pelo IBAMA. Segundo relatório de venda de agrotóxicos em 2017 (IBAMA, 2018), os três agrotóxicos mais vendidos no Brasil são o glifosato, com 173 mil toneladas, seguido do herbicida 2,4 D, que teve 57 mil toneladas vendidas. O terceiro agrotóxico mais vendido é o mancozebe, com 30 mil toneladas de ingrediente ativo vendidas em 2017. O padrão de potabilidade brasileiro contempla 27 parâmetros, porém apenas 11 deles tem comercialização permitida no País. A tabela 3-4 exibe também a venda total dos agrotóxicos permitidos em 2017.

***Tabela 3-4**: Informações sobre os agrotóxicos previstos no anexo XX da portaria de consolidação MS nº 5 2017*

Substância	Grupo Químico	Objetivo da Aplicação	Tonelada de ingrediente ativo vendido [3] (2017)
2,4 D[1]	Ácido ariloxialcanóico	Herbicida	57389
2,4,5 T [2]	Clorofenoxiácido	Herbicida	**
Alaclor [1]	Cloroacetanilida	Herbicida	*
Aldicarbe [2]	Carbamado	Inseticida, Acaricida, Nematicida	*
Aldicarbesulfona [2]	Carbamado	Inseticida, Acaricida, Nematicida	**
Aldicarbesulfóxido [2]	Carbamado	Inseticida, Acaricida, Nematicida	**
Aldrin[2]	Organoclorado	inseticida	*
AMPA [1]	Glicina substituída	herbicida	**
Atrazina[1]	Triazina	herbicida	24731
Benomil [2]	Benzimidazol	Fungicida, Acaricida	*
Carbendazim [1]	Benzimidazol	fungicida	3748

Substância	Grupo Químico	Objetivo da Aplicação	Tonelada de ingrediente ativo vendido [3] (2017)
Carbofurano [2]	Carbamado	Inseticida, Nematicida, Acaricida	*
Clordano [2]	Organoclorado	inseticida	*
Clorpirifós [1]	Organofosforado	Inseticida, Formicida e Acaricida	6471
Clorpirifós-oxon [1]	Organofosforado	Inseticida, Formicida e Acaricida	**
DDD [2]	Organoclorado	Insecticide	**
DDE [2]	Organoclorado	Inseticida	**
DDT [2]	Organoclorado	Inseticida	*
Dieldrin [2]	Hidrocarboneto Clorado	Inseticida	**
Diuron [1]	Uréia	Herbicida	6999
Endossulfan [2]	Organoclorado	Insecticide, Acaricide	*
Lindano [2]	Organoclorado	Inseticida, Acaricida	*
Mancozebe [1]	ditiocarbamato	Fungicida	30815
Metamidofós [2]	Organofosforado	Inseticida, Acaricida	*
Metolacloro [2]	Chloroacetamide	Herbicida	*
Molinato [2]	tiocarbamado	Herbicida	*
Parationa Metílica [2]	Organofosforado	Inseticida	*
Pendimentalina [2]	Dinitroanilina	Herbicida	*
Permetrina [1]	Piretróide	Inseticida e formicida	83
Profenofós [1]	Organofosforado	Inseticida e acaricida	*
Simazina Sulfato de [1]	Triazina	Herbicida	308
Tebuconazol [1]	Triazol	Fungicida	2065
Terbufós [2]	Organofosforado	Inseticida, Nematicida	*
Trifluralina [1]	Dinitroanilina	Herbicida	1940
Endossulfan Alfa [2]	Organoclorado	Inseticida, Acaricida	**
Endossulfan Beta [2]	Organoclorado	Inseticida, Acaricida	**
Endrin [2]	Organoclorado	Inseticida, Avicida,	*
Glifosado [1]	Glicina substituída	Herbicida	173151

* Produto proibido no Brasil, ** Metabólito da decomposição de agrotóxico

Fonte: (1) ANVISA, (2) UNIVERSITY OF HERTFORDSHIRE, (3) IBAMA, 2018

4 Materiais e Métodos

A metodologia aplicada neste trabalho utiliza como base os fundamentos do KDD. Neste capítulo serão descritos os métodos aplicados e considerações realizadas no trabalho. As rotinas aqui implementadas fizeram uso dos *softwares MS Excel* versão 1902 e *Python* versão 3.7.

Na primeira etapa do KDD, a seleção do banco de dados, haviam disponíveis dados de controle e de vigilância no SISAGUA. Foram escolhidos dados de controle por contemplarem mais variáveis, o que justificaria a aplicação de análise multivariada, objetivo inicial do trabalho. O Ministério da Saúde, através do *website* dados.gov.br, disponibilizou o banco de dados do SISAGUA para consulta pública em 2018. Os dados utilizados neste trabalho foram os dados de controle semestral de 2014 até 2018, que estava disponível à época de elaboração.

Na segunda etapa do KDD, foi definido o subconjunto de dados alvo do trabalho, os agrotóxicos. Dado o grande consumo de agrotóxicos no Brasil e a crescente preocupação com o assunto, este foi definido como sendo o foco do trabalho. O banco de dados escolhido para a elaboração do trabalho contemplou os 27 parâmetros de agrotóxicos do padrão de potabilidade nacional.

Na terceira etapa, o pré-processamento, foram encontrados valores não numéricos como resultado: “<LD” (Menor que o limite de detecção do método) e “<LQ” (Menor que o limite de quantificação do método). Como para a análise dos dados todos os resultados devem ser números, os resultados “<LD” e “<LQ” foram aproximados como sendo zero.

A quarta etapa abrangeu a transformação dos dados de um formato de lista para um formato multivariado. Os dados foram disponibilizados pelo governo federal em arquivos no formato CSV, sendo um arquivo para cada ano, como ilustrado na **Figura 4-1Figura 5-1**. Os arquivos originais possuíam uma coluna que discriminava o parâmetro e outra que discriminava o resultado, o que fazia com que a lista somada de 2014 a 2018 fosse muito grande, com mais de 2.000.000 de linhas. As linhas foram ordenadas de forma a abrangerem um semeste, de um ponto do tratamento (Água Bruta, Tratada e Rede de distribuição), em um semestre.

***Figura 4-1**: Disposição das colunas do banco de dados original*

$Amostra$	$Parâmetro$	$Resultado$
$A1$	$P1$	$R1$
$A1$	$P2$	$R2$

Foi necessário o uso de ferramenta robusta para transformar o banco de dados em um formato multivariado. O número máximo de linhas de uma planilha de *MS Excel*, é 1.048.576 linhas (MICROSOFT), e por isso foi utilizada a ferramenta *Power Query*, da *Microsoft*, para ler os arquivos CSV. No *MS Excel* foi criada uma tabela dinâmica que realizou a transformação, colocando cada amostra em uma linha e os diferentes resultados em colunas, conforme ilustrado na **Figura 4-2**.

***Figura 4-2**: Disposição das colunas do banco de dados Multivariado*

$Amostra$	$Parâmetro\ 1$	$Parâmeto\ 2$
$A1$	$R1$	$R2$
$A2$	$R3$	$R4$

Após observação do banco de dados no formato novo, foi necessário estabelecer uma nova etapa de pré-processamento (terceira na ordem). Percebeu-se grande número de pontos de dados que coincidiam exatamente com seus valores de LD, LQ e VMP. Quando estas coincidências ocorrem em diversos resultados de uma mesma amostra, considera-se que são erros de input no sistema, o que é plausível, devido à forma de alimentação manual e desinformação do usuário, que pode confundir estes valores associados ao resultado com o próprio resultado.

Após análise do banco de dados tratado com os critérios expostos anteriormente, a elaboração de novo critério se mostrou necessária. O novo critério é formado ao identificar os 7 valores que aparecem mais frequentemente em cada coluna do banco de dados, à exceção de zeros. O número 7 foi definido de maneira arbitrária pelo autor. Dessa maneira, são formados 27 conjuntos de valores mais frequentes, correspondentes a cada um dos parâmetros de qualidade em questão, e os 27 resultados de cada amostra são comparados com seus respectivos conjuntos de valores mais frequentes. Dessa maneira é realizada contagem, para cada linha, de quantos resultados pertencem ao conjunto de seus respectivos 7 valores mais frequentes.

Com os resultados de cada critério, estipulou-se ponto de corte para o descarte de amostras consideradas erros. Os 4 critérios citados geram valores entre 0 e 27, de acordo com o número de resultados de uma mesma amostra que atendem aquele critério específico. O ponto de corte foi estabelecido em 12, para os critérios LQ, VMP e Frequência. Amostras que tiveram 12 ou mais resultados que se enquadraram em um mesmo critério foram excluídas. O ponto de corte do critério LD foi 1 ocorrência.

Além da transformação da tabela original do MS para formato multivariado, o *Excel* foi usado para quantificar estes prováveis erros na base de dados. Depois de importada para o ambiente *Python*, as linhas que atendessem aos critérios de identificação de erros foram excluídas.

Adicionalmente foi aplicado o teste T quadrado de Hotelling, para exclusão de *outliers*, com limite = 0,01. Foram também removidas linhas em que houvesse parâmetros não alimentados, pois os métodos requerem preenchimento em todos os campos.

Na quinta e sexta etapas, a escolha dos métodos a serem aplicados levou em conta a revisão da literatura realizada pelo autor. Chegou-se à conclusão de que o método de Praus (2007), que consiste em PCA e a clusterização, seria o mais adequado para uma primeira análise do banco de dados, por serem uma forma simples de fazer uma análise exploratória dos dados.

Na sétima etapa do KDD foi aplicado o PCA sobre o banco de dados normalizado. A normalização da matriz de dados se deu com o uso da função standard scaler, da biblioteca PreProcessing do Scikit-Learn (SCIKIT LEARN), onde a matriz é escalonada para média zero e variância unitária, independente em cada variável. Após, foi realizada clusterização através de metodologia k-means, com o uso da função KMeans da biblioteca Clusters do Scikit-Learn, com o objetivo de identificar padrões. O número de componentes principais foi definido como sendo o que representasse 95% da variância dos dados. Para este estudo, 3 componentes principais representaram 94,5% da variância dos dados, e por isso o número de componentes principais usado foi 3. Estes 3 PCs serviram de *input* para aplicação do k-means. A avaliação do número ideal e qualidade dos *clusters* foi realizada com Análise de Silhueta.

Para a oitava etapa, a análise dos resultados, foi analisada a matriz de *loadings* do PCA e um bi-plot dos componentes principais exibindo os *clusters* formados pelo K-means. A análise pode ser validada pela observação dos dados originais.

5 Resultados e Discussão

Os resultados serão divididos em duas seções principais, sendo a primeira uma quantificação dos erros presentes no banco de dados, e a outra a aplicação do método de análise multivariada. A avaliação profunda da qualidade dos dados em busca de erros se mostrou necessária por estes erros afetarem os resultados da análise multivariada significativamente.

5.1 Identificação de erros no banco de dados

Devido ao fluxo de alimentação do SISAGUA, em que os dados de Controle são alimentados manualmente, ocorrem frequentemente problemas de inserção. Um dos problemas identificados é a ocorrência de dados incompletos. Outra é a inserção de valores no local do resultado da análise que não correspondem ao resultado, mas sim a algum valor associado àquele resultado, notadamente o LD, o LQ e o VMP.

A alimentação de valores associados ao resultado provavelmente ocorre devido à desinformação dos servidores responsáveis pela inserção dos dados. Possivelmente por se se depararem com um resultado “<LD” ou “<LQ”, que não é numérico, buscam algum valor numérico presente no laudo para digitar no sistema.

O Ministério da Saúde disponibiliza através do SISAGUA para as secretarias de estado e municípios relatório de implementação do VIGIAGUA estratificado por UF e Município. Um dos indicadores contemplados no relatório é o da presença de controle. O indicador é binário, bastando haver apenas uma análise cadastrada para a amostra inteira ser marcada como alimentada.

Os prováveis erros do banco de dados foram quantificados e segmentados por região do país, conforme descrito no Capítulo 3. As próximas subseções tratam dos resultados e formam um panorama da alimentação dos dados de agrotóxicos no SISAGUA.

5.1.1 Dados não alimentados no SISAGUA

A não alimentação de dados no sistema gera lacunas que não podem ser interpretadas pelo método de análise multivariada proposto neste trabalho, de maneira que no caso de haver pelo menos um resultado faltando, a linha toda acaba sendo descartada na análise. Esta inconsistência pode ter duas razões: análise não realizada pelo prestador de serviço, ou análise não alimentada no sistema de informação. No primeiro caso, a empresa de tratamento está sujeita a sanções legais por não estar executando o controle de qualidade adequado. O segundo caso é o mais preocupante, pois cria um cenário de desperdício de informação. A análise, que tem um alto custo associado, foi realizada e enviada para o setor saúde, porém a informação acabou se perdendo por não ser alimentada no sistema. A longo prazo, quando se tem o objetivo de construir uma série histórica, a consulta fica prejudicada.

A **Tabela 5-1** exibe o número de parâmetros alimentados no sistema por região do país. A região Norte é a que possui menos dados no sistema, enquanto que o sudeste é o que possui mais. As regiões Sul e Sudeste juntas contemplam aproximadamente 43 mil amostras do total de 50853. A região Sul é a que possui as análises mais incompletas no sistema, sendo que quase 5000 amostras, do total de 12000, possuem ao menos um parâmetro sem informação. No Brasil, 40% das amostras possuem ao menos um resultado sem preenchimento no sistema, o que significa que no presente trabalho elas precisaram ser desconsideradas e 20% das amostras do Brasil possuem 14 parâmetros preenchidos ou menos, ou seja, estão pela metade.

***Tabela 5-1**: Distribuição acumulada da alimentação de dados de agrotóxicos no sistema por região do Brasil de 2014 a 2018*

Parâmetros preenchidos	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
1	1219	5	125	42	576	471
até 2	1564	6	190	50	796	522
até 3	1850	7	197	57	1024	565
até 4	2055	8	255	66	1126	600
até 5	2266	8	366	75	1191	626
até 6	2448	10	377	86	1329	646
até 7	2709	10	382	97	1527	693
até 8	2957	10	387	107	1705	748
até 9	3340	11	423	123	1897	886
até 10	4224	11	666	160	2105	1282
até 11	5462	12	719	231	2364	2136
até 12	6493	16	919	375	2849	2334
até 13	7847	20	1116	594	3394	2723
até 14	10026	27	1597	895	3685	3822
até 15	11381	28	1765	1199	4074	4315
até 16	13553	30	2162	1832	4808	4721
até 17	13761	30	2169	1848	4951	4763
até 18	14070	34	2171	1858	5179	4828

Parâmetros preenchidos	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
até 19	14664	34	2173	1876	5600	4981
até 20	14821	35	2176	1881	5688	5041
até 21	15114	35	2180	1883	5882	5134
até 22	15684	36	2218	1886	6218	5326
até 23	17724	36	2505	1894	7372	5917
até 24	18363	36	2514	1897	7505	6411
até 25	19053	43	2525	1916	7820	6749
até 26	21099	59	2554	2019	9069	7398
até 27	50853	1019	2719	3850	30719	12546

5.1.2 Dados Iguais ao VMP

A alimentação de dados Iguais ao VMP, em especial, é a mais problemática de todas as hipóteses consideradas no trabalho. O valor máximo permitido é o marco determinante da não potabilidade da água, e inserí-lo no local de um resultado "<LD" ou "<LQ", causará grande distanciamento da realidade, pois um valor baixo e não preocupante é trocado por um valor alto, que se fosse real traria potencial risco à saúde.

A distribuição acumulada de alimentação de dados iguais ao VMP para as amostras consideradas no trabalho está representada na **Tabela 5-2**. Percebe-se que o Norte praticamente não possui resultados alimentados iguais ao VMP, enquanto que o Sul é a região com mais alimentações iguais ao VMP.

Apesar de ser pouco provável, não é impossível um resultado ser exatamente igual ao VMP. Não há como afirmar com certeza que uma ocorrência isolada desse tipo seja um erro, porém ao avaliar os vários resultados de uma mesma amostra, percebe-se que este tipo de erro de input de fato ocorre. Exemplo marcante disso são 8 amostras do Sudeste, e 2 do Sul, que tiveram todos os 27 parâmetros com resultados alimentados iguais ao VMP. Para o autor, qualquer valor maior do que três ou quatro resultados idênticos ao VMP já seriam sinal de alimentação precária, porém é difícil definir qual o ponto de corte para identificação de erros.

***Tabela 5-2**: Distribuição acumulada do número de resultados de agrotóxicos iguais ao VMP de 2014 a 2018*

Parâmetros iguais ao VMP	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 1	4973	3	1358	297	2295	1020
No mínimo 2	2288	0	552	156	1082	498
No mínimo 3	1301	0	509	100	219	473
No mínimo 4	618	0	6	30	160	422
No mínimo 5	581	0	0	23	142	416
No mínimo 6	543	0	0	13	121	409
No mínimo 7	528	0	0	13	114	401
No mínimo 8	482	0	0	13	72	397
No mínimo 9	469	0	0	13	70	386
No mínimo 10	399	0	0	13	68	318

Parâmetros iguais ao VMP	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 11	367	0	0	13	68	286
No mínimo 12	275	0	0	13	65	197
No mínimo 13	74	0	0	13	55	6
No mínimo 14	71	0	0	13	52	6
No mínimo 15	70	0	0	13	52	5
No mínimo 16	67	0	0	13	49	5
No mínimo 17	60	0	0	12	44	4
No mínimo 18	22	0	0	0	18	4
No mínimo 19	22	0	0	0	18	4
No mínimo 20	21	0	0	0	18	3
No mínimo 21	18	0	0	0	15	3
No mínimo 22	16	0	0	0	13	3
No mínimo 23	16	0	0	0	13	3
No mínimo 24	15	0	0	0	12	3
No mínimo 25	13	0	0	0	10	3
No mínimo 26	11	0	0	0	9	2
27	10	0	0	0	8	2
Total de Amostras	**50853**	**1019**	**2719**	**3850**	**30719**	**12546**

5.1.3 Dados Iguais ao LD

Outro critério de avaliação do banco de dados é a alimentação de valores idênticos ao LD. O LD também é informado pelo usuário que digita os valores no sistema. O problema deste tipo de ocorrência é que um resultado igual ao LD é impossível de constar em um laudo, pois não é uma quantificação. Os resultados somente se tornam numéricos a partir do LQ, dado que LD<LQ.

Apenas um resultado igual ao LD já seria suficiente para pôr em dúvida a confiabilidade do resultado. Para isso ocorrer, o usuário precisa digitar exatamente o mesmo valor em dois campos diferentes no sistema. 1257 amostras analisadas no Brasil possuem ao menos um resultado com esse problema, como pode ser visto na **Tabela 5-3**. A ocorrência desse tipo de problema de input é significativa especialmente no Sudeste, em que 45 amostras tiveram ao menos 25 parâmetros com resultado idêntico ao LD.

***Tabela 5-3**: Distribuição acumulada do número de resultados de agrotóxicos iguais ao LD de 2014 a 2018*

Parâmetros iguais ao LD	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 1	1257	0	63	77	578	539
No mínimo 2	971	0	60	54	459	398
No mínimo 3	730	0	60	49	405	216

Parâmetros iguais ao LD	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 4	656	0	60	39	389	168
No mínimo 5	622	0	60	35	375	152
No mínimo 6	582	0	60	32	347	143
No mínimo 7	551	0	58	30	342	121
No mínimo 8	533	0	58	28	331	116
No mínimo 9	501	0	58	25	304	114
No mínimo 10	472	0	58	23	278	113
No mínimo 11	443	0	58	19	256	110
No mínimo 12	329	0	23	4	200	102
No mínimo 13	289	0	15	4	172	98
No mínimo 14	250	0	0	4	158	88
No mínimo 15	199	0	0	4	123	72
No mínimo 16	148	0	0	4	110	34
No mínimo 17	124	0	0	4	98	22
No mínimo 18	105	0	0	3	91	11
No mínimo 19	98	0	0	3	87	8
No mínimo 20	80	0	0	3	70	7
No mínimo 21	74	0	0	3	65	6
No mínimo 22	71	0	0	3	65	3
No mínimo 23	58	0	0	3	52	3
No mínimo 24	54	0	0	3	48	3
No mínimo 25	48	0	0	3	45	0
No mínimo 26	1	0	0	1	0	0
27	0	0	0	0	0	0
Total de Amostras	**50853**	**1019**	**2719**	**3850**	**30719**	**12546**

5.1.4 Dados Iguais ao LQ

O resultado igual ao LQ possui uma característica parecida ao critério do LD. Para que ele ocorra, o usuário do SISAGUA precisa digitar duas vezes o mesmo valor em campos diferentes. A principal diferença é que esse critério de LQ possui número de ocorrências mais expressivo do que o de LD.

No Brasil, 5518 amostras tiveram resultado idêntico ao LQ, como visto na **Tabela 5-4**. A ocorrência de um resultado igual ao LQ sozinho não é suficiente para provocar suspeita de , porém nota-se frequência elevada de amostras com vários resultados iguais ao LQ. Mais de 2000 amostras possuem ao menos 13 resultados iguais ao LQ, o que pode ser tratado como indicativo de erro de *input*.

***Tabela 5-4**: Distribuição acumulada do número de resultados de agrotóxicos iguais ao LQ de 2014 a 2018*

Parâmetros iguais ao LQ	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 1	5518	18	102	542	3518	1338
No mínimo 2	5132	7	100	510	3333	1182
No mínimo 3	4879	6	97	501	3191	1084
No mínimo 4	4691	5	97	481	3073	1035
No mínimo 5	4389	3	97	466	2828	995
No mínimo 6	4034	3	95	455	2544	937
No mínimo 7	3750	3	92	359	2441	855
No mínimo 8	3504	3	87	329	2274	811
No mínimo 9	3221	3	87	324	2083	724
No mínimo 10	2946	3	86	310	2015	532
No mínimo 11	2701	3	86	299	1820	493
No mínimo 12	2394	1	86	279	1556	472
No mínimo 13	2104	1	85	256	1344	418
No mínimo 14	1709	1	83	199	1067	359
No mínimo 15	1150	1	83	171	585	310
No mínimo 16	950	1	83	156	505	205
No mínimo 17	848	1	80	118	483	166
No mínimo 18	659	1	71	34	455	98
No mínimo 19	599	1	45	21	438	94
No mínimo 20	520	1	16	14	400	89
No mínimo 21	501	1	9	14	390	87
No mínimo 22	471	1	7	8	375	80
No mínimo 23	426	1	0	7	351	67
No mínimo 24	341	0	0	5	313	23
No mínimo 25	113	0	0	5	97	11
No mínimo 26	3	0	0	0	2	1
27	0	0	0	0	0	0
Total de Amostras	**50853**	**1019**	**2719**	**3850**	**30719**	**12546**

5.1.5 Critério de frequência

O critério de frequência busca encontrar erros de alimentação que não coincidem exatamente com o LD, LQ e VMP. Após descarte de linhas com os critérios citados no Capítulo 4, percebeu-se ainda grande número de amostras com valores repetidos, mas que não coincidiam com LD, LQ e VMP.

A principal hipótese que explica este fenômeno, é o não-preenchimento dos valores de LQ e LD por parte do usuário. Neste cenário, o usuário insere o valor de LD ou LQ no lugar do resultado, e deixa em branco o campo de LD ou LQ que também tem alimentação

manual. Também se notou problema com a digitação de casas decimais, observando-se amostras que coincidiam com o valor de LD, LQ ou VMP, porém com uma casa decimal a mais ou a menos.

Para a execução deste critério, foi criada planilha que lista os 7 resultados mais frequentes, diferentes de zero, de cada um dos 27 parâmetros. Em seguida, para cada linha, foi realizada contagem de quantos parâmetros pertenciam ao conjunto dos 7 resultados mais frequentes para aquele parâmetro. Este critério não é suficiente para constatar um erro de inserção de dados no SISAGUA, porém serviu para o ajuste fino do tratamento de dados deste trabalho. O resultado desta contagem está exibido na **Tabela 5-5**.

São váras as amostras que atendem ao critério, e há 7543 amostras do Brasil que possuem ao menos 12 resultados que atendem ao critério de frequência. 5082 delas estão na região sudeste.

***Tabela 5-5**: Distribuição acumulada do número de resultados iguais a algum dos 7 valores mais frequentes para cada parâmetro*

Parâmetros iguais a algum dos 7 valores mais frequentes	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
No mínimo 1	12467	14	1605	693	8345	1810
No mínimo 2	11439	11	1513	648	7670	1597
No mínimo 3	11048	9	1503	623	7374	1539
No mínimo 4	10655	9	1452	598	7127	1469
No mínimo 5	10414	9	1434	592	6949	1430
No mínimo 6	10129	9	1330	583	6810	1397
No mínimo 7	9889	9	1328	571	6624	1357
No mínimo 8	9620	9	1325	552	6468	1266
No mínimo 9	9244	9	1304	534	6203	1194
No mínimo 10	8515	9	1106	519	5748	1133
No mínimo 11	7948	6	1043	503	5347	1049
No mínimo 12	7543	5	953	481	5082	1022
No mínimo 13	7246	3	870	478	4905	990
No mínimo 14	6498	3	420	470	4647	958
No mínimo 15	6080	1	258	434	4476	911
No mínimo 16	5708	1	111	398	4326	872
No mínimo 17	5316	1	111	371	4026	807
No mínimo 18	4983	1	108	238	3874	762
No mínimo 19	4581	1	106	216	3603	655
No mínimo 20	4326	1	104	199	3444	578
No mínimo 21	3834	1	101	166	3114	452
No mínimo 22	3299	1	72	118	2820	288
No mínimo 23	2853	1	4	112	2545	191
No mínimo 24	2193	1	3	110	1926	153
No mínimo 25	1899	0	2	109	1740	48
No mínimo 26	1602	0	2	101	1478	21
27	553	0	0	2	543	8

Parâmetros iguais a algum dos 7 valores mais frequentes	Total	NORTE	NORDESTE	CENTRO-OESTE	SUDESTE	SUL
Total de Amostras	**50853**	**1019**	**2719**	**3850**	**30719**	**12546**

5.2 Aplicação da análise multivariada aos dados do SISAGUA

5.2.1 Tratamento dos dados

Após a transformação do banco de dados de controle disponibilizado pelo Ministério da Saúde, a tabela obtida pelo autor contava com 77212 linhas, relativas aos 89 parâmetros previstos no padrão de potabilidade. Deliberadamente foi escolhido o subconjunto de dados foco do trabalho como sendo os 27 parâmetros de agrotóxicos previstos no padrão de potabilidade, o que fez com que o conjunto de dados fosse reduzido a 50853 linhas, devido a presença de amostras em que nenhum parâmetro de agrotóxico foi analisado.

Com o conjunto de 50853 linhas, o número de colunas de dados foi reduzido para 27 - os agrotóxicos. Começou então a análise dos critérios estabelecidos na seção 5.1. Ao aplicar o critério de LD, o conjunto foi reduzido a 49596 amostras. Após o critério LQ, restaram 47460. Com o critério de VMP, 47347. Por fim, com o critério de frequência, restaram 42747 amostras. A ordem da aplicação dos critérios não altera o número restante de amostras. O ponto de corte dos critérios LQ, VMP e Frequência foi igual a 12 ocorrências, enquanto que o do critério LD foi 1, conforme descrito no Capítulo 4.

O próximo critério de remoção de linhas aplicado foi o de remoção de linhas que não houvesse preenchimento total dos parâmetros. Dessa maneira, restaram 24568 amostras no *dataset*. Por fim, foi realizada remoção de duplicatas. Para tal, foram selecionadas apenas amostras em que houvessem resultados diferentes de zero para pelo menos um parâmetro. Assim, foram excluídas linhas em que todos os 27 resultados da amostra coincidissem exatamente com os 27 de outra, chegando a um *dataset* de 23446 amostras. Nesse momento também foi feita contagem das amostras com todos os parâmetros iguais a zero: 22689. Apenas 757 amostras do *dataset* final contém resultados, sendo todas as outras preenchidas por zeros. Os zeros não foram excluídos do *dataset.*

Nesse momento foi realizado o teste T Hotteling, para a remoção de *outliers*. Como o conjunto de dados final está majoritariamente composto por zeros, o teste foi aplicado somente às amostras em que houvesse dados diferentes de zero. O teste removeu 58 pontos, restando 699 amostras, de maneira que o *dataset* utilizado para a aplicação do método contém 23388 amostras. A **Tabela 5-6** resume o tratamento dos dados anterior à aplicação do método. Como pode haver amostras que atendam a mais de um critério, e por isso apenas o primeiro critério aplicado corresponde diretamente ao valor exibido na seção 5.1.

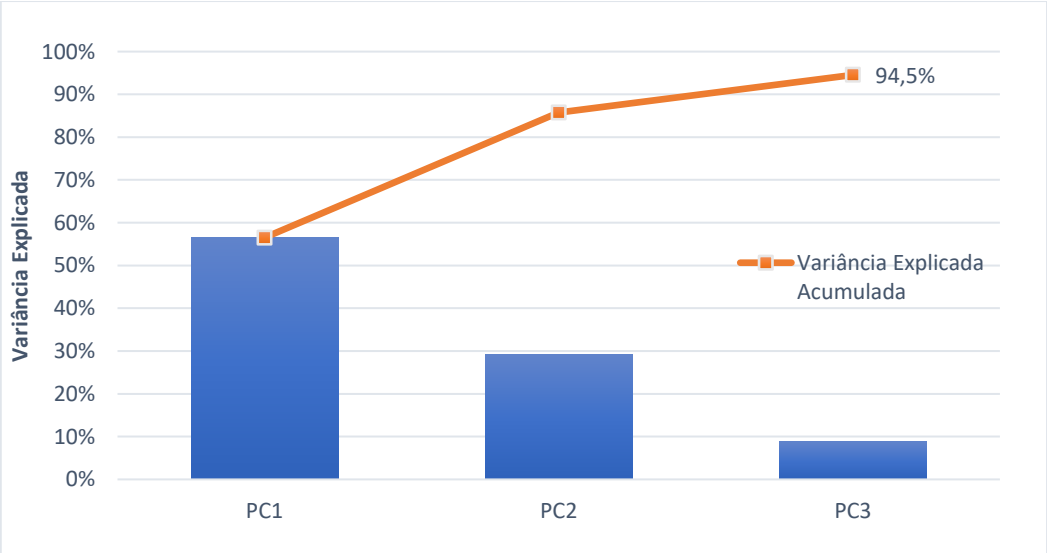
***Tabela 5-6**: Progressão do tratamento de dados*

Etapa	Linhas removidas	Linhas restantes
Conjunto Original	-	77212
Subconjunto Agrotóxicos	26359	50853
Critério LD	1257	49596
Critério LQ	2134	47460
Critério VMP	113	47347
Critério Frequência	4600	42747
Valores não preenchidos	18179	24568
Remoção de duplicatas	1122	23446
Teste T^2 Hotelling	58	23388

5.2.2 Aplicação do PCA

Com o conjunto de dados tratado, o mesmo foi normalizado e aplicado o PCA. O resultado pode ser observado na **Figura 5-1**. O método identificou que três componentes principais somados são responsáveis por explicar 94,5% da variância dos dados. O primeiro componente é responsável por aproximadamente 56% da variância explicada e o segundo por quase 30%.

***Figura 5-1**: Variância explicada para cada componente principal*



Os *Loadings* de cada componente estão representados na tabela **Tabela 5-7**. Observa-se que o primeiro componente principal é influenciado com um maior peso pelos parâmetros glifosato + AMPA e mancozebe. Os parâmetros carbendazim + benomil e tebuconazol também apresentam significativa contribuição ao PC1. O segundo componente principal possui influência com ordens de grandeza elevadas dos parâmetros atrazina, endrin, lindano e parationa metílica. Já o terceiro é identificado como sendo somente o clordano.

***Tabela 5-7**: Loadings de cada componente principal*

Parâmetro	PC1	PC2	PC3
2,4 D + 2,4,5 T	0,014971	0,010053	0,002613
Alaclor	0,00334	0,001925	0,004327
Aldicarbe + Aldicarbesulfona + Aldicarbesulfóxido	0,025872	-0,01165	0,013412
Aldrin + Dieldrin	0,000244	-0,00032	-0,00031
Atrazina	0,001926	8,50E+09	0,00276
Carbendazim + benomil	0,118604	-0,0805	0,036919
Carbofurano	0,012103	-0,00461	-0,00026
Clordano	0,000277	0,000312	-3,78E+11
Clorpirifós + clorpirifós-oxon	0,017544	-0,00922	0,059057
DDT + DDD + DDE	0,000396	-0,00011	0,000564
Diuron	0,064759	-0,01699	0,133907
Endossulfan (a, ß e sais)	0,001993	0,001724	0,002436
Endrin	0,000329	5,42E+11	0,000234
Glifosato + AMPA	0,693653	0,716921	-0,05434
Lindano (gama HCH)	0,000744	7,09E+10	0,000239
Mancozebe	0,681571	-0,68237	-0,22741
Metamidofós	0,01189	-0,00409	0,028962
Metolacloro	0,013707	-0,00346	0,044028
Molinato	0,002214	0,000123	0,006019
Parationa Metílica	0,005743	-8,76E+10	0,015118
Pendimentalina	0,004626	0,000492	0,013735
Permetrina	0,004116	0,001607	0,010643
Profenofós	0,051557	-0,01729	0,203356
Simazina	0,001033	0,000246	0,000761
Tebuconazol	0,177745	-0,11374	0,936639
Terbufós	0,002406	-0,00125	0,002478
Trifluralina	0,003632	-0,0006	0,018253

É possível realizar busca na literatura para avaliar as características em comum dos parâmetros originais que tem peso maior em cada PC. As características consideradas são objetivo da aplicação, característica físico-química e volume de venda. Serão analisados somente os *loadings* nesta subseção.

O PC1 é identificado como o grupo da soja, pois é influenciado por agrotóxicos que possuem amplo espectro de aplicação, mas possuem em comum a autorização de aplicação à soja (ANVISA). Todos são fungicidas, à exceção do glifosato e seu metabólito AMPA, que é um herbicida. A soja é a cultura de maior produção no brasil, e o glifosato, que pode ser aplicado na soja, é o agrotóxico mais vendido. O fato de o componente mais importante estar relacionado com a soja e o glifosato é um resultado compreensível.

Os parâmetros que influenciam o PC2 são inseticidas organoclorados (endrin e lindano), inseticida organofosforado (parationa metílica) e herbicida (atrazina). Assim como no PC1, os parâmetros foram aglomerados por objetivo de uso na sua maioria. Nota-se que a atrazina possui venda bastante expressiva, sendo a quarta maior venda da lista de agrotóxicos do padrão de potabilidade. Para este PC, os agrotóxicos destacados possuem ordem de grandeza em seus *loadings* da ordem de 9, 10 e 11, muito acima de qualquer outro.

O PC3 é representado pelo clordano, que possui ordem de grandeza 11. Como não há outro parâmetro com loading tão elevado quanto o clordano, parece não haver correlação entre a ocorrência de clordano com a ocorrência de outros agrotóxicos. O clordano é um inseticida organoclorado.

5.2.3 Clusterização

5.2.3.1 Análise de Silhueta

A análise de Silhueta foi realizada para K = 3 e K = 4 *clusters*, e apresentou melhor resultado para K = 4 *clusters*. Ambas as análises retornaram um *score* aproximadamente igual, em torno de 0,7 numa escala de -1 a 1. Os gráficos de silhueta estão disponíveis nas **Figura 5-2** e **Figura 5-3**.

No caso de haver 3 *clusters*, houve diminuição na largura das regiões hachuradas no gráfico, significando maior distância entre os pontos e os centroides dos *clusters*. O *cluster* **a** sempre é o que possui a maior altura de todos, significando que ele aglomera mais amostras do que os outros. A diferença entre K = 3 e K = 4 está na divisão do *cluster* **b**, em **b** e **d**. O *cluster* **c** permanece inalterado.

Figura 5-2: *Análise de Silhueta com K = 3 clusters*

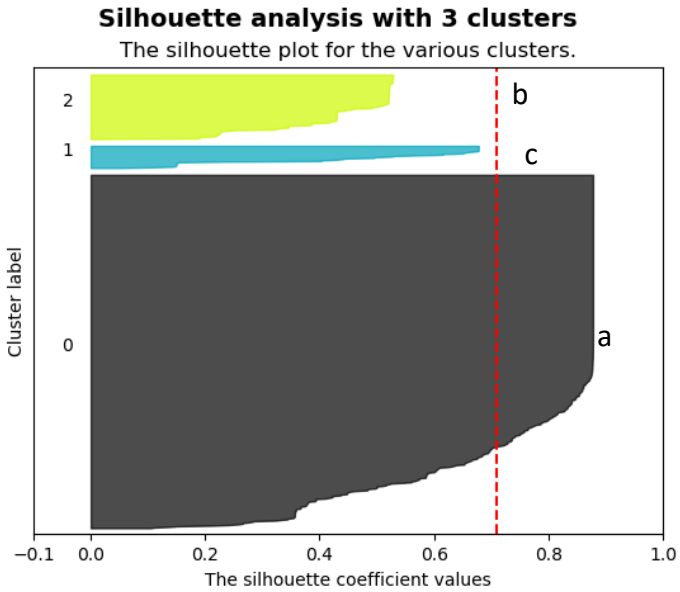
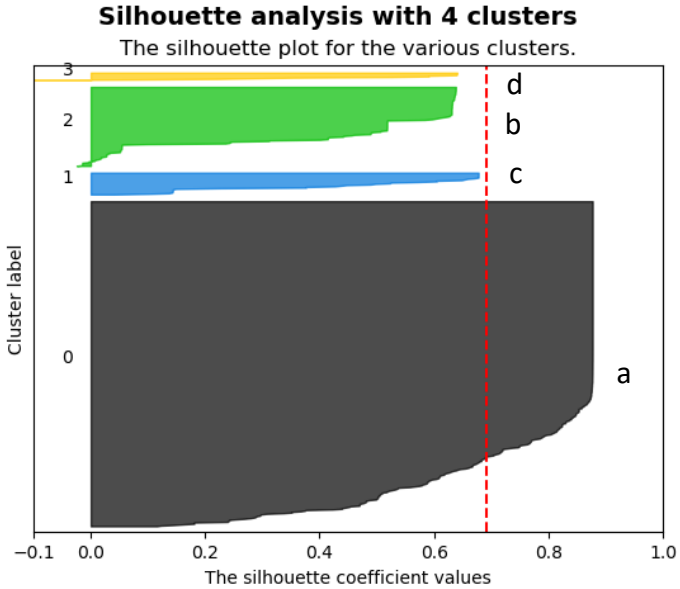


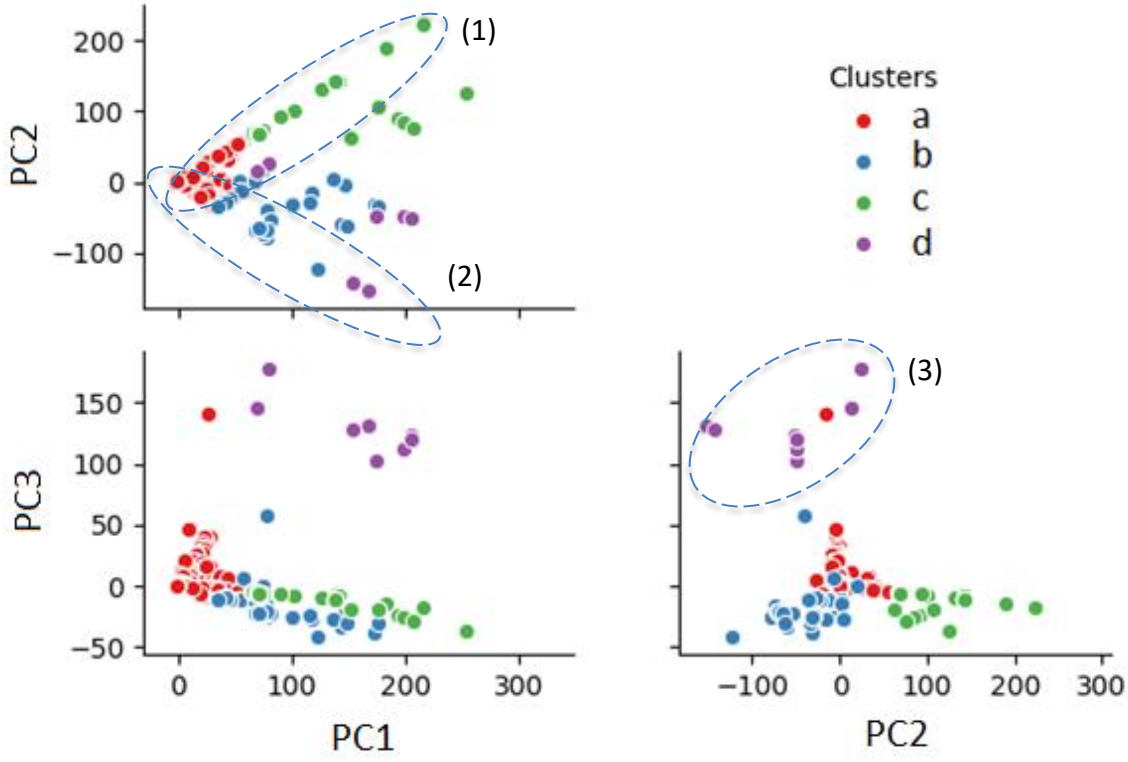
Figura 5-3: *Análise de Silhueta com K = 4 clusters*



5.2.3.2 K-means

Com o objetivo de aprofundar a análise, a partir do resultado do PCA foi aplicado o K-means e quatro *clusters* foram revelados. A **Figura 5-4** é um bi-plot dos três primeiros componentes principais, identificando os quatro *clusters* com cores diferentes. O *cluster* **a** corresponde a 23207 amostras, e engloba todas as amostras que contém apenas zeros. Do total, há 518 amostras com resultado diferente de apenas zeros no *cluster* **a**. No *cluster* **b**, foram aglomeradas 133 amostras, no *cluster* **c**, 36, e no *cluster* **d**, 12 amostras.

***Figura 5-4**: Bi-plot dos Componentes principais com representação dos clusters identificados pelo K-means*



Os *clusters* não estão bem demarcados nos gráficos, apresentando regiões comuns em na dispersão de seus pontos. Apesar disso, a tendência que marca a aglomeração dos pontos nas regiões dos gráficos correspondentes aos *clusters* pode ser avaliada.

O *cluster* **a** está centralizado em torno de valores de PC1, PC2 e PC3 iguais a zero. Ele pode ser interpretado como sendo representante das amostras que possuem quantificações de poucos agrotóxicos e em baixas concentrações. Isso é devido ao fato de não haver nada que as desloque fortemente do zero nas direções de PC1, PC2 e PC3.

Nota-se que PC2 é suficiente para distinguir as amostras dos *clusters* **b** e **c**, porém há região comum entre esses dois *clusters* e o *cluster* **a**. O *cluster* **b** é marcado pelo valor negativo de PC2, enquanto o *cluster* **c** possui valor de PC2 sempre positivo.

Ao analisar os valores originais das amostras, nota-se que não são os parâmetros com ordem de grandeza próxima de 10 que fazem com que os PCs 2 e 3 assumam valores mais elevados, mas sim aqueles que possuem *loading* entre 0,5 e 1. Isso mostra que a análise dos *loadings* sozinha não foi suficiente para identificar o comportamento dos dados.

O *cluster* **b** é marcado por amostras com elevada concentração de mancozebe. O mancozebe possui loading significativo positivamente no PC1 e negativamente no PC2, O que faz com que os pontos com alta concentração de mancozebe se desloquem em linha na direção (2) marcada no gráfico.

O *cluster* **c** é formado por amostras com concentração elevada de glifosato, que possui contribuição positiva forte para o PC1 e o PC2. A linha que descreve o crescimento da concentração de glifosato está marcada como (1) no gráfico.

O *cluster* **d** é revelado quando se analisa o PC3, e percebe-se uma região de deslocamento marcada como (3). Esta região está relacionada com a concentração elevada de tebuconazol, que possui loading elevado no PC3. No plot PC2 *versus* PC3, nota-se que amostras com PC2 equivalente à faixa dos *clusters* **a** e **b** são deslocadas pelo tebuconazol, com elevação do PC3, porém as amostras do *cluster* **c**, do glifosato, não são. Este comportamento é resultado de uma tendência de não haver tebuconazol nos pontos em que é identificado o glifosato.

Uma hipótese que justifica a tendência observada é a necessidade de aplicação de tebuconazol apenas nos locais em que o Glifosato não é aplicado. O tebuconazol e o glifosato tem como efeito comum o controle de fungo chamado ferrugem de soja (SOARES et al., 2008). O Glifosato tem venda de 173 mil toneladas por ano, enquanto o tebuconazol tem venda de 2 mil toneladas (IBAMA, 2018), o que mostra a diferença de escala de aplicação dos dois.

O 2,4 D, apesar de ser o segundo agrotóxico mais vendido no Brasil, não possui grande influência nos resultados observados neste trabalho. De fato, ao observar os dados originais, são poucas as detecções de 2,4 D frente aos outros parâmetros citados nesta subseção.

6 Conclusões e Trabalhos Futuros

Em um país que é marcado pelo uso de agrotóxicos em larga escala, é muito importante realizar o acompanhamento dos impactos que esta aplicação pode trazer à saúde das pessoas. Este estudo buscou analisar e extrair conhecimento útil a partir dos dados correspondentes à contaminação de agrotóxicos em amostras de água para consumo humano registrados na base de dados do SISAGUA.

A revisão bibliográfica teve sucesso em encontrar publicações de outros lugares do mundo, porém percebe-se que não é um campo muito explorado. O presente trabalho alcançou o seu objetivo principal ao aplicar, pela primeira vez, técnicas de análise multivariada a resultados de análises de água para consumo humano no Brasil.

A partir da aplicação do PCA, fica claro que a matriz de l*oadings* pode não ser suficiente para identificar padrões do universo amostral. Foi a partir da busca de *clusters*, realizada com k-means, e plot dos PCs que os principais achados do trabalho foram revelados. Mesmo assim, os *loadings* encontrados caracterizam o PC1 como o grupo da soja, por possuir principalmente fungicidas aplicados nessa cultura, enquanto que o PC2 foi chamado de grupo dos inseticidas organoclorados e organofosforados, mais a atrazina.

Foi possível encontrar significados físicos para os resultados de uma metodologia não supervisionada. Descobriu-se que os agrotóxicos que representam a maior variância dos dados são exatamente os mais vendidos no Brasil. O glifosato e o mancozebe possuíram forte influência nos PC1 e PC2, deslocando os pontos no gráfico e formando *clusters* por si sós. Há ainda uma região em comum em que a separação dos clusters não é muito clara, que representa pontos com glifosato e mancozebe ao mesmo tempo, ou seja, a ocorrência de um não está necessariamente correlacionada com a do outro.

Outra correlação interessante encontrada foi a dependência do glifosato e do tebuconazol, revelada através do k-means. Descobriu-se tendência de não ser encontrado tebuconazol nas amostras em que o glifosato está presente, sendo justificado pelo efeito fungicida que ambos os agrotóxicos possuem.

Além das correlações entre os dados, foi realizada análise da qualidade dos dados do SISAGUA. Foi encontrada significativa parcela de valores não preenchidos no sistema, o que inviabiliza o uso da amostra inteira para o tipo de análise realizada neste trabalho. Duas causas para a ocorrência de valores não preenchidos no sistema foram citadas, sendo a primeira a não realização de alguma análise, e outra a não alimentação do resultado por parte do usuário do sistema.

A não alimentação de resultados no sistema gera desperdício e perda de informação. É muito importante que todos os resultados sejam incluídos no sistema, pois somente assim é possível realizar análise histórica do comportamento dos dados. De nada adianta haver cobrança pela realização das análises, se os resultados estão sendo perdidos.

Foi graças à nova disposição dos dados, com os diferentes parâmetros em colunas, que a análise dos erros foi possível. Resultados iguais ao VMP, LD e LQ não são valores expúrios que poderiam ser removidos com uma simples busca por *outliers*. Foi necessário analisar vários resultados de uma vez, e tratar a amostra como um todo para perceber este tipo de erro.

A quantidade de erros de alimentação no SISAGUA não é desprezível, e através de critérios estabelecidos pelo autor, foram identificados prováveis erros de *input* no sistema, que tiveram que ser removidos para a elaboração do estudo.

Tanto os erros de input quanto a não alimentação de dados são causados pelo processo de carga dos dados no sistema. Já há *webservice* que possibilita a carga automática de dados no SISAGUA, sem passar por digitação manual. O que ocorre é que a aplicação não está disseminada por todo o país. No Rio Grande do Sul, por exemplo, a Companhia Riograndense de Saneamento (CORSAN) ainda envia seus resultados em relatórios PDF, que são digitados por funcionários dos municípios.

Este estudo pode ser o ponto de partida para um novo olhar, multivariado, sobre os dados do SISAGUA. O potencial de aplicação é muito grande, podendo a mesma análise ser realizada para outros parâmetros, que não agrotóxicos, e novos comportamentos serem identificados. Há ainda dados suficientes para diminuir a abrangência geográfica da análise e fazer investigações a nível municipal e estadual. Podem ser ainda aplicados métodos mais complexos, baseados em redes neurais, por exemplo.

Do ponto de vista de metodologia científica, as primeiras oito etapas do KDD são executadas pelo autor. Elas servem de subsídio para a tomada de ação por parte do Ministério da Saúde e outros braços do Estado. Os resultados do trabalho podem ser o ponto de partida para a nona etapa, a implementação de melhorias no programa VIGIAGUA a nível nacional.

Como recomendação para o setor saúde do Brasil, estabelecer um prazo ou meta de implementação do *webservice,* para companhias de saneamento, eliminando assim a etapa manual da inserção de dados no sistema. Somente tendo confiança nos resultados é que podemos fazer análises relevantes.

7 Referências

ABDI, Hervé; WILLIAMS, Lynne J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, [s. l.], v. 2, n. 4, p. 433–459, 2010. Disponível em: <http://doi.wiley.com/10.1002/wics.101>

ANVISA. **Monografias de Agrotóxicos.** Disponível em: <http://portal.anvisa.gov.br/agrotoxicos>

BRASIL. **Anexo XX da Portaria de Consolidação nº 5**. Brasília: Ministério da Saúde, 2017.

BRASIL. Secretaria de Vigilância em Saúde. **Diretriz Nacional do Plano de Amostragem da Vigilância da Qualidade da Água para Consumo Humano**. Brasília: Ministério da Saúde, 2016.

BRASIL, **Lei Federal nº 7.802**. Brasil: 1989

CHOWDHURY, Shakhawat; CHAMPAGNE, Pascale; JAMES MCLELLAN, P. Factors influencing formation of trihalomethanes in drinking water: Results from multivariate statistical investigation of the Ontario drinking water surveillance program database. **Water Quality Research Journal of Canada**, [s. l.], v. 43, n. 2–3, p. 189–199, 2008.

COUSINOU, M. C. et al. Multivariate analysis of cosmic void characteristics. **Astronomy and Computing**, [s. l.], v. 27, p. 53–62, 2019. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2213133718300684>

CONAMA. **Resolução nº 357**. Brasil: 2005.

CONAMA. **Resolução nº 396**. Brasil: 2008.

DARLINGTON, J.; GHANEM, M.; TO, H. W. **Structured parallel programming**. [s.l: s.n.].

GELADI, Paul. Chemometrics in spectroscopy. Part 1. Classical chemometrics. **Spectrochimica Acta Part B: Atomic Spectroscopy**, [s. l.], v. 58, n. 5, p. 767–782, 2003. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0584854703000375>

HAND, David J. Data *Cluster*ing: Theory, Algorithms, and Applications by

Guojun Gan, Chaoqun Ma, Jianhong Wu. **International Statistical Review**, [s. l.], v. 76, n. 1, p. 141–141, 2008. Disponível em: <http://cds.cern.ch/record/1560002/files/9780898716238_TOC.pdf>

IBAMA. **Consolidação de dados fornecidos pelas empresas registrantes de produtos técnicos, agrotóxicos e afins, conforme art. 41 do Decreto n° 4.074/2002.** Disponível em: <https://www.ibama.gov.br/agrotoxicos/relatorios-de-comercializacao-de-agrotoxicos>

JIAWEI HAN, MICHELINE KAMBER, Jain Pei. **Data Mining – Concepts & Techniques**. [s.l: s.n.].

LE GRESLEY, Adam; PERON, Jean-Marie R. A semi-automatic approach to the characterisation of dark chocolate by Nuclear Magnetic Resonance and multivariate analysis. **Food Chemistry**, [s. l.], v. 275, p. 385–389, 2019. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0308814618316613>

LOPES, Carla Vanessa Alves; ALBUQUERQUE, Guilherme Souza Cavalcanti De. Agrotóxicos e seus impactos na saúde humana e ambiental: uma revisão sistemática. **Saúde em Debate**, [s. l.], v. 42, n. 117, p. 518–534, 2018.

MANNING-DAHAN, Tyler. PCA and Autoencoders. [s. l.], 2017. Disponível em: <http://tylermd.com/pdf/pca_ae.pdf>

MICROSOFT. **Especificações e Limites do Excel 2016**. Disponível em: <https://support.office.com/pt-br/article/especifica%C3%A7%C3%B5es-e-limites-do-microsoft-excel-1672b34d-7043-467e-8e27-269d656771c3 >

OLIVEIRA, Aristeu De et al. Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (Sisagua): características, evolução e aplicabilidade*. **Epidemiologia e Serviços de Saúde**, [s. l.], v. 28, n. 1, p. 1–13, 2019.

OLKIN, I.; SAMPSON, A. R. Multivariate Analysis: Overview. In: **International Encyclopedia of the Social & Behavioral Sciences**. [s.l.] : Elsevier, 2001. v. 54p. 10240–10247.

PEDREGOSA *et al.* **Scikit-learn: Machine Learning in Python**. JMLR 12. pp. 2825-2830, 2011.

PRAUS, P. Urban water quality evaluation using multivariate analysis. **Acta Montanistica Slovaca**, [s. l.], v. 12, n. 2, p. 150–158, 2007.

RIO GRANDE DO SUL. **Portaria n º 320**. Porto Alegre: Secretaria Estadual da Saúde, 2014

ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of *cluster* analysis. **Journal of Computational and Applied Mathematics**, [s. l.], v. 20, n. 4, p. 53–65, 1987. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0377042787901257>

SCIKIT LEARN. **PreProcessing Library - Standard Scaler.** Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

SOARES, Rafael Moreira et al. Utilização de glifosato para o controle de ferrugem da soja. **Pesquisa Agropecuaria Brasileira**, [s. l.], v. 43, n. 4, p. 473–477, 2008.

TAVARES, Edson Marcos Leal Soares; SOARES, Patrícia Silva. **GRÁFICO DE CONTROLE MULTIVARIADO T2 DE HOTELLING - INSTRUMENTO DE ANÁLISE DA QUALIDADE NUMA INDÚSTRIA DE ALUMÍNIO**. Simpósio de Pesquisa Operacional da Marinha. Rio de Janeiro: Brasil, 2006.

UNIVERSITY OF HERTFORDSHIRE. **Pesticide Properties DataBase (PPDB).** 2016. Disponível em: <http://sitem.herts.ac.uk/aeru/ppdb/en/index.htm>

WHO. **Guidelines for drinking-water quality**. Genebra: World Health Organization, 2011.