UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA – ITEC
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AGRUPAMENTO DE FORNOS DE REDUÇÃO DE ALUMÍNIO UTILIZANDO OS ALGORITMOS *AFFINITY PROPAGATION,* MAPA AUTO–ORGANIZÁVEL DE KOHONEN (SOM), *FUZZY C–MEANS* E *K–MEANS*

FLÁVIA AYANA NASCIMENTO DE LIMA

DM 37/2017

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém – Pará – Brasil
2017

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA – ITEC

PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FLÁVIA AYANA NASCIMENTO DE LIMA

AGRUPAMENTO DE FORNOS DE REDUÇÃO DE ALUMÍNIO UTILIZANDO OS ALGORITMOS *AFFINITY PROPAGATION*, MAPA AUTO–ORGANIZÁVEL DE KOHONEN (SOM), *FUZZY C–MEANS* E *K–MEANS*

Dissertação de Mestrado submetida à banca examinadora do Programa de Pós–Graduação em Engenharia Elétrica da UFPA para a obtenção do grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém – Pará – Brasil
2017

Lima, Flávia Ayana Nascimento de.

Agrupamento de fornos de redução de alumínio utilizando os algoritmos *Affinity Propagation*, Mapa auto–organizável de Kohonen (som), *Fuzzy C–Means* e *K–Means* / Flávia Ayana Nascimento de Lima; orientador, Roberto Célio Limão de Oliveira — 2017.

Dissertação (Mestrado) – Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós–Graduação em Engenharia Elétrica (PPGEE) Belém, 2017.

1. Agrupamento. 2. Redução do Alumínio. 3. Mineração de Dados. I. Lima, Flávia Ayana Nascimento de. II. Oliveira, Roberto Célio Limão de. III. Universidade Federal do Pará – UFPA. IV. Agrupamento de fornos de redução de alumínio utilizando os algoritmos *Affinity Propagation*, Mapa auto–organizável de Kohonen (som), *Fuzzy C–Means* e *K–Means*.

CDD: 006.312

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA – ITEC
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FLÁVIA AYANA NASCIMENTO DE LIMA

Título: Agrupamento de fornos de redução de alumínio utilizando os algoritmos *Affinity Propagation*, Mapa auto–organizável de Kohonen (SOM), *Fuzzy C–Means* e *K–Means*.

DEFESA DE MESTRADO

Esta Dissertação foi julgada e aprovada para obtenção do título de **Mestre em Engenharia Elétrica na área de concentração em Computação Aplicada do Programa de Pós–Graduação *Strictu Sensu* em Engenharia Elétrica da Universidade Federal do Pará – ITEC – UFPA.**

Belém–PA, _____ / _____ / 2017

BANCA EXAMINADORA

_____

Prof. Dr. Roberto Célio Limão de Oliveira – Orientador (PPGEE – UFPA)

_____

Prof. Dr. Diego Lisboa Cardoso – Co-orientador (PPGEE–UFPA)

_____

Prof. Dr. Jefferson Magalhães de Morais – Examinador Externo (PPGCC–UFPA)

_____

Prof. Dr. Adrião Duarte Dória Neto – Examinador Externo (UFRN)

Dedico esta Dissertação de Mestrado a minha querida e amada mãe, que com sua infinita sabedoria me ensinou que jamais devemos desistir dos nossos sonhos, mesmo com todas as dificuldades, adversidades e obstáculos que a vida nos prepara.

Mãe, olhe sempre por mim aí do céu e obrigada por tudo. Te amo!

AGRADECIMENTOS

Agradeço primeiramente a Deus, que é a fortaleza e o alicerce para todos os sonhos.

Agradeço a minha amada mãe, que do céu continua me amparando e me guiando para o caminho do bem, fazendo com que eu encontre forças divinas para seguir em frente, mesmo com toda dor que sua falta me faz. Mãe, te amo mais que a mim mesma.

Agradeço ao professor Roberto Célio Limão de Oliveira, querido orientador, que me guiou durante toda a jornada desta Dissertação. Professor, sem o seu auxílio esse sonho jamais se tornaria realidade. Obrigada de coração.

Agradeço a Carnot Luiz Braun Guimarães, que me apresentou ao mestrado da UFPa e me incentivou a seguir em frente.

Agradeço ao meu querido amigo Alan Marcel, parceiro de disciplinas e dupla incansável para a realização desta Dissertação. Amigo, sem palavras para agradecer e dizer o quanto você foi essencial para que esse sonho se tornasse realidade.

Agradeço também ao meu amigo Fábio Mendes, que mesmo na Guatemala me ajudou muito nesta jornada. Amigo, muito obrigada!

Agradeço ao meu amigo Alexandre Oliveira, que esteve presente em minha trajetória, nos momentos bons e ruins da minha vida. Amigo, obrigada por tudo e principalmente, obrigada por me entender e estar presente em todos os momentos mais importantes.

Agradeço a minha querida prima Liciana Alice, mais que prima, minha irmã, que mesmo longe me apoia em todos meus caminhos e sempre dá seu toque especial em todos os momentos mais importantes da minha vida. Prima, te amo muito e obrigada por tudo.

Agradeço aos meus amigos e familiares que torcem por mim: a minha vó Imelda, meu querido pai Renato, minha vó Dayse, minha sogra Laura, meu sogro Vincenzo, meus irmãos, minhas tias, principalmente àqueles que conhecem minha trajetória no mestrado e sabem tudo que eu passei para chegar até aqui. Obrigada a todos pelo apoio e carinho.

E por fim, agradeço a Deus por ter me enviado um ser humano indescritível chamado José Vincenzo Procopio Filho, meu amor, minha vida, a pessoa que hoje é o motivo que me faz seguir em frente. Amor, obrigada por sempre me apoiar, enxugar minhas lágrimas e me dizer: não desista, estou contigo. Você é mais que maravilhoso, você é meu presente de Deus. Obrigada por existir. Te amo!

RESUMO

O constante avanço da tecnologia requer medidas que beneficiem as indústrias em busca do lucro e da competitividade. Em relação à indústria de minerais, o processo de fundição de alumínio geralmente possui grande número de células, também chamado de forno ou cuba de redução, produzindo alumínio em um procedimento contínuo e complexo. Um monitoramento analítico é essencial para aumentar a vantagem competitiva dessa indústria, por exemplo, durante a operação, algumas células compartilham comportamentos semelhantes às outras, formando grupos ou *clusters* de células. Esses *clusters* dependem de padrões de dados geralmente implícitos ou invisíveis para a operação, mas que podem ser encontrados por meio da análise de dados. Neste trabalho, são apresentadas quatro técnicas de agrupamento, o *Affinity Propagation*, o mapa auto–organizável de Kohonen (SOM), o algoritmo difuso *Fuzzy C–Means* (FCM) e o *K–Means*. Essas técnicas são utilizadas para encontrar e agrupar as células que apresentam comportamentos semelhantes, de acordo com sete variáveis tais como as que consiste no processo de redução do alumínio. Este trabalho visa trazer o benefício do agrupamento, principalmente pela simplificação da análise da linha de produção do alumínio, uma vez que um grande número de células pode se resumir em um único grupo, o que pode fornecer informações mais compactas para o controle e a modelagem dos dados. Este benefício de identificar os dados que possuem características semelhantes e agrupá–los faz com que a análise dos grupos se torne mais simples para quem irá manusear esses dados futuramente. Nesse trabalho de dissertação também será feito a identificação da quantidade ideal de grupo em cada técnica utilizada.

Palavras–Chave: Agrupamento, Redução do Alumínio, Mineração de Dados.

ABSTRACT

The continuous development of technology accounts for measures that provide industries benefits to grant them profitability and competitive advantage. In the mineralogy field, aluminum smelting usually requires substantial number of cells, also known as reduction pots, to produce aluminum in a continuous and complex process. Analytical monitoring is essential for those industries' competitive advantage, given that during operation some cells show behavior similar to others, thereby forming clusters of cells. These clusters depend on data patterns usually implicit or invisible for the operation, but can be found by data analysis techniques. In this work four clustering techniques are presented to that end: the Affinity Propagation; the Kohonen Self Organizing Map; the Fuzzy C–Means; and the K–Means Algorithm. These techniques are used to find and group cells that share similar behavior, by analysing seven variables which are closely related to the aluminum reduction process. This work aims at addressing the benefits of clustering, especially by simplifying the aluminum potline analysis, once a large group of cells might be summarized in one sole group, what can provide more compact yet rich information for data driven modeling and control. Moreover, the identification of similar data patterns in clusters makes the task of those who is going to be in charge of analyzing these dats. This work also identifies the ideal cluster size for each technique applied.

Keywords: Clustering, Aluminum Reduction, Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

GLOSSÁRIO

Affinity Propagation	Algoritmo baseado no conceito de "passagem de mensagens" entre pontos de dados. Ao contrário dos algoritmos de agrupamento como *K–Means* ou K–Medoids, a propagação por afinidade não exige que o número de *clusters* seja determinado ou estimado antes da execução o algoritmo.
Ânodo	Elétrodo no qual há oxidação (perda de elétrons).
Big Data	Grande quantidade de dados, que necessitam de ferramentas especialmente preparadas para lidar com grandes volumes, para que toda e qualquer informação nestes meios possa ser analisada e aproveitada.
Cátodo	Elétrodo de uma célula eletroquímica onde se dá a redução de uma espécie química (ganho de elétrons).
Criolita	Mineral do grupo dos halóides, composto basicamente por: sódio, alumínio e flúor, com a fórmula: Na_3AlF. É maciço, de brilho vítreo e graxo, e algumas vezes nacarado.
Cluster	Grupo de coisas, dados ou de atividades semelhantes que se desenvolvem conjuntamente.
Dendrograma	(dendro = árvore) é um tipo específico de diagrama ou representação icónica que organiza determinados fatores e variáveis.
Distância Euclidiana	Distância entre dois pontos, que pode ser provada pela aplicação repetida do teorema de Pitágoras. Aplicando essa fórmula como distância, o espaço euclidiano torna–se um espaço métrico.
Filtragem de dados	Processo de separação dos dados considerados importantes em relação aos dados considerados com informações incorretas ou contendo ruídos.
Forno Eletrolítico	Local onde ocorre o processo da passagem de corrente elétrica por uma solução química.

***Fuzziness* (Fuzzificação)**	Grau de imprecisão, confusão, difusão.
Fuzzy C–Means	Método de agrupamento o qual permite que uma parte de dados pertença a dois ou mais *clusters*, de acordo com seu grau de pertinência e com a similaridade entre os dados.
Histograma	Diagrama constituído por retângulos ou linhas desenhados a partir de uma linha de base, em que a posição deles ao longo dessa linha representa o valor ou a amplitude de uma das variáveis, e a sua altura, o valor correspondente de uma segunda variável.
Javascript	Linguagem de programação baseada em scripts e padronizada pela ECMA International (associação especializada na padronização de sistemas de informação).
K–Means	Método de agrupamento (*clustering*) que objetiva particionar *n* observações dentre *k* grupos, onde cada observação pertence ao grupo mais próximo da média.
***Machine Learning* (Aprendizado de Máquina)**	Método de análise de dados que automatiza o desenvolvimento de modelos analíticos.
Mapa auto–organizável de Kohonen	Este algoritmo é capaz de diminuir a dimensão de um grupo de dados, conseguindo manter a representação real com relação as propriedades relevantes dos vetores de entrada, tendo–se como resultado um conjunto das características do espaço de entrada.
Outlier	Dados que estão muito distantes dos demais em uma série estatística, e que são chamados comumente de “ponto fora da curva” ou dados espúrios.
Resistência	Capacidade em se opor à passagem de corrente elétrica, mesmo quando exista uma diferença de potencial aplicada.
RStudio	*Software* livre destinado ao ambiente de desenvolvimento integrado para R, uma linguagem de programação para gráficos e cálculos estatísticos.

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

AHP	*Analytic Hierarchy Process* (Processo de Análise hierárquica).
Al $(OH)_3$	Hidróxido de Alumínio.
Al_2O_3	Óxido de Alumínio (Alumina).
AlF/AlF_3	Fluoreto de alumínio.
% ALF / m_ALF	Percentual de Fluoreto de alumínio em excesso no banho químico.
ALF_3A / m_ ALF_3A	Quantidade de fluoreto adicionado no banho.
AP	*Affinity Propagation* (Propagação por afinidade).
BD	Banco de Dados.
BMU	*Best matching unit* (Melhor unidade de correspondência).
CaO	Óxido de cálcio.
CaF/CaF_2	Fluoreto de cálcio.
CSS	*Cascade Style Sheet* (folha de estilo). O CSS é composto por "camadas" e utilizado para definir a apresentação (aparência) em páginas da internet que adotam para o seu desenvolvimento linguagens de marcação (como XML, HTML e XHTML). O CSS define como serão exibidos os elementos contidos no código de uma página da internet e sua maior vantagem é efetuar a separação entre o formato e o conteúdo de um documento.
CSV	*Comma–separated values* (Valores separados por vírgula).
DC	Discriminante Canônica.
DOG	*Difference of Gaussians* (Diferença Gaussiana).
EMF	Tensão efetiva.
FCM	*Fuzzy C–Means*.
Fe_2O_3	Óxido férrico.
FPC	*Flexible Procedures for Clustering* (Procedimentos flexíveis para agrupar).

HTML	*HyperText Markup Language* (Linguagem de Marcação de Hipertexto).
IncTM / m_ IncTM	Incremento de resistência por temperatura.
KDD	*Knowledge Discovery in Databases*.
LVQ	*Learning Vector Quantization* (Aprendizagem de quantização vetorial).
MD	Mineração de Dados.
MDS	*Multidimensional Scaling* (Escalonamento multidimensional).
ML	*Machine Learning* (Aprendizado de Máquina).
NME	Nível de metal.
QALr / m_ QALr	Quantidade de alumina alimentada.
R	RStudio.
RMR	Resistência de Forno.
RNA	Redes Neurais Artificiais.
SiO_2	Dióxido de silício.
SOFM	*Self–organizing feature map* (Mapa auto–organizado de características).
SOM	*Self–Organizing Map* (Mapa auto–Organizável).
TiO_2	Dióxido de titânio.
TMP / m_TMP	Temperatura.
%TOV / m_TOV	Percentual de tempo em alimentação *over–feeding* (super–alimentado, no contexto da dissertação, significa forno com excesso de alumina).
%TUN / m_TUN	Percentual de tempo em alimentação *under–feeding* (sub–alimentação, no contexto da dissertação, significa forno com falta de alumina).
VIDA	Tempo de operação do forno.
VQ	*Vetor Quantization* (Quantificação Vetorial).

SUMÁRIO

1. INTRODUÇÃO

Atualmente as empresas manuseiam em seus sistemas de informações grandes volumes de dados, que em sua maioria são usados para a tomada de decisão e o aprimoramento dos processos realizados. Diante deste cenário, são apresentadas algumas perguntas (Li, Yang e Song, 2009): 'O que fazer com os dados armazenados?', 'Como analisar e utilizar de maneira útil todo o volume de dados disponível?'. 'A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas?' (Bramer,2007). Dentro deste contexto é que surgiu a área de Mineração de Dados. Com uma adequada administração das informações e com técnicas como a mineração de dados é possível encontrar na imensidão de dados, as informações mais essenciais, que dificilmente seriam descobertas sem o auxílio delas, identificando relações não aparentes ou de difícil detecção em sua forma tradicional.

Esta dissertação de mestrado aborda umas das tarefas do processo de mineração de dados: a fase de agrupamento (Larose, 2005). Serão realizados experimentos com quatro técnicas de agrupamento: *Affinity Propagation*, o Mapa Auto–Organizável de Kohonen (SOM), *Fuzzy C–Means* (FCM) e *K–Means*. A ideia central do trabalho é realizar agrupamentos em uma base de dados de fornos de redução de uma fábrica de alumínio, com o intuito de identificar aqueles fornos que possuem comportamentos semelhantes.

A mineração de dados na indústria tem seu uso cada vez mais frequente, visto que a exploração dos dados é uma tarefa comumente utilizada em diversas áreas, e embora se tenha uma série de sistema de mineração de dados comerciais disponíveis hoje, ainda há muitos desafios neste domínio.

Namikka e Gibbon (2002) afirmam que os processos industriais são geralmente vastos, complexos e geram grandes quantidades de dados dimensionais. E esses dados contêm informações potencialmente úteis que podem ser imprescindíveis no controle e otimização dos procedimentos, caso as ferramentas apropriadas sejam aplicadas de forma correta. Os analistas de sistemas industriais estão, portanto, cada vez mais atentos aos dados e métodos de tratamento para auxiliar na análise e desempenho do processo para fins de melhoria, eficiência, produtividade e qualidade do produto.

A indústria de minerais é altamente diversificada, o que torna difícil fazer generalizações sobre sua produção e utilização. Qualquer proposta política ou ideia para

mudá-la ou regulamentá-la deve ser baseada nas diferentes características dos diversos setores da indústria.

De acordo com o Ministério de Relações Exteriores, Comércio Internacional e Cultural (2010), o alumínio em sua forma bruta é o terceiro elemento mais abundante na crosta terrestre depois do oxigênio e do silício, e constitui 7,3% de sua massa. Na sua forma natural, só existe em uma combinação estável com outros materiais (em particular, sais e óxidos) e foi descoberto a sua existência no ano de 1808. A partir de então ainda exigiu muitos anos de pesquisa e testes para isolar o alumínio puro do mineral no seu estado original tornando possível seu modo de produção, comercialização e processamento. Por conseguinte, o alumínio começou a ser produzido com a finalidade comercial na segunda metade do século XIX, sendo considerado ainda novo, em termos comparativos, tendo em conta que a humanidade vem usando outros elementos por milhares de anos antes da descoberta do alumínio. Apesar disso, a produção de alumínio atualmente supera em quantidade a soma dos restantes metais não ferrosos, como o cobre, o chumbo e o estanho.

A indústria de alumínio é uma atividade econômica que gera excelentes ganhos e lucros em muitos continentes, sendo inclusive a fortaleza da economia dos países com padrões de vida mais elevados do mundo (IAI, 2010; CVG Venalum, 2008).

O alumínio é a principal matéria-prima de muitas indústrias, cujos produtos são usados por quase todas as famílias, trazendo formas de empregos fornecidos pelas indústrias e movimentando a economia dos países. A tecnologia da indústria do alumínio e os avanços científicos ajudaram a estimular o crescimento das empresas de muitos países, inclusive o Brasil, além de ajudar a competir com outros produtores líderes.

Desde material de construção até embalagens para comida, o mercado de alumínio fornece aos usuários a tecnologia necessária para construir e desenvolver centenas de indústrias. Ele compete com o aço em várias utilidades, porém tem a vantagem de ser mais leve, resistente a corrosão, um bom condutor de eletricidade, além de ser um metal forte combinado com outros elementos. Este metal possui uma série de características diferenciais das que o define: é leve, tem um terço do peso do aço, não enferruja, é condutor de calor, não é tóxico, não é magnético, é maleável, entre outras características (Argentina, 2010).

Soares (2009) descreve em sua Dissertação de mestrado o projeto de um sensor virtual capaz de estimar a temperatura dos fornos de redução de alumínio, através de variáveis que compõe este processo. Os resultados obtidos foram autenticados pela equipe de processos de uma indústria de alumínio. Com a utilização deste sensor apresentado por Soares (2009), a

fábrica de alumínio ganhou vários benefícios, como a redução de gastos com termopares, que são sensores reais utilizados para a medição direta da temperatura dos fornos; economizará também com energia elétrica, já que o sensor agora é virtual; fará economia de tempo para a realização da tarefa; além de preservar a saúde dos trabalhadores que necessitavam enfrentar condições de periculosidade para realizar a medição da temperatura dos fornos.

Souza (2011) relata em sua Dissertação de Mestrado a estimação da porcentagem de flúor em alumina fluoretada originado de uma planta de tratamento de gases através de um sensor virtual neural capaz de simular o comportamento de um sensor de qualidade de alumina, criado através da técnica de inteligência computacional conhecida como Rede Neural Artificial (RNA). Através deste sensor exposto por Souza (2011), a equipe responsável pelos testes e análises da fábrica de alumínio teve uma opção mais rápida para simular o comportamento da planta de tratamento de gases sem a necessidade de parar com a produtividade da fábrica, o que colabora com a geração de alumina fluoretada de melhor qualidade, contribuindo também com a preservação dos fornos de redução do alumínio, aumentando assim seu tempo de vida e utilidade na fábrica em que este trabalho foi baseado.

Xiangtao et. al. (2006) apresentaram no *Light Metals* (TMS) 2006 um artigo onde os conceitos de *data warehouse* e *data mining* foram introduzidos em sistemas existentes de controle de eletrólise de alumínio, desenvolvendo uma ferramenta de mineração de dados chamada "minerador de dados de eletrólise de alumínio". Esta ferramenta consiste em três componentes principais de mineração: análise de associação cinza, análise de *cluster* baseada em componentes conectados e análise de coeficientes de compensação. Esses modelos de componentes analisam detalhadamente os dados armazenados em um sistema de controle, obtendo conhecimento, ajustando vários parâmetros operacionais e de controle *on-line* e fornecendo suporte de decisão para operadores. Para diferentes usos e perfis e com uma interface de fácil manuseio, os gerentes e operadores podem analisar e extrair minuciosamente os dados armazenados em um sistema de controle e podem ajustar as condições técnicas das células para alcançar a eficiência de produção ideal pelo conhecimento descoberto.

Zhuo et. al. (2008) criaram um modelo para o controle do processo industrial baseado em mineração de dados, que ao analisar os dados históricos usando o conhecimento de técnicas de descoberta conseguiram resolver os problemas de difícil diagnóstico e previsão quando se refere ao estado da célula de eletrólise de alumínio, onde através de parâmetros de controle, o modelo foi usado para realizar a otimização da tomada de decisões e a regularidade de evolução do estado do alumínio, aumentando a eficiência da eletrólise,

prevendo tendências e acidentes que poderiam ocorrer, agrupando o estado das células, reduzindo o consumo de energia e tornando o processo industrial mais inteligente.

Li, Yang e Song (2009) também criaram um modelo de processo de mineração de dados para a eletrólise de alumínio para padronizar o processo de mineração industrial, onde o sistema inteligente aplicado foi usado como o objetivo da mineração de dados. Neste modelo, tanto a mineração de dados quanto o sistema inteligente aplicado são reconhecidos como de importância equivalente, onde o desenvolvimento de sistemas inteligentes aplicado é o objetivo do processo de mineração de dados e também é usado para melhorar a qualidade dos resultados da mineração, mostrando através dos resultados experimentais que o modelo é efetivo para certas extensões.

O processo de mineração de dados envolve o uso de diversas técnicas, e dentre elas está a de identificação de agrupamentos ou *clustering*, que em meio a tantas finalidades, também pode ser utilizada na indústria de minerais como o alumínio, onde através do banco de dados (BD) sua função é formar grupos de registros similares, que compartilham um certo número de propriedades, e desta forma são considerados homogêneos. Assim, o particionamento de BD's volumosos, com diferentes tipos de dados, constituídos de agrupamentos com dados homogêneos relacionados, resulta na aquisição de padrões de comportamento do banco de dados em análise.

A grande vantagem do uso das técnicas de clusterização (ou agrupamento) é que, ao agrupar dados similares, pode-se descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados. Isso fornece um maior entendimento do conjunto de dados original, além de possibilitar o desenvolvimento de esquemas de classificação para novos dados e descobrir correlações interessantes entre os atributos dos dados que não seriam facilmente visualizadas sem o emprego de tais técnicas. Alternativamente, Clusterização pode ser usada como uma etapa de pré-processamento para outros algoritmos, tais como caracterização e classificação, que trabalhariam nos *clusters* identificados. A ideia desse trabalho é utilizar o resultado do agrupamento para melhor selecionar dados de treino de uma rede neural artificial, que é utilizada para modelar o comportamento dinâmico de algumas variáveis do forno de redução de alumínio.

1.1.Objetivo da Dissertação

1.1.1. Objetivo Geral

a) Identificar agrupamentos de fornos de uma fábrica de alumínio através de técnicas de mineração de dados.

1.1.2. Objetivos Específicos

a) Usar técnicas de agrupamento para identificar fornos com comportamentos semelhantes em uma fábrica de alumínio;
b) Verificar a quantidade ideal de grupos para cada técnica utilizada;
c) Comparar o desempenho das técnicas de agrupamento utilizadas.

1.2.Estrutura da Dissertação

Esta dissertação de mestrado está dividida em 7 capítulos.

O capítulo 1 traz a introdução do trabalho, explicando a ideia de mineração de dados e seu uso na indústria do alumínio primário, e o contexto que irá ser abordado ao longo do trabalho.

O capítulo 2 apresenta o processo da produção do alumínio, trazendo conceitos de redução do alumínio, explicando o processo de reação do forno eletrolítico e demonstrando as variáveis utilizadas em todo o processo de modelagem e análise dos dados.

O capítulo 3 aborda o conceito da análise de agrupamento, explicando o método das medidas de agrupamento, os métodos hierárquicos e não– hierárquicos, e trazendo também o conceito de aprendizado não–supervisionado.

O capítulo 4 descreve os algoritmos utilizados nos agrupamentos: o *Affinity Propagation*, o algoritmo Kohonen (SOM), o algoritmo *Fuzzy C–Means* (FCM), e o agrupamento de fornos utilizando o algoritmo *K–Means*.

O capítulo 5 explica a metodologia utilizada no trabalho, mostrando a seleção e extração dos dados, descrevendo a base de dados, as variáveis utilizadas, a disposição dos fornos nas salas de redução e a estatística básica dos dados.

O capítulo 6 apresenta os resultados obtidos da base real de dados e a comparação entre os resultados das técnicas dos algoritmos *Affininity Propagation*, do Mapa Auto – Organizável de Kohonen, *Fuzzy C–Means* e *K–Means*, para medidas utilizando os princípios aritméticos da média, mediana e desvio padrão com filtro e sem filtro.

E por fim, o capítulo 7 traz a conclusão dos experimentos realizados e propostas para trabalhos futuros.

2. A PRODUÇÃO DE ALUMÍNIO

O alumínio é um metal não ferroso, luminoso, com baixo ponto de fusão, macio e deformável, com alta condutividade elétrica e térmica, com ampla utilização na indústria e produzido pelo homem com minerais extraídos da natureza (Argentina, 2010).

Para a produção de alumínio é utilizado a bauxita, um mineral que contém hidróxido de alumínio ($Al(OH)_3$) e outras impurezas como: Fe_2O_3; SiO_2; CaO; TiO_2, etc. Este mineral do qual se extrai o alumínio é abundantemente encontrado principalmente em áreas tropicais e subtropicais, tais como: África, Antilhas, América do Sul e Austrália. Há também algumas minas de bauxita na Europa. A bauxita é refinada para obter óxido de alumínio (alumina) e através de um processo eletrolítico, a alumina é reduzida a alumínio metálico.

As plantas produção de alumínio primário estão localizadas em todo o mundo, geralmente em áreas onde existem abundantes recursos de energia elétrica barata. Estimativas demonstram que são necessários de duas a três toneladas de bauxita para produzir uma tonelada de alumina, e aproximadamente duas toneladas de alumina para produzir uma tonelada de alumínio. A Figura 1 mostra de forma resumida as etapas do processo produtivo da alumina e do alumínio primário.

Figura 1 – Resumo do Processo de Produção da alumina e do alumínio primário.
Disponível em: http://slideplayer.com.br/slide/3344007/
Acesso em: 13 mar. 2017

Na natureza, o alumínio está sempre ligado a algum outro elemento químico na forma de sais ou óxidos, o que torna necessário um processo de separação. Por meio do processo Hall–Héroult, conhecido e utilizado no mundo todo, a quebra por eletrólise da molécula da alumina (Al_2O_3) que contém o alumínio (Figura 2) é feita de tal forma que se consegue obter um alumínio com 99% de pureza (Grjotheim e Kvande, 1993).

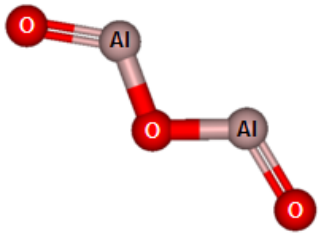


Figura 2 – Molécula de alumina.
Fonte: Souza (2011).

Essa quebra da molécula, representada pela Equação 1 a seguir, requer um gasto exorbitante de energia, sendo necessária a adição de alguns elementos químicos no forno de redução de alumínio que ajudam a reduzir a temperatura para facilitar a eletrólise, economizando energia (Grjotheim e Kvande, 1993). Além disso, esses elementos são de fundamental importância para a manutenção da estabilidade química do forno, pois formam o banho eletrolítico (Souza et al., 2011).

$$2Al_2O_3 + 3C^+ \rightarrow 4Al^+ + 3CO_2 \quad (1)$$

O processo Hall–Héroult é contínuo, ou seja, em condições normais, funciona 24 horas ao dia e sete dias na semana. A alumina é dissolvida em um banho de criolita a uma temperatura em torno de 960ºC sob a passagem de corrente elétrica contínua. No caso da fábrica em questão, é utilizada uma corrente entre 160 kA e 180 kA. Na fábrica, os fornos, também conhecidos como cubas ou, em inglês, *pots* ou *cells*, são dispostos em série, assim, teoricamente, a corrente elétrica é a mesma em todos os fornos. O metal produzido é retirado dos fornos e encaminhado ao lingotamento – no qual o alumínio líquido é solidificado em formas produzindo–se os lingotes, que logo depois são empilhados e estocados para exportação (Kvande, 2014).

Cada etapa citada no parágrafo anterior é alvo de constante controle no interior da fábrica. Além disso, os institutos de pesquisa, principalmente as universidades, têm atuado em

conjunto com os membros de equipes da fábrica, produzindo, através de parcerias em projetos, melhorias que tendem a tornar mais eficiente o processo de fabricação de alumínio.

2.1.REDUÇÃO DO ALUMÍNIO

O alumínio obtido dos fornos eletrolíticos contém uma certa quantidade de impurezas, alumina, criolita e gases, de modo que, para obter uma elevada pureza de alumínio (99,85–99,9%) é necessário que o processo seja submetido a uma redução, tendo duas maneiras principais para esta prática (Bruzos, n.d):

1. **Cloração:** Por este processo, se insufla cloro na massa de alumínio fundido a uma temperatura entre 750–770ºC, durante cerca de 10–15 minutos. No período da insuflação, as impurezas reagem com o cloro e se separam do alumínio, mesmo que uma parte de mais ou menos 1% do alumínio também reaja e se separe, produzindo perdas do material.
2. **Redução eletrolítica:** Para reduzir o alumínio pelo método eletrolítico, as barras de alumínio impuro se colocam como ânodos em um banho de sais de cloro e de flúor. As impurezas do ânodo ficam na solução ou formam a lama anódica. Altos níveis de pureza podem ser obtidos pela técnica de redução eletrolítica em soluções aquosas, com o teor do metal de valor de 99,9% ou maior.

2.2.O FORNO ELETROLÍTICO

O forno, esquematizado pela Figura 3, é o local onde a reação eletrolítica ocorre. É uma construção de aço especialmente preparada para receber os elementos utilizados no processo: banho eletrolítico, alumina, gases, aditivos químicos, corrente elétrica e estruturas de carbono para condução de energia.

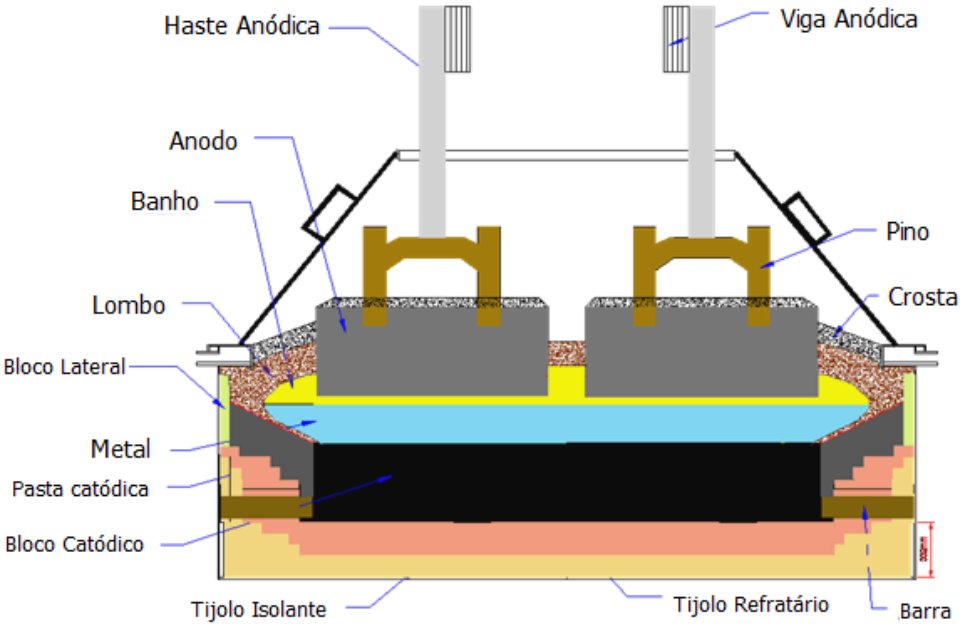


Figura 3 – Forno de Redução de Alumínio e seus Componentes.
Fonte: Grjotheim e Kvande, 1993 (adaptação).

A corrente elétrica atravessa o forno do ânodo ao cátodo. O número de ânodos por forno é proporcional a corrente de operação e o tamanho do mesmo. No caso da indústria em pauta, são 18 ânodos em cada forno. O banho eletrolítico fica entre os ânodos e o cátodo, que é o meio onde ocorre a eletrólise. O alumínio produzido é atraído para o pólo negativo (cátodo), ficando depositado no fundo do forno. Uma crosta sólida se forma ao redor do forno e na sua superfície. Essa crosta é importante, pois isola o forno não permitindo troca de calor excessiva com o meio externo, além disso, ela ajuda a diminuir as emissões de gases para o meio ambiente. Periodicamente, ela é rompida para troca de ânodos.

A temperatura de operação do banho é de 920 a 1000ºC, entretanto, o adequado é manter a temperatura o mais próximo possível da ideal: 960ºC. A criolita pura possui um ponto de fusão alto (1009ºC) e, para reduzi–lo, são adicionados alguns aditivos químicos, principalmente o fluoreto de alumínio (AlF_3), o fluoreto de cálcio (CaF_2) e a própria alumina (Al_2O_3). A finalidade dos aditivos químicos também é manter a composição química do forno estável.

A resistência comum de um forno está em torno de 15 µOhm, mantendo a tensão em torno de 4,2V. A tensão efetiva, ou seja, aquela que é gerada pela própria reação química e resulta na produção do alumínio corresponde a 1,65V por forno. É fundamental que a resistência se mantenha estável para que a tensão efetiva não se altere, pois causaria uma queda na produtividade. Uma resistência muito alta pode provocar superaquecimento no forno, comprometendo o ânodo e em casos extremos o cátodo (Grjotheim e Kvande, 1993). O

comprometimento do cátodo provoca o desligamento do forno. Já uma resistência muito baixa provoca uma queda na eficiência de corrente. Observa–se que a relação tensão versus resistência é regida pela Primeira Lei de Ohm, descrita através da equação 2:

$$V = R.i \tag{2}$$

Para manter a resistência estável, é necessário controlar a distância ânodo–cátodo a um nível que mantenha a resistência próxima de seu *setpoint*. Os ânodos são consumidos continuamente pelo eletrólito, fazendo com que a distância com o cátodo aumente, elevando assim a resistência do forno. Existe um sistema de controle de resistência capaz de monitorar continuamente as variáveis online do forno, identificando se há necessidade de subir ou de descer o ânodo. Ressalta–se que esse sistema de controle normalmente não movimenta um ânodo individualmente, mas todos ao mesmo tempo.

Na fábrica estudada, os fornos estão dispostos em quatro reduções e cinco linhas de fundição. Nas linhas de fundição estão ordenados os fornos eletrolíticos, onde cada redução contém um total de 240 fornos. Visando a otimização do controle administrativo e uma melhor manutenção, os fornos são divididos em sessões, as quais são submetidas aos mesmos turnos e iguais procedimentos de manutenção programada. No total, a fábrica conta com 960 fornos, dispostos em oito salas com 120 fornos cada uma. A Figura 4 mostra o layout onde encontra-se organizado a área de redução I que contém os fornos de redução de alumínio da fábrica que este trabalho foi baseado.

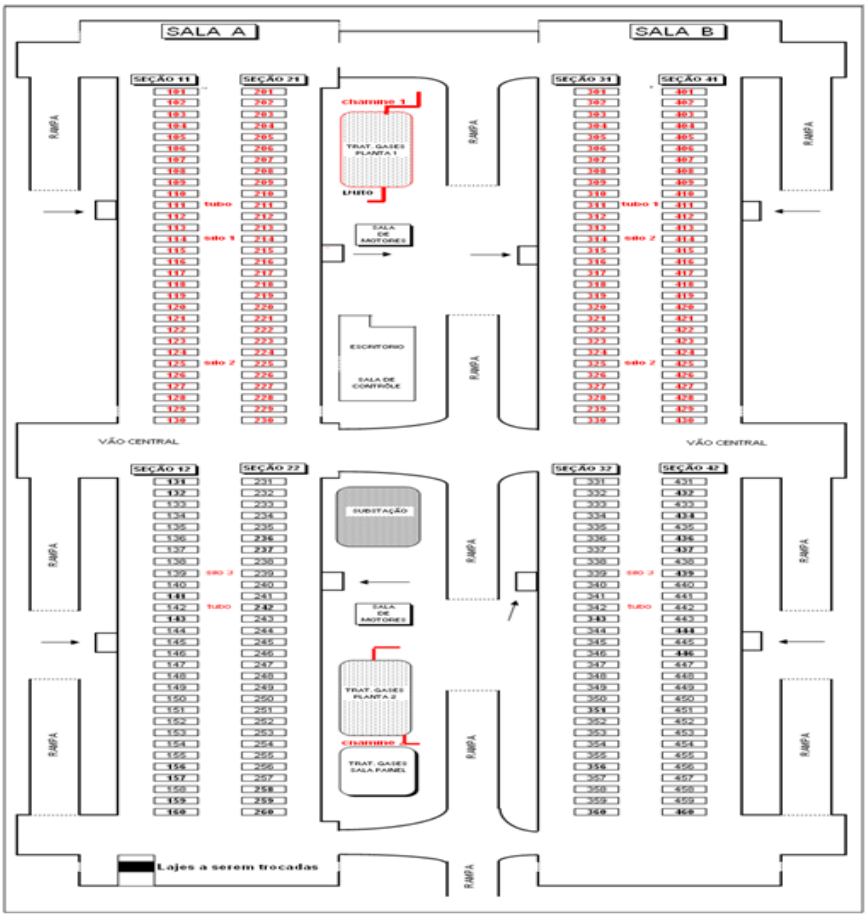


Figura 4 – Layout da Área de Redução I.

Cada forno possui um sistema de controle individual, de modo que alguns fornos se comportam de maneira semelhante. Diante deste acontecimento, é necessário conhecer quais fornos tem comportamento iguais para assim tratá-los e estudá-los de maneira igual, levando em conta o objetivo de "desenvolver um modelo" baseado em dados reais dos fornos, para que ao invés de utilizar todos os 960 fornos, através do agrupamento seja possível utilizar apenas '*n*' fornos, sendo um forno de cada agrupamento, de acordo com as suas características e similaridades. Além disso, de acordo com a técnica de agrupamento utilizada, também é possível "extrair conhecimento" dos dados por meio de regras associadas a eles.

Nesta dissertação será dado ênfase a extração de conhecimento do funcionamento dos fornos. Por este motivo, o objetivo deste trabalho é realizar o agrupamento (*clustering*) dos fornos de redução de alumínio - levando em consideração dados reais disponibilizados por uma fábrica produtora deste metal - através da técnica de inteligência computacional conhecida como SOM (*Self Organizing Map* - Mapa Auto-Organizável), Propagação por afinidade (*Affinity Propagation*), *Fuzzy C-Means* e *K-Means*. Após o agrupamento, será possível descobrir regras associativas, escondidas nos dados, que ajudam a explicar o processo de produção de alumínio realizado pela fábrica.

2.3.VARIÁVEIS MAIS UTILIZADAS EM MODELAGEM E ANÁLISE DE COMPORTAMENTO

A escolha das variáveis se deu através de uma análise com a equipe de processo de uma determinada fábrica de alumínio, e de uma pesquisa bibliográfica sobre o controle químico do banho eletrolítico, para que fosse possível a compreensão do comportamento dos fornos e de qual modo cada componente interferia de forma direta e indireta na temperatura dos fornos, para assim definir quais variáveis possuíam maior grau de influência nesse processo. De acordo com (McFadden et al.,2001), a temperatura está fortemente relacionada com a composição química do eletrólito, que envolve as variáveis de Fluoreto de Alumínio (ALF), Fluoreto de Cálcio (CAF) e alumina (Al_2O_3). Haupin e Kvande (1993) em seu trabalho sobre balanço térmico, identificaram que o tempo de funcionamento (VIDA) e a alumina fluoretada pelas plantas de tratamento de gases possuem alguma relação direta com o consumo de fluoreto em excesso no banho, que impacta diretamente na temperatura dos fornos. Frost e Karri (1999) desenvolveram um trabalho de estimação da quantidade de fluoreto em excesso nos fornos, utilizando Redes Neurais Artificiais (RNA), na qual faz uso da Resistência de Forno (RMR), da tensão efetiva (EMF) e do nível de metal (NME) na inferência do valor de fluoreto (ALF), que possui forte correlação com a temperatura (Soares, 2009).

É inviável considerar todas as variáveis do processo para o agrupamento, por esta razão, as sete variáveis abaixo foram escolhidas por possuírem uma relação satisfatória com a clusterização dos fornos quanto à química de banho, fazendo com que se tenha uma boa qualidade e confiança nos dados utilizados. Estas variáveis foram as mesmas que Soares (2009) usou no modelo neural de sua dissertação de mestrado para prever a temperatura do forno, quais sejam:

a) Temperatura (TMP);
b) Fluoreto de alumínio (% de ALF no Banho);
c) Quantidade de fluoreto adicionado no banho (ALF_3A);
d) Quantidade de alumina alimentada (QALr);
e) Incremento de resistência por temperatura (IncTM);
f) Percentual de tempo em alimentação *under–feeding* (%TUN);
g) Percentual de tempo em alimentação *over–feeding* (%TOV).

3. MINERAÇÃO DE DADOS

De acordo com Castro e Ferrari (2016), a mineração de dados (do inglês *Data mining*) se originou em 1990 como campo de pesquisa, no entanto, suas aplicabilidades nas áreas da matemática, estatística e computação deram início muito antes deste período, e ganhou destaque com o surgimento da expressão *Big Data* e com a publicação do relatório denominado *Big Data: The next Frontier for Innovation, Competition, and Productivity* pelo *McKinsey Global Institute* (2011). Ela é a principal responsável pela análise e preparação do *Big Data* com sua grande massa de dados. A técnica de mineração equivale a extração de minerais preciosos, assim como a extração do ouro em jazidas. O termo "mineração de dados" (MD) foi originado em decorrência ao processo de mineração de pedras preciosas, já que o mesmo explora uma base de dados, que pode ser comparado a uma mina, fazendo o uso e algoritmos, que são equiparados às ferramentas utilizadas na extração dos minerais preciosos.

Esse processo de exploração de grandes quantidades de dados é uma das tecnologias mais promissoras da atualidade, já que empresas gastam centenas de milhões para realizar a coleta dos dados, e em algumas das vezes nenhuma informação é obtida, necessitando assim a realização de um tratamento e análise de um especialista. Ela pode ser definida inicialmente como um processo de descobrimento de novas e significativas relações, padrões e tendências ao examinar grandes quantidades de dados (César, 2007).

A mineração de dados trabalha com o processo de extração do conhecimento em base de dados (*Knowledge Discovery in Databases* – KDD). Castro e Ferrari (2016) afirma que embora muitos autores utilizem Mineração de Dados como sendo sinônimo de KDD, em 1995 foi proposto na primeira conferência internacional de KDD, sediada na cidade de Montreal - Canadá, onde a nomenclatura descoberta de conhecimentos em bases de dados seria atribuída a todo processo de extração de conhecimento a partir de um banco de dados, e não somente a etapa de Mineração de Dados.

Embora o processo de KDD seja mais extenso (Castro e Ferrari, 2016), nesta dissertação de mestrado serão consideradas as seguintes etapas, que de forma sintetizadas são as etapas principais do processo de extração do conhecimento:

1. **Seleção da base de dados:** Conjunto de dados informativos e organizados em um mesmo contexto. Esta etapa consiste em buscar o objetivo e as ferramentas do processo de mineração, identificando os dados a serem extraídos, buscando os atributos apropriados de entrada e a informação de saída para representar a tarefa.

Isto significa que antes de iniciar o processo é necessário saber o que se quer obter como resposta ou resultado, e quais são os dados que facilitarão essa informação em busca da meta a ser alcançada.

2. **Pré-Processamento (Limpeza dos dados):** Etapa que antecede a mineração de dados, onde as informações julgadas desnecessárias são removidas. Esta etapa consiste em limpar dados considerados "sujos", que são os dados incompletos, onde há valores perdidos de atributos; ruídos, que são os valores incorretos ou inesperados; e dados inconsistentes, os quais contêm valores e atributos com nomes diferentes. Esta etapa é necessária para que os dados considerados sujos (espúrios) não venham a contribuir para uma análise inexata, ocasionando resultados incorretos.
3. **Mineração de dados:** Esta etapa do processo consiste na busca dos padrões de interesse e no emprego de algoritmos capazes de obter conhecimentos a partir dos dados pré-processados. Esta é a fase onde são aplicados métodos inteligentes, com objetivo de extrair padrões antes desconhecidos e potencialmente úteis, que até então estavam contidos na base de dados, mas de difícil detecção em razão do grande conjunto de informações.
4. **Validação do conhecimento e interpretação dos resultados:** É a etapa onde os padrões obtidos são estudados para verificar se são realmente interessantes e úteis, com base em algumas medidas e realizando uma avaliação dos resultados obtidos. Consiste também em entender os resultados da análise. Esses padrões identificados pelo sistema são interpretados em conhecimento e dão suportes à tomada de decisão humana, ou seja, são usados para sintetizar e responder problemas e questões que antes eram complexas.

A figura 5 demonstra as etapas do processo de extração do conhecimento através do uso de uma base de dados, com a finalidade de extrair informações até então desconhecidos e potencialmente relevantes. Através da figura 5 é possível notar que as etapas estão correlacionadas e interligadas, de modo que é necessário considerar essas inter-relações e suas influências no resultado final.

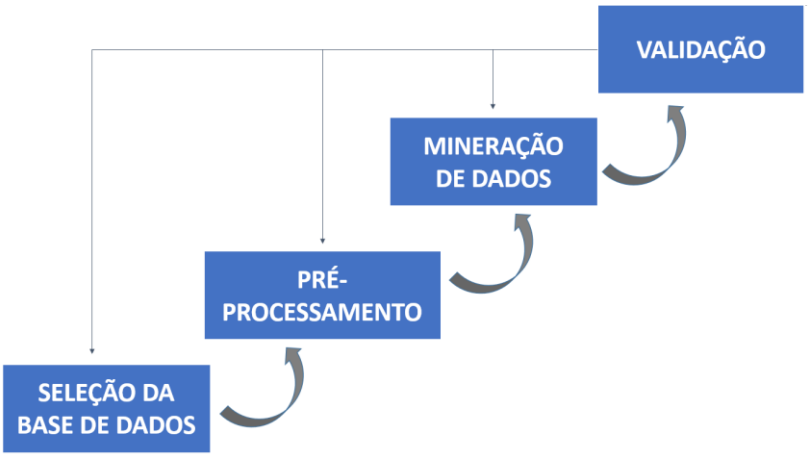


Figura 5 – Processo de descoberta de conhecimento com o uso de bases de dados.

Na prática, os métodos da mineração de dados estão divididos em aprendizado supervisionado, que consiste nas atividades preditivas, e o aprendizado não – supervisionado, o qual onde é encontrado as atividades descritivas. O aprendizado não – supervisionado não necessita de algum atributo ensinando como deve ser feito, ou seja, o experimento terá que descobrir sozinho as informações dos dados que estão sendo apresentados e codificá-las nas saídas. Já no aprendizado supervisionado existe algum atributo, que geralmente são dados já treinados, que auxiliam e avaliam se está coerente a resposta da rede em relação ao padrão atual de entrada.

Nesta dissertação de mestrado será utilizado o aprendizado não-supervisionado, dado que a clusterização de dados ou análise de agrupamento é uma prática de mineração de dados que tem por objetivo encontrar similaridades entre as *n* amostras da base de dados, usando algoritmo de aprendizado não-supervisionado.

A figura 6 mostra as atividades e tarefas da mineração de dados. Vale ressaltar que a mineração de dados possui várias etapas, quais sejam: agrupamento, classificação, estimação e predição; no entanto esta dissertação irá abordar apenas a etapa de agrupamento.

Figura 6 – Atividades e tarefas da Mineração de dados.

3.1.ANÁLISE DE AGRUPAMENTO

A análise de *cluster*, conhecida como análise de agrupamento, é uma técnica estatística multivariada que busca agrupar elementos (ou variáveis) tentando alcançar a máxima homogeneidade em cada grupo e a maior diferença entre os demais grupos. A análise de agrupamento (em Inglês, *cluster analisys*) é uma técnica multivariada que permite agrupar os casos ou variáveis de um arquivo de dados em função do grau de similaridade entre eles. O dendrograma é uma representação gráfica que melhor ajuda a interpretar o resultado de uma análise de agrupamento. Ele representa matemática e ilustrativamente todo o procedimento de agrupamento através de uma estrutura de árvore (Everitt et al., 2001). Os nós do dendrograma representam agrupamentos, e são compostos pelos grupos e/ou objetos (grupos formados apenas por ele mesmo) ligados a ele (nó). Caso o dendrograma seja cortado em um nível de distância desejado, se obtém uma classificação dos números de grupos existentes nesse nível e dos indivíduos que os formam. O número de grupo dos indivíduos é obtido pelo corte do dendrograma em um nível desejado, e então cada componente conectado forma um grupo.

O processo de agrupamento é uma das mais antigas funções cerebrais desenvolvidas pelo homem. No século V a.c. os filósofos gregos já refletiam sobre esta função cerebral de agrupamento. Em geral, se pode afirmar que o homem identifica objetos, observa e mede características, e também realiza agrupamentos de objetos com base nessas características para encontrar alguma finalidade específica que tenha sido levantada.

A técnica de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação, pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado).

Incorporando um projeto de mineração de dados com diferentes agentes tecnológicos em uma indústria se contribui com o dinamismo do processo por se tratar de um modelo de aprendizagem contínuo. A modernização tecnológica nos modelos de manufatura ligada a facilidade de processamento e análise de informação se convertem em recursos valiosos para a compreensão do comportamento dos sistemas e o contínuo melhoramento do processo (Harding, 2008).

Desde a década de 1970, as áreas de inteligência artificial, reconhecimento de padrões e *Machine Learning* (ML) têm trabalhado na modelagem do processo de agrupar objetos de acordo com suas características e geração de algoritmos que permitam realizar tais

agrupamentos de maneira automática; e com base nestes estudos foram desenvolvidos vários métodos de agrupamento, também conhecido como clusterização. Estes métodos formam um conjunto de padrões, onde cada padrão representa um objeto com uma quantidade fixa e igual de atributos, que representam as características do objeto, e proporciona uma partição do conjunto de padrões indicando a pertinência de cada padrão a cada um dos grupos encontrados.

As técnicas de análise de agrupamento exigem de seus usuários a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da medida de dissimilaridade, bem como pela definição do número de grupos (Gower e Legendre, 1986; Jackson et al., 1989; Duarte et al., 1999).

Como técnica de agrupamento de variáveis, a análise de agrupamento é semelhante à análise fatorial, mas enquanto que a fatoração é inflexível em algumas de suas suposições (linearidade, normalidade, variáveis quantitativas, etc.) e sempre estima da mesma maneira a matriz de distância, o agrupamento é menos restritivo em suas suposições (não exige linearidade, não exige simetria, permite variáveis categóricas, etc.) e admite vários métodos de estimativa da matriz de distância.

Já como técnica de agrupamento de casos, a análise de *cluster* é semelhante à análise discriminante. No entanto, enquanto que a análise discriminante realiza a classificação tomando como referência um critério ou uma variável dependente (grupos de classificação), a análise de agrupamento permite detectar o número ideal de grupos e sua composição, de acordo com a similaridade existente entre eles. Além disso, a análise de agrupamento não assume nenhuma distribuição específica para as variáveis.

4. AGRUPAMENTO UTILIZADO NO FORNO DE REDUÇÃO DE ALUMÍNIO

4.1. AGRUPAMENTO DE FORNOS VIA *AFFINITY PROPAGATION*

Em estatística e em mineração de dados, o *Affinity Propagation* (AP), que em Português é denominado de Propagação por Afinidade, é um algoritmo de *cluster* baseado no conceito de "passagem de mensagens" entre pontos de dados. Ao contrário dos algoritmos de *clustering* como *K–Means* ou *k–medoids*, o *Affinity Propagation* não exige que o número de *clusters* seja determinado ou estimado antes de executar o algoritmo. Semelhante ao *k–medoids*, o algoritmo é responsável por encontrar "exemplares", membros do conjunto de entrada que são representativos de *clusters*, que serão descritos ao longo deste capítulo (Brendan e Dueck, 2007).

A técnica é responsável por selecionar um determinado número de *clusters* de acordo com a base de dados existente. O *Affinity Propagation* usa como conjunto de dados as principais semelhanças entre os mesmos, onde as semelhanças s (i, k) indicam quão adequado são os dados de *k* para cada ponto de *i*. Quando o objetivo é minimizar os erros quadrados, cada similaridade é estabelecida como sendo o inverso do erro quadrado (distância euclidiana).

Sendo s (k, k) um número real, se pode inferir que para cada ponto *k*, seus pontos serão escolhidos como pontos principais. Estes pontos são denominados pontos exemplares. O número de pontos exemplares será o número de *clusters*, influenciados pelos valores de entrada exemplares. Em princípio, se sugere que todos os dados possam ser eleitos como tais, mas este ponto pode ser transformado para produzir o número de *clusters*. O valor compartilhado pode ser a mediana (caso se trate de um número moderado de *clusters*), ou seus mínimos (se for o resultado de um número pequeno de *clusters*).

O algoritmo funciona como se segue: dado o conjunto de dados bi–dimensionais, em que a Distância Euclidiana é utilizada como medida de similaridade. Cada cor é estipulada ao ponto dependendo da evidência de ser o centro do *cluster*, e as distâncias entre um ponto i e um ponto *k* são medidas mediante a força que pode ser transmitida entre si. A responsabilidade r (i, k) é enviada entre pontos, indicando qual é o ponto forte em relação a um outro ponto exemplar. A disponibilidade a (i, k) é enviada a partir dos candidatos aos pontos para indicar o grau em que estes podem ser o centro do *cluster*. Em seguida, é mostrado o efeito do valor da preferência de entrada (comum para todos os pontos de dados) no número de exemplares identificados (número de grupos).

4.1.1. TÉCNICA DE AGRUPAMENTO VIA *AFFINITY PROPAGATION*

A técnica *Affinity Propagation* (AP) foi empregada para realizar o agrupamento dos fornos. Esta técnica é responsável por selecionar um determinado número de *clusters* de acordo com a base de dados existente. Os parâmetros para realizar experimentos através da função **ap*cluster*** são:

- Matriz de similaridade;
- Matriz de dados.

O *Affinity Propagation* é uma técnica relativamente recente para agrupamento de dados (Bodenhofer et al., 2017). Possui como grande diferencial a não necessidade de fixar o número de *clusters* a priori. Este algoritmo identifica exemplares entre os dados e forma *clusters* em volta destes exemplares. Ele trabalha, simultaneamente, considerando todos os dados como exemplar em potencial e trocando mensagens entre os dados até que um bom conjunto de exemplares e grupos seja alcançado (Bodenhofer et al., 2017).

Um programa na linguagem R foi desenvolvido para ler a matriz de dados. Em seguida, a função que implementa o AP, chamada (*apcluster*) e várias saídas foram calculadas e analisadas, de acordo com o código descrito na Tabela abaixo:

Tabela 1 – Código–Fonte programado em R para agrupamento usando *Affinity Propagation*.

```
#carrega biblioteca para uso do AP
library("apcluster")
#importa os dados do arquivo CSV para a variável "Fornos"
Fornos = read.table("planilhaDados.csv", sep=";", header=T)
#executa o algoritmo AP com os seguintes parâmetros:
#1) matriz de similaridade (r=7, pois são sete variáveis)
#2) matriz de dados
resultado <- apcluster(negDistMat(r=7), Fornos)
#mostra os resultados provenientes do AP
resultado
#mostra o gráfico do agrupamento de cada instância comparando as variáveis em par
plot(resultado, Fornos)
```

4.2. AGRUPAMENTO DE FORNOS USANDO O MAPA AUTO–ORGANIZÁVEL DE KOHONEN (SOM)

Em 1982 Teuvo Kohonen apresentou um modelo de rede denominado mapa auto–organizável ou SOM (*Self–Organizing Maps*), com base em determinadas evidências descobertas a nível cerebral. Este tipo de rede possui um aprendizado não–supervisionado competitivo (Kohonen, 1982).

Os mapas auto–organizados de atributos, conhecidos como *Self–Organizing Maps* (SOM) representam um tipo especial de quantificação vetorial na qual existe uma ordem ou topologia imposta aos vetores codificados. O objetivo do SOM é representar dados de alta dimensionalidade em uma estrutura de baixa dimensão, usualmente 1 ou 2 dimensões, que capturam a estrutura original dos dados. Agrupamentos distintos dos dados são mapeados a distintos agrupamentos de vetores codificados em uma matriz. O SOM é citado como um método para exploração de dados (Webb,2002).

Wehrens e Buydens (2007) comparam o SOM proposto por Kohonen em 2001 ao Escalonamento Multidimensional – (MDS), mas que o SOM ao invés de tentar reproduzir distâncias, busca reproduzir uma dada topologia, ou em outras palavras, tenta manter os mesmos vizinhos. Desta forma se dois objetos estão próximos em alta dimensionalidade, então a posição dos mesmos em baixa dimensão também será similar. Outra característica do SOM é mapear os objetos em uma grade regular de unidades.

Para treinar um mapa de Kohonen, primeiro é necessário que se tenha um conjunto de dados, que serão divididos em três conjuntos: treinamento, teste e validação. O treinamento, como o próprio nome infere, fará o treinamento do mapa em análise, o teste é feito para selecionar um mapa entre vários, e posteriormente ocorre a validação, cujo o objetivo é definir o erro final. Esta é apenas uma das possíveis maneiras de treinar um mapa. Para definir o erro total de um mapa com determinados parâmetros, é possível usar outra metodologia diferente chamada deixar–k–fora (*leave–k–out*), a qual consiste em dividir o conjunto de dados em *k* partes, onde se usa k–1 para treinamento e 1 para teste; o procedimento se repete para as *k* partes em que se divide o conjunto de treino.

Raramente os conjuntos servem em sua forma primitiva; muitas vezes eles possuem valores em escalas e variabilidades muito diferentes, e até mesmo são categóricos. Por isso, normalmente, antes do treino, é necessário que se faça algum tipo de pré–processamento, de

modo que todas as variáveis tenham aproximadamente a mesma categoria e o mesmo desvio padrão. A maneira mais segura ocorre da seguinte forma:

Convertendo variáveis categóricas de *n* categorias em *n* variáveis (também pode ser convertido em *n–1* variáveis). Os valores destas variáveis dependem de como será feito o pré–processamento do resto do arquivo: normalmente é colocado o valor máximo se a variável seja correspondente a essa categoria e o valor mínimo caso contrário. Por exemplo, em uma variável categórica com três categorias, a segunda categoria seria codificada como [min, max, min] sendo [–1, 1, –1].

Normalizando o restante das colunas, subtraindo a média e dividindo pelo desvio padrão. Assim, todas as variáveis terão um intervalo entre –1 e 1.

Outro possível pré–processamento é a aplicação de logaritmos, se o intervalo de variação passar por várias ordens de grandeza, ou subtraindo o valor mínimo e dividindo pelo intervalo de variação, para dar variáveis diferentes no intervalo [0,1]. No entanto, quando se trata de variáveis muito diferentes, especialmente no caso em que a distribuição das variáveis seja muito diferente, os modelos encontrados não são suficientemente bons, devendo submetê–los a algum pré–processamento adicional.

Uma vez que o conjunto esteja bem pré–procesado, em um formato que seja adequado para passá–lo para o programa correspondente, é realizado a parte do treinamento em si, e a arquitetura da rede escolhida (hexagonal ou quadrado). Em suma, consiste em realizar a inicialização dos vetores do mapa de Kohonen. Pode ser feito de várias maneiras diferentes, tanto aleatoriamente dentro do intervalo de variação de cada variável, e com uma distribuição determinada: uniforme, gaussiana; ou aleatoriamente com valores pequenos, muito menor do que o intervalo de variação, usando vetores do conjunto de treinamento, o qual não se faz necessário a preocupação com o intervalo de variação; ou ainda usando algum algoritmo, tal como *K–Means*, que permite ter inicialmente uma população de vetores cobrindo mais ou menos espaço de entrada. E consiste também em gerar o treinamento, escolhendo aleatoriamente um vetor do conjunto. Calcula–se o vetor mais próximo entre aqueles do mapa, e se denomina o vencedor; alterando o valor desse vetor e de outros vetores vizinhos, de forma que eles se aproximem mais ao vetor de entrada. O tamanho da vizinhança diminuirá ao longo do treinamento; e essa é a peça chave da auto–organização.

Os mapas auto–organizáveis de Kohonen são um algoritmo, às vezes agrupados dentro das redes neuronais, que a partir de um processo de treinamento agrupa os dados; este agrupamento faz com que a projeção desses dados sobre o mapa distribua suas características

de uma forma gradual. O Mapa Kohonen SOM (*self–organizing map*, mapa de auto–organização) ou SOFM (*self–organizing feature map*, mapa auto–organizado de características) é usado para diferentes aplicações:

- *Clustering*: É possível agrupar dados do conjunto de entrada, atendendo a diferentes critérios.
- Visualização: Este agrupamento, como se realiza de uma forma ordenada, permite visualizá–lo e descobrir características novas, ou relações que não haviam sido planejadas com antecedência. Também permite visualizar a evolução temporal de um conjunto de dados.
- Classificação: Uma vez calibrado o mapa, ou atribuído algum tipo de etiqueta a cada *cluster*, pode ser usado para classificar em dados desconhecidos.
- Interpolação de uma função: Atribuir valores numéricos para cada um dos nós da rede de Kohonen, podem ser atribuídos esses valores numéricos aos vetores de entrada.
- Quantização vetorial, ou seja, aplicação de uma entrada contínua a uma saída que está discretizada, isto é, obter a partir de um vetor qualquer o vetor mais próximo de um conjunto previamente estabelecido.

Os SOMs são algoritmos não–supervisionados, e como o próprio nome infere, é sem supervisão e totalmente dependente dos próprios dados, ou seja, não existe um agente externo indicando a resposta desejada para os padrões de entrada. Isso significa que eles não usam a etiqueta dos dados para o treinamento; no entanto, na maioria dos casos, se usam algum tipo de etiqueta para exibir os dados adequadamente. Este tipo de aprendizado também é conhecido como aprendizado auto–supervisionado ou de auto–organização por que não requer saída desejada e/ou não precisa usar supervisores para seu treinamento. O critério de validação envolve índices que afirmam a qualidade dos dados distribuídos entre os grupos. São caracterizados também como algoritmos competitivos, porque o treinamento é feito em um só "neurônio" de cada vez, o que significa que a representação de cada região do espaço de entrada é concentrada por neurônios não distribuídos (como muitas vezes acontece em outras redes neurais, tais como o perceptrón multicamada, treinadas pelo algoritmo *Backpropagation*). O aprendizado competitivo consiste na disputa entre os neurônios de uma camada x pelo privilégio de permanecerem ativos, tal que o neurônio com maior atividade seja o único a participar do processo de aprendizado. Assim, enquanto uma rede neural

baseada em aprendizado Hebbiano, vários neurônios de saída podem estar simultaneamente ativos, no caso do aprendizado competitivo, somente um neurônio de saída fica ativo a cada vez.

Kohonen foi o pioneiro no desenvolvimento da teoria das redes competitivas. Esta rede é uma estrutura de duas camadas de neurônios. A primeira camada é a de entrada e seus neurônios estão completamente interconectados aos neurônios da segunda camada denominada competitiva, que é organizada em um arranjo dependente do objeto a ser mapeado.

Conforme Kohonen (1987), o esquema básico de um modelo de Kohonen faz com que neurônios da camada de saída disputem entre si a representação da informação apresentada aos neurônios de entrada. Havendo um neurônio vencedor, este é reajustado para responder ainda melhor ao estímulo recebido. Dentro deste modelo não–supervisionado, não somente o vencedor, mas também os seus vizinhos (dentro de um senso físico) são ajustados. O processo de aprendizagem do SOM segue da seguinte forma:

Passo 1. Um vetor x é selecionado de forma aleatória de acordo com o conjunto de dados, e a partir dele é calculada sua distância (similaridade) aos vetores da Tabela de codificação, utilizando, por exemplo, a distância euclidiana:

$$\|x - m_c\| = min_j \left\{ \left\| x - m_j \right\| \right\} \qquad (3)$$

Passo 2. Uma vez que se já encontrado o vetor mais próximo ou BMU (*best matching unit* – melhor unidade de correspondência) o resto dos vetores do *codebook* é atualizado.

O BMU e os seus vizinhos (no sentido topológico) se movem próximo do vetor x no espaço de dados. A magnitude desta atração é determinada pela taxa de aprendizagem.

Enquanto o processo é atualizado e novos vetores são atribuídos ao mapa, a taxa de aprendizagem diminui gradualmente em direção a zero, e junto com ela, também diminui o raio de vizinhança.

A regra de atualização para o dado vetor de referência *i* é:

$$m_j\,(t+1) = \begin{cases} m_j\,(t) + \alpha\,(t)\left(x(t) - m_j\,(t)\right) & j \in N_c\,(t) \\ m_j\,(t) & j \notin N_c\,(t) \end{cases} \qquad (4)$$

Os passos 1 e 2 são repetidos até que o treinamento termine. O número de passos de treinamento deve ser definido a priori, para que seja possível o cálculo da taxa de convergência da função de vizinhança e da taxa de aprendizagem.

Uma vez que o treinamento seja concluído, o mapa se ordena em sentido topológico: *n* vetores topologicamente próximos interagem com *n* neurônios adjacentes ou até mesmo com o mesmo neurônio.

4.2.1. MÉTODOS DE PROJEÇÃO MULTIDIMENSIONAL

O mapa de Kohonen também pode ser usado para a projeção multidimensional, como já foi visto em alguns algoritmos de aprendizado não– supervisionado. Na verdade, é um método de projeção não–linear. Os métodos de projeção lineares, tais como análises de componentes principais, tentam encontrar um subespaço de poucas dimensões que contenha a variação máxima da amostra de entrada; uma vez encontrada, se pode projetar as amostras sobre esse espaço, e usá–las para classificação supervisionada (Wehrens e Buydens, 2007).

No entanto, nem sempre se pode projetar linearmente algumas dimensões. Nestes casos, deve ser utilizado um sistema de projeção não–linear. Um deles é chamado de escalonamento multidimensional (*Multidimensional Scaling*, MDS), o qual utiliza as distâncias entre os diferentes elementos da amostra e os projeta a um espaço de dimensão inferior, geralmente dois. Este algoritmo tenta manter as relações topológicas, assim como o algoritmo denominado mapa Sammon, um algoritmo que projeta um conjunto bidimensional de dados n–dimensionais, mantendo as relações métricas, ou seja: o que está próximo em n–dimensões também deve estar próximo em 2 dimensões.

Atualmente estão sendo desenvolvidos algoritmos mais poderosos, tais como o de análise de componentes curvilíneos, desenvolvido a partir do mapa de Kohonen, e que melhora o mapa de Sammon no sentido de favorecer a reprodução de distâncias curtas na projeção em relação a distâncias longas.

4.2.2. MEDIDAS DE QUALIDADE E PRECISÃO DO MAPA

Uma vez que se tenha treinado o mapa, é importante saber se houve a adaptação adequada aos dados de treinamento. Como medidas de qualidade dos mapas é considerada a precisão da projeção e a preservação da topologia.

A medida de precisão da projeção descreve como os neurônios se adaptam ou respondem aos dados. Normalmente, o número de dados é maior que o número de neurônios e o erro de precisão é sempre diferente de 0.

Para calcular a precisão da projeção é utilizado o erro médio de quantização sobre o conjunto completo dos dados. Essa medida está baseada no cálculo da média das distâncias entre os dados e o vetor que representa a região onde os dados estão localizados, e é calculada de acordo com a equação 5:

$$\varepsilon_q = \frac{1}{n}\sum_{j=1}^{n} \left\|\overrightarrow{x_j} - \overrightarrow{C_i}\right\| \qquad (5)$$

onde *n* é o número total de amostras quantizadas, $\overrightarrow{x_j}$ é um dado e $\overrightarrow{C_i}$ é o vetor representante do grupo ou classe à qual o dado $\overrightarrow{x_j}$ pertence, sendo $i = \arg min\, {}^{c}_{i=1} d\left(\overrightarrow{C_i}, \overrightarrow{x_j}\right)$.

4.2.3. TÉCNICA DE AGRUPAMENTO VIA REDE NEURAL ARTIFICIAL (RNA)

Uma das técnicas empregadas para realizar o agrupamento dos fornos foi a Rede Neural Artificial (RNA). O algoritmo de treinamento da rede é o Mapa Auto–organizável de Kohonen, cuja sigla SOM e, *Self–organized map*, em inglês; e um programa na linguagem R foi desenvolvido para ler a matriz de dados de um arquivo CSV (*Comma–separated values*), o qual possui os dados separados por vírgula. Em seguida, o programa realiza a normalização dos dados entre 0 e 1 e executa o algoritmo de Kohonen para encontrar os grupos (*clusters*) de fornos de redução de alumínio. Por conseguinte, quatro gráficos foram gerados para analisar o mapa resultante. Este programa se encontra abaixo e mais detalhes estão descritos em comentários que começam com o caractere “#”.

Tabela 2 – Código–Fonte programado em R para agrupamento usando Kohonen (SOM).

```
#carrega as bibliotecas utilizadas pelo programa
library("kohonen");
library("cluster");

#importa dados de um arquivo CSV para o R
Fornos <– read.csv2(file='planilhaMédias_EstimadorTemp.csv');
#realiza a normalização dos dados entre 0 e 1
Fornos.sc <– scale(Fornos);
#executa o algoritmo SOM de Kohonen com os seguintes parâmetros:
#1) recebe dados normalizados
#2) grade/mapa 3x3 (resultando em 9 grupos); cada grupo possui geometria hexagonal
#3) número de épocas/iterações: 1000
Fornos.som <– som(data = Fornos.sc, grid = somgrid(3, 3, "hexagonal"), rlen=1000);

#gera gráfico da relação entre a distância média para o grupo mais próximo e a iteração
plot(Fornos.som, type = "changes", main = "Distância média x Iteração")

#gera gráfico de agrupamento dos Fornos de acordo com a matriz de dados utilizada
plot(Fornos.som, main = "Agrupamento das Fornos")

#gera gráfico da quantidade de fornos por grupo
plot(Fornos.som, type = "counts", main = "Quantidades por grupo")

#gera gráfico da qualidade de cada grupo
plot(Fornos.som, type = "quality", main = "Qualidade por grupo")
```

4.3. AGRUPAMENTO DE FORNOS USANDO *FUZZY C–MEANS* (FCM)

O método de clusterização *Fuzzy C–Means* foi desenvolvido por *Dunn* em 1973 e melhorado por Bezdek em 1981.

Em 1973, *Dunn* apresentou um método de agrupamento que combinavam os conceitos dos métodos baseados em função objetivo com os métodos da lógica *Fuzzy*, também conhecida como lógica difusa. Desta maneira, um padrão poderia ter certo grau de pertinência aos diferentes subgrupos resultantes, ao invés de simplesmente possuir uma pertinência discreta (0 ou 1).

Uma das tarefas da mineração de dados é a identificação de grupos ou *clusters* naturais nos conjuntos de dados. Em muitas situações cotidianas ocorrem casos em que um elemento está muito perto de dois *clusters* simultaneamente, de tal modo que se torna difícil agrupar esse elemento em um ou outro grupo. Isto ocorre devido à frequência relativa com que um conjunto de dados em particular apresenta características pertencentes a diferentes *clusters* e como consequência não é facilmente classificado. Para solucionar tais inconvenientes, foi desenvolvido um algoritmo conhecido como *Fuzzy C–Means* (FCM), o qual consiste em uma extensão difusa do conhecido *C–Means* e atribui a cada elemento um determinado valor, de acordo com o grau de pertinência e semelhança com as características do *cluster* em questão e, portanto, o mesmo elemento pode pertencer parcialmente a mais de um *cluster*.

O algoritmo *Fuzzy C–Means* (FCM) é o método mais utilizado para realizar agrupamento difuso. Ele permite encontrar um conjunto (C) de protótipos representativos de cada *cluster* e os graus de pertinência de cada dado. Outra importante função humana também presente na análise de dados é a seleção de atributos para as tarefas de agrupamento e classificação. Este algoritmo é uma técnica de mineração de dados que permite encontrar agrupamentos naturais em um conjunto de dados e pode ser aplicado em diversas áreas, como organização e classificação de dados, reconhecimento de padrões, estudo do clima, diagnóstico de doenças, bioinformática, genética (Seo et al., 2006), cancelamento de ruído e interferência de um sinal, estudo de séries temporais, estudo da rentabilidade econômica de uma empresa (Díaz e Morillas, 2004), suporte à decisão, segmentação de mercado e clientes (Weber, 2000), entre outras (Javier, 2003; Jantzen, 1998; Zha e Wei–Yip, 2005). Ele atribui a cada dado um grau de pertinência dentro de cada *cluster*, e como resultado, os dados podem pertencer parcialmente a mais de um grupo.

A aplicação do método difuso permite extrair algumas conclusões adicionais. Como é sabido, a lógica *Fuzzy* rompe com o princípio aristotélico do terceiro excluído, (em latim, *principium tertii exclusi ou tertium non datur*) é a terceira de três clássicas Leis do Pensamento. Ela afirma que para qualquer proposição, ou esta proposição é verdadeira, ou sua negação é verdadeira. Este princípio indica que um elemento pode ou não pertencer a um determinado conjunto, sendo vedada qualquer outra possibilidade. No entanto, ao trabalhar com a lógica *Fuzzy*, a mesma permite que um elemento pertença parcialmente a um determinado conjunto, trabalhando assim com graus de pertinência de diferentes elementos para diferentes conjuntos.

Ao contrário do algoritmo *C–Means* clássico que trabalha com uma partição dos dados, o FCM realiza uma partição suave do conjunto de dados, onde em tais partições os dados pertencem com um certo grau a todos os *clusters*.

Uma partição suave é formalmente definida como:

Seja X o conjunto de dados e x_i um elemento pertencente a X. Pode–se dizer que uma partição P = {C_1, C_2, ..., C_c} é uma partição suave de X se e somente se forem cumpridas as seguintes condições:

$$\forall x_i \in X \quad \forall C_j \in P \quad 0 \leq \mu c_j(x_i) \leq 1 \tag{6}$$

$$\forall x_i \in X \quad \exists C_j \in P \quad tal\ que\ \mu c_j(x_i) > 0 \tag{7}$$

onde C_j (x_i) μ denota o grau em que X_i pertence ao *cluster* C_j (John e Reza, 1999).

Um tipo de partição suave especial é aquela em que a soma dos graus de pertinência de um ponto específico em todos os *clusters* é igual a 1.

$$\sum_j \mu c_j(x_i) = 1 \quad \forall x_i \in X \tag{8}$$

Uma partição suave que atende a essa condição adicional é chamada de partição suave restrita. O algoritmo FCM produz uma partição suave restrita e para fazer isso, a função objetivo J se estende em duas maneiras: primeiro, na equação (9) são incorporados os graus de pertinência *Fuzzy* de cada dado em cada *cluster*; e em seguida, é introduzido um parâmetro adicional *m* que serve de peso expoente na função de pertinência. Assim, a função objetivo estendida J_m é disposta na seguinte equação:

$$J_m(P,V) = \sum_{i=1}^{k} \sum_{xk \in X} \left(\mu C_i(xk)\right)^m \|xk - v_i\|^2 \tag{9}$$

onde P é uma partição *Fuzzy* do conjunto de dados X formada por (C_1, C_2, ... C_k), o parâmetro *m* é um peso que determina o grau em que os membros parciais de um *cluster* afetam o resultado (John e Reza, 1999; George e Yuan, 1995).

Da mesma forma que o *C–Means* clássico, o FCM também tenta encontrar uma boa partição mediante a busca dos protótipos v_i que minimize a função objetivo J_m e, adicionalmente, o FCM também deve buscar as funções de pertinência μc_i que minimizem J_m. Estas condições são apresentadas no seguinte teorema que serve como a base do algoritmo FCM.

4.3.1. TEOREMA *FUZZY C–MEANS* (FCM)

Uma partição *Fuzzy* {C1, C2, ... Ck} pode ser um mínimo local da função objetivo J_m se e somente se estiverem de acordo com as seguintes condições (John e Reza, 1999; George e Yuan, 1995):

$$\mu c_i(x) = \frac{1}{\sum_{j=1}^{k} \left(\frac{\|x - v_i\|^2}{\|x - v_j\|^2} \right)^{\frac{1}{m-1}}} \quad 1 \le i \le k, x \in X \tag{10}$$

$$v_i = \frac{\sum_{x \in X} \left(\mu c_i(x)\right)^2 x}{\sum_{x \in X}^{n} \left(\mu c_i(x)\right)^m} \quad 1 \le i \le k \tag{11}$$

Depois de determinar o número de agrupamentos, o valor de *m* e o critério de parada, o algoritmo FCM executa duas etapas: primeiro calcula as funções de pertinência de acordo com a equação (10), e em seguida, atualiza os protótipos utilizando a equação (11). As duas etapas são repetidas iterativamente até alcançar o critério de parada.

4.3.2. VALIDAÇÃO DO *CLUSTER*

A validação de um algoritmo de agrupamento (*clustering*) é estimada por um critério objetivo para determinar quão boa é a partição gerada pelo algoritmo. Estes critérios são importantes porque permitem comparar os resultados de diversos algoritmos e permitem determinar o melhor número de *clusters*.

As medidas de validação de uma partição suave restrita são compostas de três categorias (John e Reza, 1999):

4.3.2.1. MEDIDAS BASEADAS NO GRAU PERTINÊNCIA

Estas medidas calculam certas propriedades das funções de pertinência em uma partição suave restrita. Uma destas medidas é o coeficiente de partição, introduzido por Jim Bezdek, no ano de 1973. Este coeficiente mede o grau de fuzzificação (*fuzziness*) do *cluster*. Este grau de fuzzificação indica o grau de pertinência do indivíduo aos *clusters*. Quanto mais difusos ou imprecisos são os *clusters*, pior é a partição.

Quando cada indivíduo possui um grau de pertinência igual a 1/c, onde c é o número total de *clusters*, se tem o que se chama de completa *fuzziness*, ou seja, cada indivíduo tem igual pertinência em todos os *clusters*. Por outro lado, quando cada indivíduo possui pertinência igual a 1 em algum *cluster*, se tem o chamado partição completa *hard*. O *Coefiente de Partição Dunn* mede o quão *hard* é uma partição fornecida por um algoritmo de clusterização *fuzzy*, o qual é definido de acordo com a seguinte fórmula (Rocha e Peres, 2012):

$$F_c = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^2}{n} \qquad (12)$$

onde u_{ik} representa o grau de pertinência do indivíduo *k* ao *cluster i*, *n* representa o número total de indivíduos no conjunto de dados e *c* representa o número total de *clusters*.

As medidas de validação baseadas no grau de pertinência, como o coeficiente de partição, podem ser utilizadas para comparar as partições alternativas com o mesmo número

de *clusters*, mas tem a limitação de melhorar à medida que se aumenta o número de *clusters*, menos ainda consideram propriedades geométricas da partição como a separação entre os *clusters*, por estas razões foram desenvolvidas as medidas geométricas (John e Reza, 1999).

4.3.2.2. MEDIDAS BASEADAS NA GEOMETRIA

Uma medida que considera os aspectos mais compactos e separados dos *clusters* com melhor partição foi introduzido por X.Xie e G.Beni, e sua fórmula segue a seguinte equação:

$$V_{XB} = \left(\frac{\sum \sigma^i}{n}\right)\frac{1}{d_{min}^2} \tag{13}$$

onde σ i é a variação do *cluster* c_i, definida como:

$$\sigma^i = \sum_j \mu c_i(x_j)\left\|x_j - v_i\right\|^2 \tag{14}$$

n é a cardinalidade do conjunto de dados e d_{min} é a menor distância entre os centros dos *clusters*, sendo definida como:

$$d_{min} = \min_{\substack{i \,.\, j \\ i \neq j}} \left\|x_j - v_i\right\| \tag{15}$$

O primeiro termo da equação 14 mede se um *cluster* não é compacto, e o segundo termo é uma medida da não–separação entre *clusters*. Por tanto, o produto dos dois reflete o grau em que os *clusters* em uma partição suave não são compactos e não estão bem separados (John e Reza, 1999).

4.3.2.3. MEDIDAS BASEADAS NO DESEMPENHO

Estas medidas avaliam uma partição com base em seu desempenho em relação a um objeto pré–definido, como o erro mínimo em uma prova de classificação.

4.3.3. TÉCNICA DE AGRUPAMENTO VIA *FUZZY C–MEANS* (FCM)

Outra técnica empregada para realizar o agrupamento dos fornos foi o *Fuzzy C–Means* (FCM). Os parâmetros para realizar os experimentos através do FCM foram:

- Matriz de dados;
- Número de centros (grupos);
- Cálculo para distância entre os pontos do grupo.

O algoritmo *Fuzzy C–Means* foi um dos primeiros algoritmos propostos para análise de agrupamento. Ele foi desenvolvido com o objetivo de solucionar o empecilho da incerteza e imprecisão, dado um elemento poder pertencer a vários conjuntos, também denominado *cluster*. Com isso, o algoritmo permite com que os dados pertençam parcialmente a dois ou mais *clusters*, de acordo com seu grau de pertinência (Rocha e Peres, 2012).

Para a realização do agrupamento dos dados, o algoritmo *Fuzzy C–Means* utiliza-se uma função de minimização das distâncias entre os dados e os centros dos grupos aos quais tais dados pertencem. Neste estudo, a medida de similaridade é sempre aplicada visando medir o quão similar é um vetor de dados e um vetor que representa uma classe ou grupo. Como dito anteriormente, a distância Euclidiana é a métrica escolhida para ser a base do cálculo de similaridade entre vetores, que é calculada como:

$$d = \left(\underrightarrow{v_i}, \underrightarrow{v_j}\right) = \left(\sum_{l=1}^{p} \left(v_{il} - v_{jl}\right)^2\right)^{\frac{1}{2}} \tag{16}$$

A função de minimização é dada por:

$$J_{CM}(U_h, C) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d\left(\underset{c_i}{\rightarrow},\underset{x_j}{\rightarrow}\right)^2 \quad (17)$$

onde $d\left(\underset{c_i}{\rightarrow},\underset{x_j}{\rightarrow}\right)$ é a distância entre o vetor de dados $\underset{x_j}{\rightarrow}$ e o protótipo do grupo $\underset{c_i}{\rightarrow}$, *c* é o número de grupos a ser determinado pelo algoritmo, *n* é o número de dados do conjuto e U_h é uma matriz binária, chamada de "matriz de partição", de dimensão *c* x *n*, definida como:

$$U_h = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ u_i + 1{,}1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ u_{c,1} & u_{c,2} & \cdots & u_{c,n} \end{bmatrix} \quad (18)$$

Um programa na linguagem R também foi desenvolvido para ler a matriz de dados de um arquivo CSV, realizar a normalização dos dados entre 0 e 1 e executar o algoritmo FCM para encontrar os grupos (*Clusters*) dos fornos de redução de alumínio, seguindo o código–fonte demonstrado abaixo:

Tabela 3 – Código–Fonte programado em R para agrupamento usando *Fuzzy C–Means* (FCM).

```
#carrega biblioteca para uso do FCM
library("e1071")
#carrega biblioteca para uso da função de normalização de dados
library("scales")
#importa dados de um arquivo CSV para o R
Fornos <– read.csv2(file='planilhaMédias_SEMFILTRO.csv');
#normaliza os dados entre 0 e 1 (função rescale) e prepara a matriz x somente com os dados
x <– rbind(rescale(Fornos$m_TMP, to=c(0, 1)), rescale(Fornos$m_ALF, to=c(0, 1)), rescale(Fornos$m_ALF3A, to=c(0, 1)), rescale(Fornos$m_QALr, to=c(0, 1)), rescale(Fornos$m_IncTM, to=c(0, 1)), rescale(Fornos$m_TUN, to=c(0, 1)), rescale(Fornos$m_TOV, to=c(0, 1)) )
#transpõe a matriz de dados
x <– t(x)
#executa o algoritmo FCM com os seguintes parâmetros:
#1) x: dados normalizados
#2) 3: número de grupos
#3) 500: número máximo de épocas/iterações
#4) verbose=TRUE: imprime informações enquanto realiza as operações do FCM
#5) euclidean: cálculo da distância entre pontos
#6) method="cmeans": método de clustering que equivalente ao Fuzzy C–Means
result<–cmeans(x, 3, 500, verbose=TRUE, dist="euclidean", method="cmeans")
#mostra o gráfico do agrupamento de cada instância comparando as variáveis por par
plot(Fornos, col=result$cluster)
#constrói um gráfico 3D que mostra o valor de pertinência para cada cluster
s3d <– scatterplot3d(result$membership, color=result$cluster, type="h", angle=55, scale.y=0.7, pch=16, main="Pertinence")
#exibe os resultados do agrupamento, inclusive o agrupamento resultante dos 960 fornos
print(result)
```

4.4. AGRUPAMENTO DE FORNOS VIA *K–MEANS*

O algoritmo *K–Means*, criado por MacQueen em 1967 é o algoritmo de agrupamento mais conhecido e utilizado por ser uma aplicação simples e eficaz.

A técnica *K–Means* também foi empregada para realizar o agrupamento dos fornos. O *K–Means* é um procedimento de classificação de um conjunto de objetos dentro de um determinado número *K* de *clusters*, sendo *K* já determinado a priori. Esta técnica aproxima por etapas sucessivas um certo número (pré–definido) de *clusters*, usando centróide dos pontos o qual deve representar. Além disso, processa os padrões sequencialmente (de modo que requerem armazenamento mínimo). No entanto, é influenciado pela ordem de apresentação dos padrões (os primeiros padrões determinam a conFiguração inicial dos agrupamentos) e o seu comportamento depende fortemente do parâmetro *K*.

O nome *K–Means* se dá devido a representação de cada um dos *K clusters* pela média (ou média ponderada) de seus pontos, ou seja, pelo seu centróide. A representação mediante centróides tem a vantagem de conter um significado gráfico e estatístico imediato. Por tanto, cada *cluster* é caracterizado por seu centro ou centróide localizado no centro ou no meio dos elementos que compõe o *cluster*.

A análise do algoritmo *K–Means* é um método de agrupamento de casos, com base nas distâncias existentes entre eles, de acordo com um conjunto de variáveis. As versões antigas do *K–Means* começaram com a atribuição dos *K* primeiros casos em relação aos centros dos *K* agrupamentos (os centros multivariados dos agrupamentos são chamados centróides). Na versão atual, o *K–Means* inicia selecionando os *K* casos mais distantes entre si (o usuário deve inicialmente determinar o número *K* de agrupamentos que deseja obter). E, em seguida, se inicia a leitura sequencial dos arquivos de dados, atribuindo cada caso ao centro mais próximo e atualizando o valor dos centros a medida que sejam incorporados novos casos. Uma vez que todos os casos sejam atribuídos a um dos *K* agrupamentos, se inicia um processo iterativo para calcular os centróides finais desses *K* agrupamentos.

A análise por agrupamento *K–Means* é especialmente útil quando se dispõe de um grande número de casos. Existe a possibilidade de utilizar a técnica de maneira exploratória, classificando os casos e realizando iterações para encontrar a localização do centróide, ou apenas como técnica de classificação, classificando os casos a partir de centróides conhecidos e fornecidos pelo usuário. Quando se utiliza como uma técnica exploratória, na maioria dos casos o usuário desconhece o número ideal de agrupamentos, sendo necessário repetir a

análise com diferentes números de agrupamentos e comparar as soluções obtidas; nestes casos também pode ser usado o método de análise de agrupamento hierárquico com uma sub–amostra de casos. Este algoritmo é um dos algoritmos mais simples e conhecidos de agrupamento, já que segue uma forma fácil e simples para dividir uma determinada base de dados em *K* grupos (fixados a priori). O algoritmo é baseado na minimização da distância interna (a soma das distâncias dos padrões atribuídos a um agrupamento ao centróide do referido agrupamento). Na verdade, este algoritmo minimiza a soma das distâncias ao quadrado de cada padrão em relação ao centróide de seu agrupamento.

A ideia principal é definir *K* centróides (um para cada grupo) e em seguida, pegar cada ponto da base de dados e situá–lo na classe de seu centróide mais próximo. O próximo passo é recalcular o centróide de cada grupo e redistribuir todos os objetos de acordo com o centróide mais próximo. O processo é repetido até que não haja alteração no grupo de um passo para o seguinte (Duda et al., 2001).

4.4.1. ETAPAS DO ALGORITMO *K–MEANS*

O algoritmo é realizado basicamente em quatro etapas:

- Etapa 1: Escolher aleatoriamente K objetos que formam os clusters iniciais. Estes pontos representam os centróides iniciais dos grupos. Para cada cluster K, o valor inicial do centro é igual a xi, sendo xi os únicos objetos do referido agrupamento pertencentes ao *cluster*.
- Etapa 2: Reatribuir objetos ao cluster. Para cada objeto x, o protótipo ao qual é atribuído a ele é exatamente o que estiver mais próximo do objeto, de acordo com uma medida de distância (geralmente é usada a medida euclidiana).
- Etapa 3: Uma vez que todos os objetos sejam agrupados, recalcula–se os centros dos K clusters (os centróides).
- Etapa 4: Repetir as etapas 2 e 3 até que não haja mais realocações e os centróides se mantenham estáveis. Isto produz uma classificação dos objetos em grupos, o qual permitem dar uma métrica entre eles. Embora o algoritmo sempre realize suas iterações até o fim, não há garantia de se obter a solução ótima. Com efeito, o algoritmo é muito sensível à escolha aleatória dos K centróides iniciais. A razão pela qual se utiliza o algoritmo K–Means inúmeras vezes sobre um mesmo conjunto de dados é para tentar minimizar este efeito supracitado e para que se obtenha melhores resultados, sabendo que os centros iniciais são os mais espaçados possíveis.

5. METODOLOGIA

5.1.SELEÇÃO E EXTRAÇÃO DOS DADOS

Como a produção de alumínio funciona 24 horas por dia, 7 dias na semana e 365 dias ao ano, ou seja, não para, faz–se necessário monitorar as variáveis que fazem parte do processo para controla–las, já que a produção de alumínio é um processo complexo e requer monitoramento contínuo. Este monitoramento é realizado em grande parte pela análise de dados extraídos por sensores. Estes dados são gravados em um banco de dados, o qual mantêm dados históricos de vários anos. A partir da Tabela 4 é possível verificar um resumo quantitativo dos dados por ano.

Tabela 4 – Registros por ano.

Ano	Quantidade
2006	348422
2007	348060
2008	348225
2009	346615
2010	347070
2011	348886
2012	216496

A média de cada variável foi extraída diretamente do banco de dados, excluindo valores nulos e menores ou igual a zero. Porém, no caso da temperatura, valores menores ou igual a 500 °C não foram considerados no cálculo da média desta variável, já que a grande maioria dos valores de temperatura registrados é muito superior a este patamar (ficando normalmente entre 950 e 970 °C, valores fora deste intervalo são considerados inválidos).

Exposto a estratégia de extração de dados, a matriz de dados a seguir (Tabela 5) foi construída e utilizada no software R para descobrir possíveis agrupamentos.

Tabela 5 – Matriz de dados resumida
(onde para cada forno foi calculada a 'média' de cada uma das variáveis a partir dos dados mencionados na Tabela 4. Cumpre informar que nesse trabalho também serão utilizados a 'mediana' e o 'desvio padrão').

Forno	m_TMP	m_ALF	m_ALF$_3$A	m_QALr	m_IncTM	m_TUN	m_TOV
FORNO1	964,3916	10,60201	37,52616	2391,912	0,141808	49,83600	52,41743
FORNO2	963,9601	10,73222	40,78312	2407,156	0,142151	50,06887	52,43101
FORNO3	964,1590	10,69377	42,81260	2378,871	0,158403	52,39175	52,31439
FORNO4	966,4861	10,30771	43,96218	2416,324	0,140115	49,11494	52,48892
FORNO5	965,0304	10,46070	42,99012	2379,370	0,137985	52,56182	52,50260
FORNO6	964,6234	10,69133	41,63504	2403,531	0,156319	48,92635	52,35775
FORNO7	966,5804	10,36257	42,45724	2378,455	0,137950	51,96111	52,35636
FORNO8	964,4563	10,38669	41,63597	2406,509	0,140249	50,10236	52,51386
FORNO9	961,7376	10,54372	37,88627	2423,455	0,168148	46,79282	52,59740
FORNO10	963,0284	10,62071	40,72819	2411,199	0,182923	49,46980	52,13562
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
FORNO958	964,4633	10,91633	34,51881	2473,864	0,174027	52,72807	51,92830
FORNO959	962,8184	10,60088	31,10854	2461,267	0,232770	55,76427	52,27551
FORNO960	963,2912	10,84042	30,15310	2465,113	0,225874	54,49614	51,74637

A partir desta grande quantidade de amostras armazenadas, dois conjuntos de dados diferentes foram criados:

- Sem filtro;
- Com filtro.

O conjunto de dados sem filtro leva em consideração somente as amostras que possuem valor maior que zero. Padrões com valores nulos não são considerados. A Tabela 6 exibe a quantidade de dados por ano. Verifica–se que este conjunto de dados é composto por mais de dois milhões de amostras no total.

Tabela 6 – Quantidade de registros por ano (sem filtro).

Variável	Quantidade
2006	348269
2007	347951
2008	348009
2009	346375
2010	346827

2011	348614
2012	216092

Outro conjunto de dados utilizado foi o “com filtro”. O conjunto de dados com filtro exclui os dados *outliers*, ou seja, registros nos quais ao menos uma variável se encontra fora da faixa normal de operação. Os dados contidos nele são aqueles que se encontram dentro das faixas de valores, para cada uma das sete variáveis, e estão demonstrados na Tabela 7. Os padrões que não estão dentro do intervalo especificado foram descartados.

Tabela 7 – Faixa de valores do conjunto “com filtro” por variável.

Variável	Faixa de Valor
TMP	Entre 940 e 990 ºC
% de ALF no Banho	Entre 2,65 e 18,5
ALF_3A	Entre 1 e 100
QALr	Entre 1820 e 3000
IncTM	Entre 0 e 1
%TUN	Entre 10 e 140
%TOV	Entre 20 e 80

Por intermédio da Tabela 8 é possível verificar a quantidade de dados por ano, levando em consideração a filtragem. Nota–se que este conjunto de dados possui mais de 340 mil registros no total.

Tabela 8 – Quantidade de registros por ano (com filtro).

Variável	Quantidade
2006	38096
2007	66711
2008	57603
2009	55642
2010	62995
2011	23539
2012	38096

Histogramas foram gerados para visualizar um sumário dos dados para cada variável com e sem filtro. Ao realizar as comparações entre os dois conjuntos de dados diferentes, verifica-se, que os dados com filtro possuem um intervalo de valores menor que os sem filtro, contribuindo para uma melhor distribuição de valores. Isto tende a facilitar o agrupamento dos dados, de acordo com o demonstrado na Figura 7 a seguir:

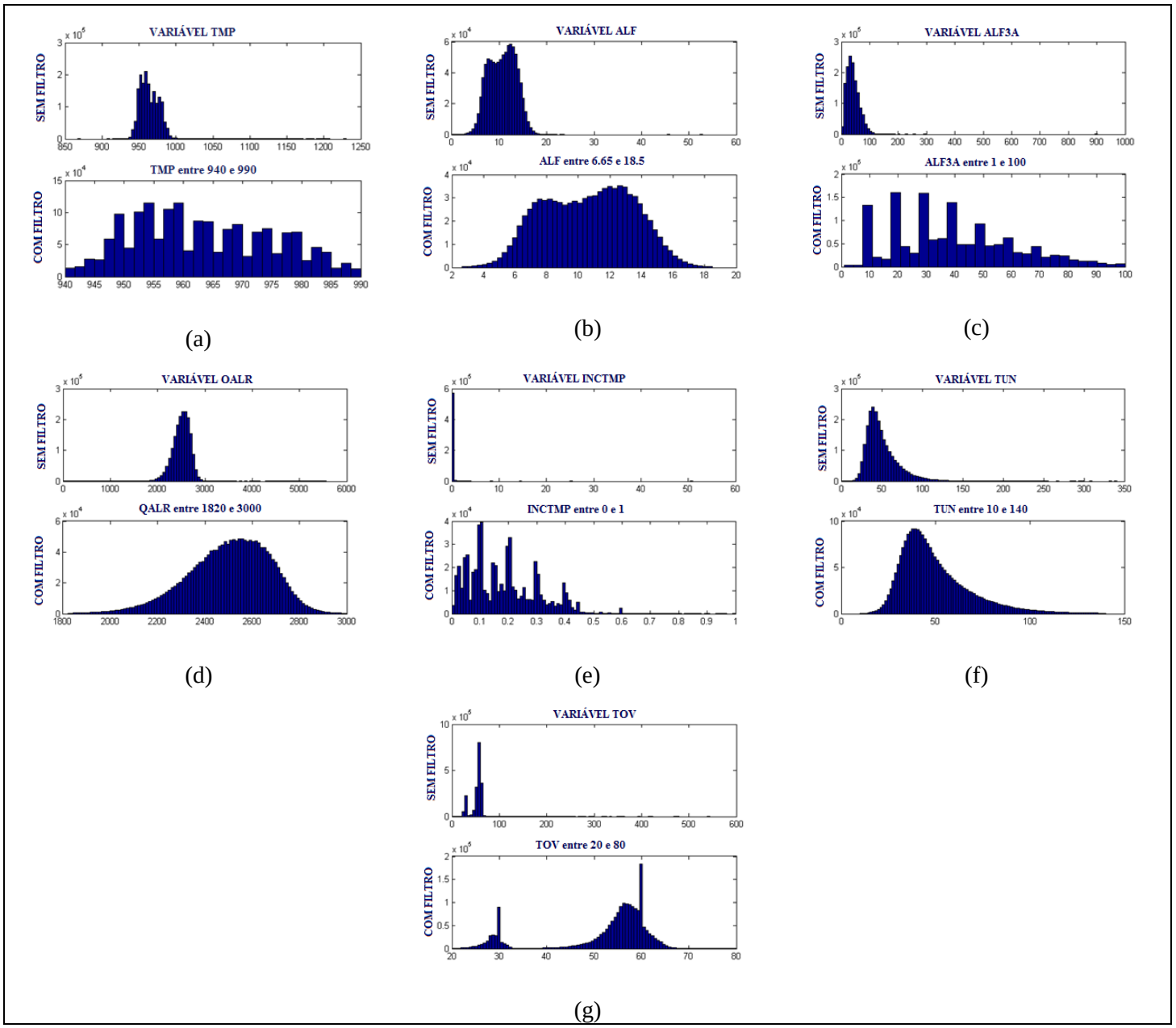


Figura 7 – Histogramas de cada variável conjunto de dados sem filtro e com filtro.
(a) Variável TMP. (b) Variável ALF. (c) Variável ALF_3A. (d) Variável QALr.
(e) Variável IncTM. (f) Variável %TUN. (g) Variável %TOV

Além de estabelecer dois tipos diferentes de conjunto de dados, três cálculos estatísticos foram utilizados (média, mediana e desvio padrão) para dados com filtro e sem filtro, resultando em seis combinações diferentes de experimentos elencados na Tabela 9. A média foi utilizada devido a sua influência nos dados espúrios. Por outro lado, mediana foi usada pela razão de ter pouca influência nos dados espúrios, e o desvio padrão foi utilizado por levar em consideração a média. Ressalta–se que todos os experimentos utilizaram distância Euclidiana.

Tabela 9 – Experimentos realizados.

Experimento	Cálculo Estatístico	Filtro
#1	Média	Com filtro
#2		Sem filtro
#3	Mediana	Com filtro
#4		Sem filtro
#5	Desvio Padrão	Com filtro
#6		Sem filtro

A Figura 8 representa a legenda das cores dos agrupamentos de 2, 5 e 13 *clusters*, a qual foi utilizada nas técnicas dos algoritmos *Affinity Propagation*, Kohonen (SOM), *Fuzzy C–Means* e *K–Means*.

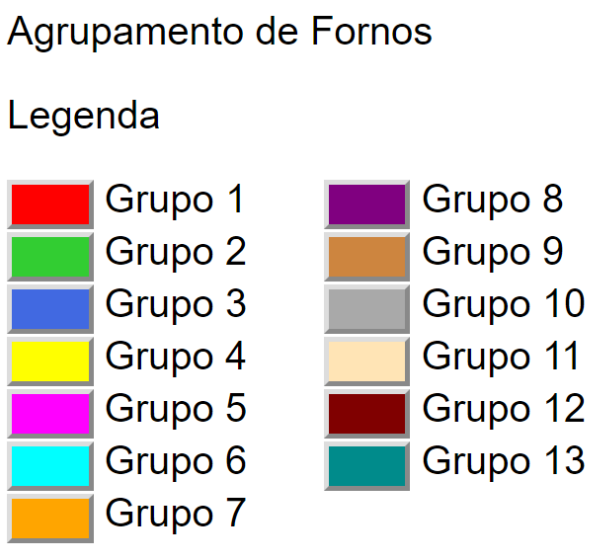


Figura 8 – Legenda das cores utilizadas nos agrupamentos dos fornos nas técnicas dos algoritmos *Affinity Propagation*, Kohonen SOM, *Fuzzy C–Means* e *K–Means*.

5.2.ESCOLHA DA MEDIDA DE ASSOCIAÇÃO

Para que seja possível a união de variáveis ou indivíduos, é necessário ter algumas medidas numéricas que caracterizem as relações entre eles. Em outras palavras, para agrupar indivíduos, é necessário a definição de uma medida de similaridade ou dissimilaridade. Com base nessa medida os objetos similares são agrupados e os demais são colocados em grupos separados (Aaker et al., 2001).

A medida de associação quantifica a relação entre uma dada exposição e uma consequência. Na prática, as medidas de dissimilaridade têm papel central nos algoritmos de agrupamentos. Através delas são definidos critérios para avaliar se dois pontos estão próximos, e, portanto, podem fazer parte de um mesmo grupo, ou não. Segundo Barroso & Artes (2003), há dois tipos de medidas de semelhança: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e medidas de dissimilaridade (quanto maior o valor, menor a semelhança entre os objetos). De um modo geral, as medidas de similaridade e de dissimilaridade são inter–relacionadas e, facilmente, transformáveis entre si (Bussab et al., 1990). Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Segundo Clifford & Stephenson (1975), tais coeficientes podem ser, facilmente, convertidos para coeficientes de dissimilaridade: se a similaridade for denominada *s*, a medida de dissimilaridade será o seu complementar (1 – s). A maioria dos métodos de análise de agrupamento requer uma medida de similaridade ou dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica (Doni, 2004).

Cada medida reflete a associação em um sentido particular e é necessário escolher uma medida adequada para o problema específico a ser tratado. A medida de associação pode ser uma distância ou uma similaridade.

Quando se escolhe uma distância como medida de associação (por exemplo, a distância Euclidiana) os grupos formados irão conter indivíduos semelhantes, de modo que a distância entre eles deve ser pequena.

Por outro lado, quando se escolhe uma medida de similaridade (por exemplo, o coeficiente de correlação) os grupos formados irão conter indivíduos com uma alta semelhança entre eles.

Vale ressaltar que neste trabalho foi utilizado como medida de associação a distância Euclidiana, para o agrupamento de 2, 5 e 13 *clusters* de acordo com as técnicas dos algoritmos de Kohonen, o SOM, do algoritmo *Fuzzy C–Means*, *K–Means*, e o algoritmo *Affinity Propagation*, que não considera o número de *clusters* a priori. Estes agrupamentos foram realizados com a finalidade de detectar o melhor número de *clusters* em cada grupo de acordo com as características de cada técnica empregada.

6. RESULTADOS

A preparação dos experimentos foi aplicada ao banco de dados obtidos de 2006 até 2012, de acordo com a Tabela 4, a qual foi ilustrada no capítulo 5 da metodologia.

O primeiro passo foi adquirir os dados, em seguida, foi realizado a filtragem desses dados de acordo com os histogramas ilustrados na Figura 7, também demonstrado no capítulo da metodologia, o qual serviu para eliminar os dados espúrios (*outliers*), ou seja, valores fora do padrão e que porventura pudessem vir a prejudicar os resultados do agrupamento.

O próximo passo foi calcular a média, a mediana e o desvio padrão de cada uma das sete variáveis de cada forno (Temperatura (TMP); Fluoreto de alumínio (% de ALF no Banho); Quantidade de fluoreto adicionado no banho (ALF_3A); Quantidade de alumina alimentada (QALr); Incremento de resistência por temperatura (IncTM); Percentual de tempo em alimentação *under–feeding* (%TUN); Percentual de tempo em alimentação *over–feeding* (%TOV)), utilizando os dados com filtro e os dados sem filtro.

A filtragem dos dados resulta em seis matrizes, sendo elas: média com filtro, média sem filtro, mediana com filtro, mediana sem filtro, desvio padrão com filtro e desvio padrão sem filtro. Cada matriz dispôs de 960 linhas, que representam a quantidade total de fornos, e 7 colunas, que representam as sete variáveis analisadas. Cada matriz (para a média, mediana, desvio padrão, com filtro e sem filtro) serviu como entrada do programa RStudio, realizando assim cada experimento, ou seja, executando o *script* desenvolvido em R com base nas planilhas calculadas com a média, mediana e desvio padrão e de acordo com a cada técnica envolvidas no trabalho para gerar os resultados.

A seguir têm-se os resultados dos experimentos dos algoritmos *Affinity Propagation*, Mapa Auto – Organizável de Kohonen – SOM, algoritmo *Fuzzy C-Means* – FCM e *K-Means*, para que seja visualizado seus comportamentos de forma individual e verificar o número ideal de *clusters* em cada técnica utilizada.

Uma vez realizado o experimento, o último passo é interpretar os resultados obtidos, os quais serão explicados a seguir.

6.1. *AFFINITY PROPAGATION*

Depois que o programa foi executado para cada um dos experimentos (cada uma das seis matrizes), seguindo os parâmetros dispostos na Tabela 9 de experimentos realizados, os gráficos gerados pelo programa podem ser vistos através das Figuras 9 (a) à 9 (f), os quais mostram a relação de agrupamento entre pares de variáveis. Nota–se que em todas as imagens há uma separação bem definida na quarta coluna e na quarta linha. Como a quarta variável é a QALr, então, na maioria dos casos, quando se compara outra variável do processo com esta, o agrupamento é mais preciso em relação às outras comparações entre variáveis. Este comportamento foi o mesmo observado em resultados semelhantes obtidos através de outras técnicas de agrupamento, como na técnica FCM, a qual será descrita na seção 6.3.

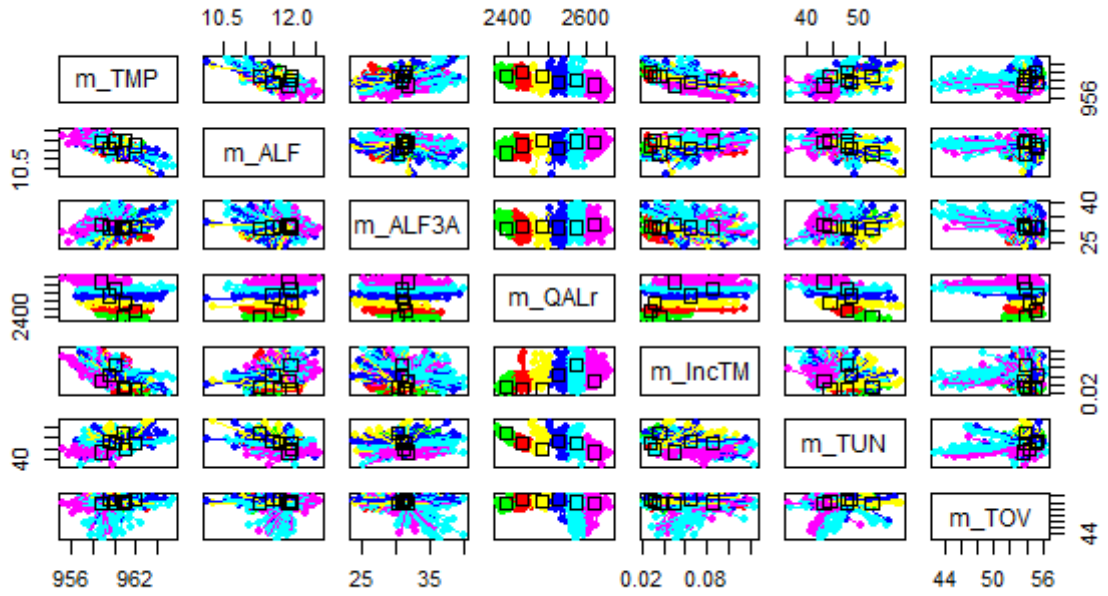


Figura 9 (a) – Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante com o uso do *Affinity Propagation*. Experimento #1: Média com filtro.

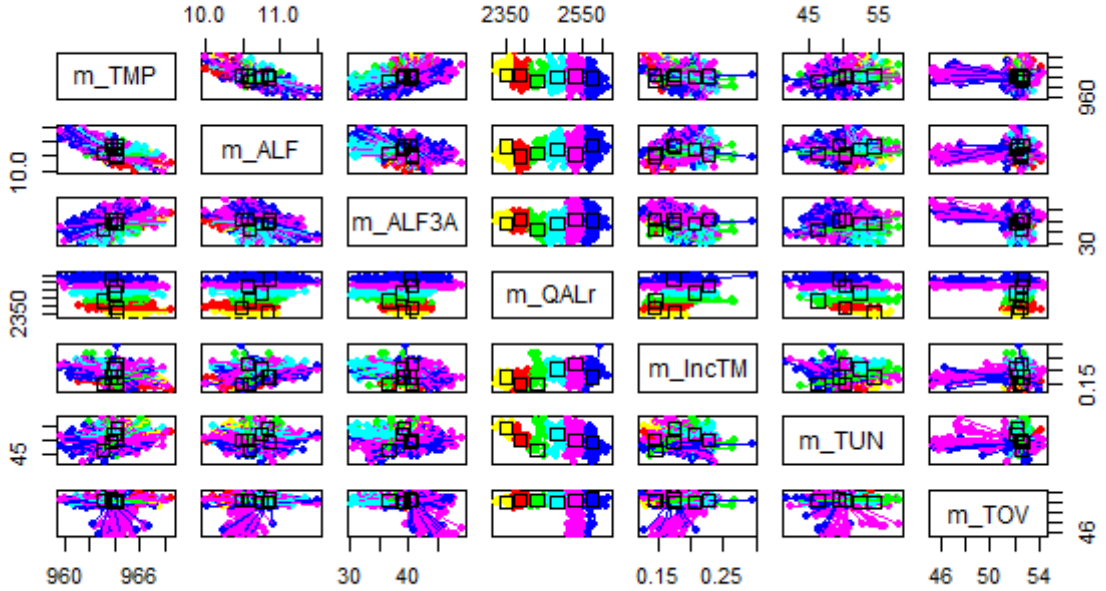


Figura 9 (b) Experimento #2: Média sem filtro.

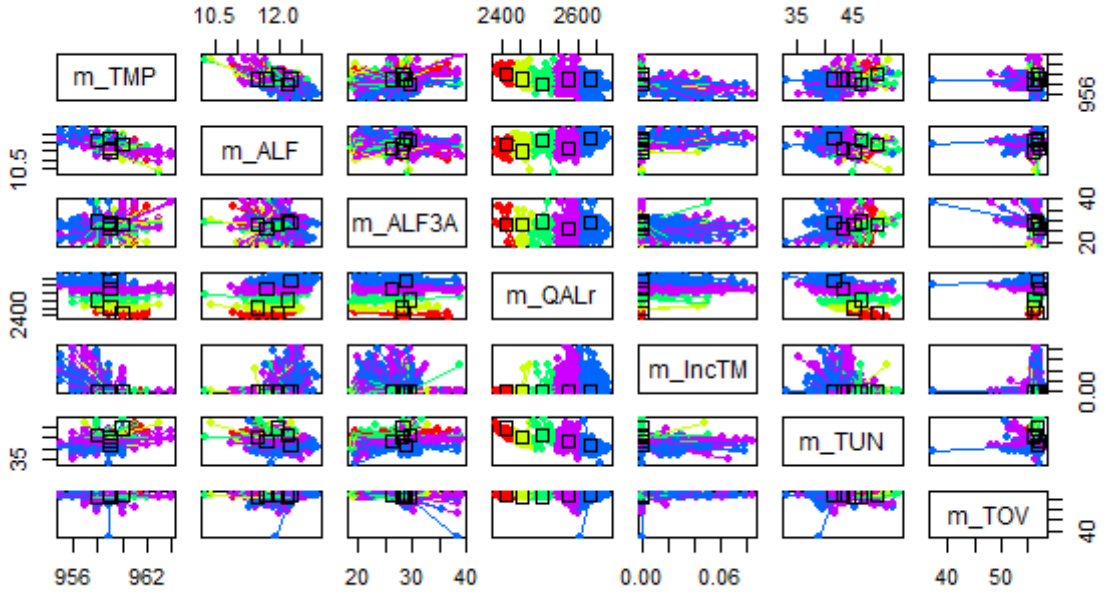


Figura 9 (c) Experimento #3: Mediana com filtro.

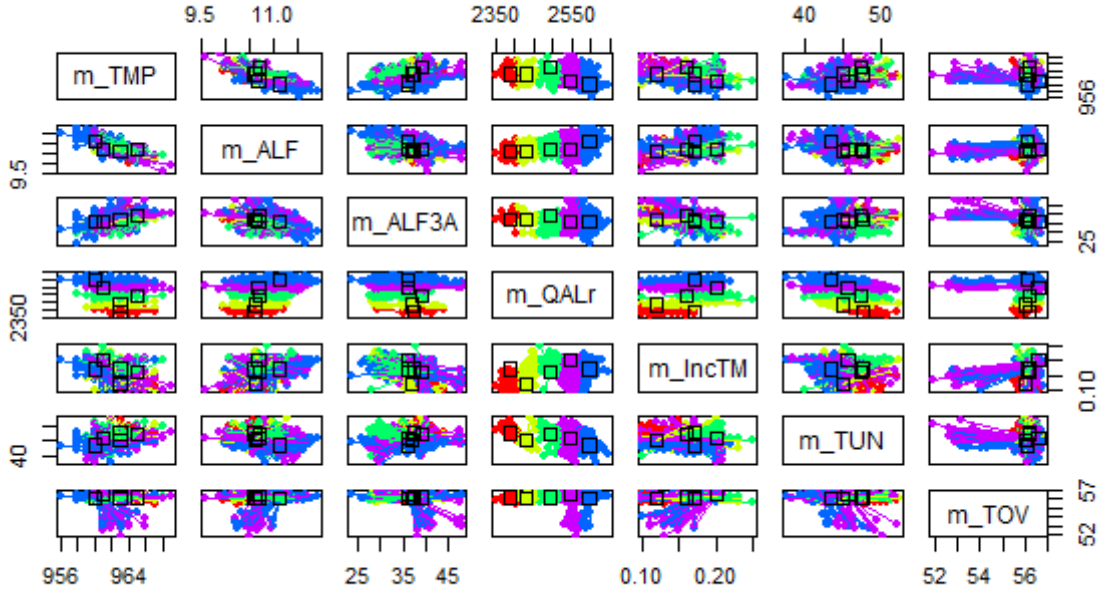


Figura 9 (d) Experimento #4: Mediana sem filtro.

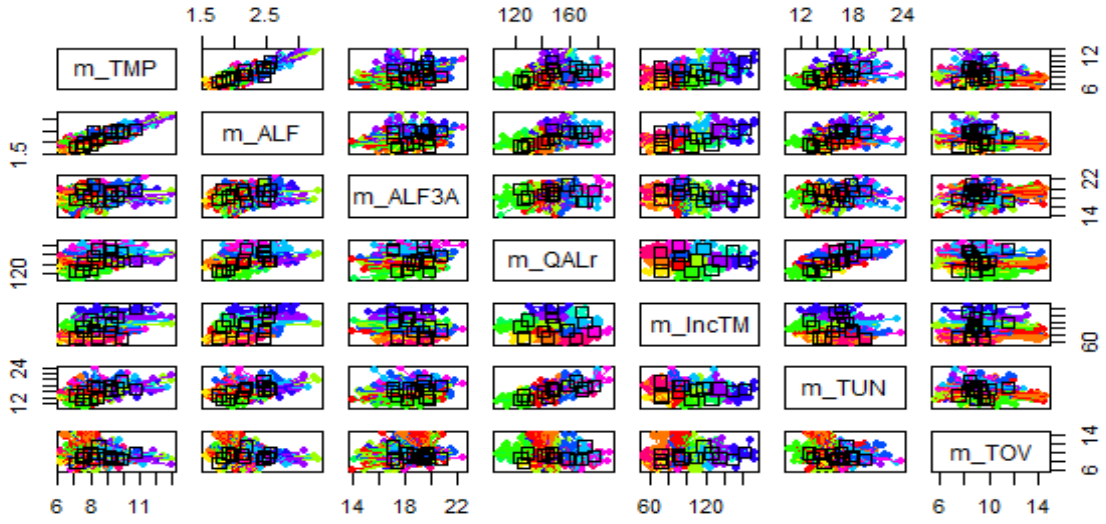


Figura 9 (e) Experimento #5: Desvio Padrão com filtro.

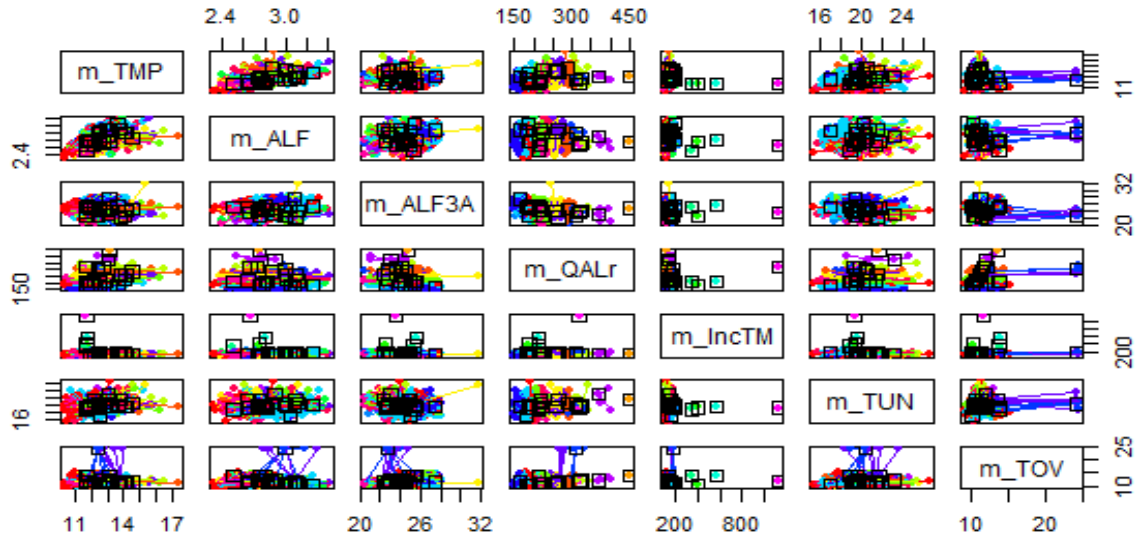


Figura 9 (f) Experimento #6: Desvio Padrão sem filtro.

Através da análise das Figuras 9 (a) à 9 (d), é possível inferir que o melhor agrupamento realizado é encontrado na variável central m_QAlr, a qual corresponde a quantidade de alumina alimentada, o que se remete uma forte influência desta variável em toda análise de agrupamento. Esta conclusão é possível ser visualizada nos experimentos com média e mediana com filtro e sem filtro, visto que os agrupamentos centrais, que equivalem exatamente aos agrupamentos onde a variável m_QAlr possui maior influência, possuem os grupos mais organizados e uniformes se comparados aos agrupamentos das outras variáveis, que por sua vez estão com os grupos misturados e de difícil compreensão. Já as Figuras 9 (e) e 9 (f) possuem todos os seus agrupamentos de forma misturada e de difícil separação. As Figuras 10 (a) à 10 (f) a seguir mostram a análise grupo *versus* localização física do forno conseguida pelo algoritmo *Affinity Propagation* com os agrupamentos para média com filtro e sem filtro, mediana com filtro e sem filtro e desvio padrão com filtro e sem filtro, calculando o número de *clusters* que cada medida aritmética resultou.

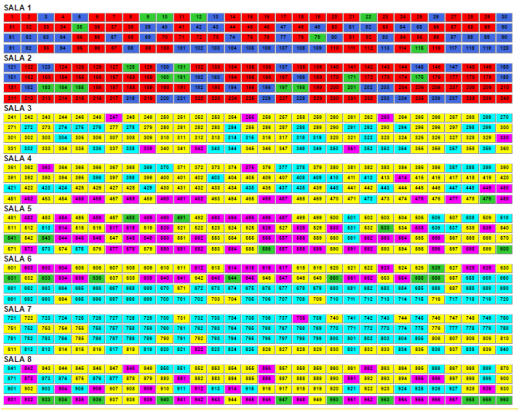


Figura 10 (a) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #1: Média com filtro.

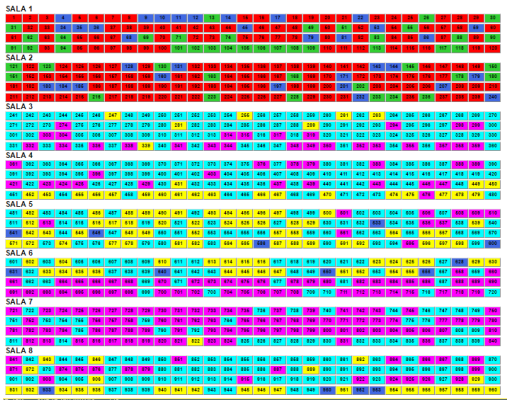


Figura 10 (b) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #2: Média sem filtro.

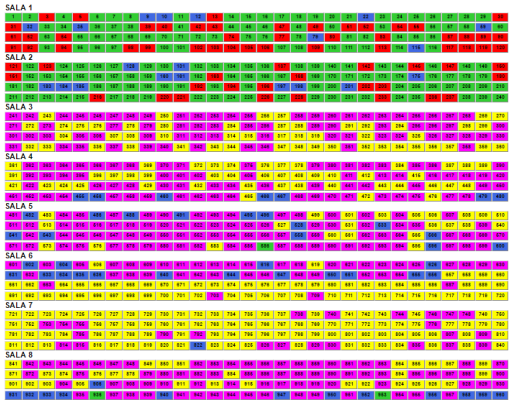


Figura 10 (c) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #3: Mediana com filtro.

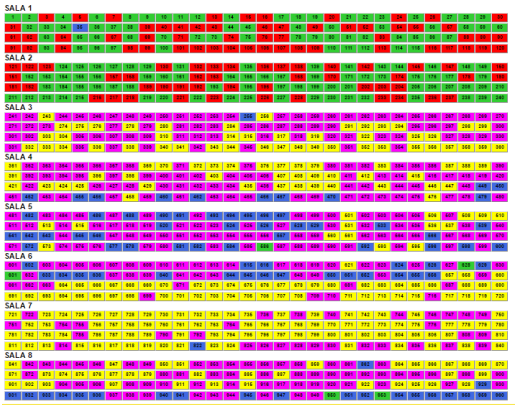


Figura 10 (d) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #4: Mediana sem filtro.

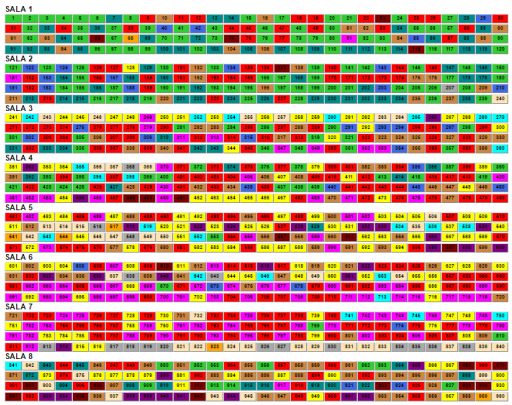


Figura 10 (e) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #5: Desvio Padrão com filtro.

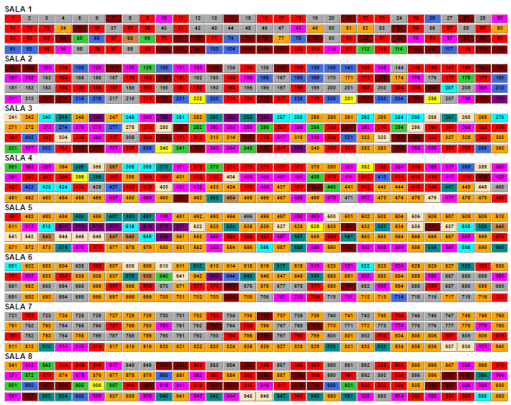


Figura 10 (f) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*. Experimento #6: Desvio Padrão sem filtro.

As cores apresentadas através da análise grupo *versus* localização das figuras 10 (a) até 10 (f) representam o número de clusters resultantes da técnica *Affinity Propagation*. Através da análise das Figuras 10 (a) à 10(f) é possível notar que a média com filtro e sem filtro resultou em seis *clusters*, enquanto que a mediana com filtro e sem filtro resultou em cinco *clusters*, o desvio padrão com filtro resultou em treze *clusters* e o desvio padrão sem filtro resultou em dezenove *clusters*.

Outra saída importante do programa é o número de grupos encontrados pela execução do algoritmo *Affinity Propagation* para cada experimento. A Tabela 10 resume estes valores e ainda mostra a iteração de convergência.

Tabela 10 – Número de grupos e iteração de convergência.

Experimento	Número de grupos	Iteração de convergência
#1	6	174
#2	6	204
#3	5	178
#4	5	167
#5	13	244
#6	19	241

Através da Tabela 10, verifica–se que para os experimentos #1 (média com filtro) e #2 (média sem filtro) o número de grupos calculado é seis com 174 e 204 iterações de convergência, respectivamente; já os experimentos #3 (mediana com filtro) e #4 (mediana sem filtro) o número de grupos encontrado é cinco com 178 e 167 iterações de convergência, respectivamente; e para os experimentos #5 (desvio padrão com filtro) e #6 (desvio padrão sem filtro) o número de *clusters* é maior que 10 com 244 e 241 iterações de convergência, respectivamente. No caso do último experimento (#6), há diversos exemplares (grupos) compostos por uma quantidade baixa de fornos, os quais podem ser mesclados com outro grupo análogo, o que reduziria o número de *clusters*.

6.2.SOM

A topologia da rede neural utilizada pode ser vista por intermédio da Figura 11. Nota–se que há sete neurônios na camada de entrada, onde cada neurônio representa uma variável do processo. Além disso, é possível verificar que há nove neurônios na camada de saída, os quais representam os grupos de fornos.

Foram feitos experimentos de 2 a 13 *clusters* (que implica na variação no número de neurônios na saída do SOM de 2 a 13), usando a média, mediana e desvio padrão para o estudo geral das clusterizações e o melhor agrupamento ocorreu com 9 *clusters* conforme demonstrado na Figura 11 abaixo, a qual representa a topologia da Rede Neural contendo 9 *clusters* com as sete variáveis envolvidas no processo:

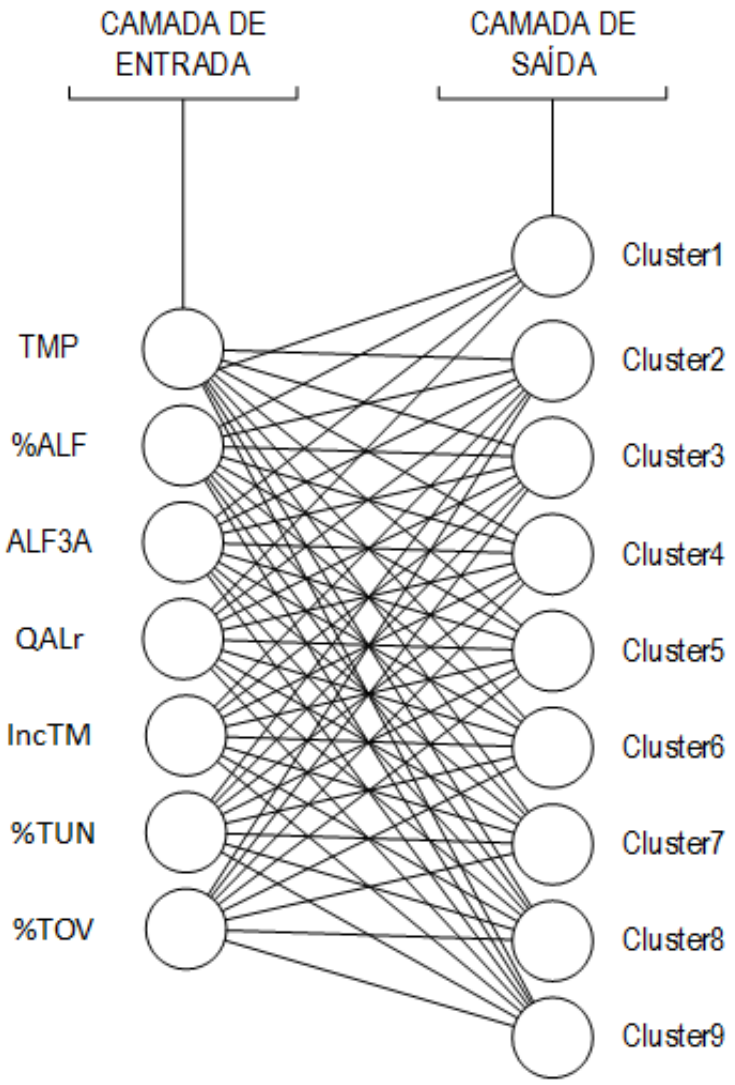


Figura 11 – Topologia da Rede Neural.

O gráfico representado pela Figura 12 a seguir, mostra a relação entre a distância média para o grupo mais próximo, à medida que as iterações acontecem. Verifica–se que após a iteração 200, as distâncias permanecem no mesmo patamar (abaixo de 0,02) até a última iteração, demonstrando uma distância média pequena.

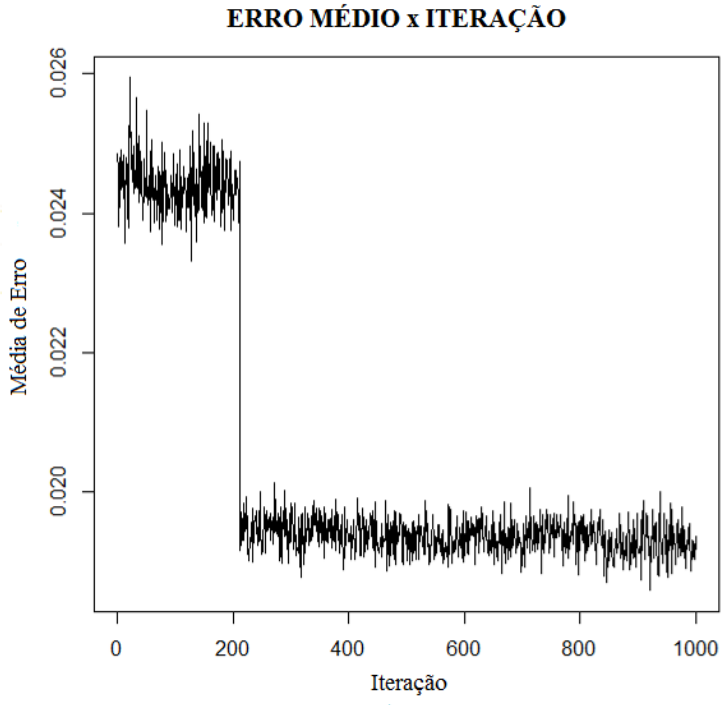


Figura 12 – Evolução do Erro Médio Quadrático por Iteração.

O principal resultado a ser analisado é o gráfico que mostra os grupos encontrados pelo mapa de Kohonen (Figura 13). Cada círculo representa um grupo de fornos e, dentro deles, há um gráfico de pizza que mostra o nível de determinada variável dentro do *cluster*. No primeiro grupo, é possível verificar que existe grande influência das variáveis de cor amarela, laranja e verde escuro (ALF_3A, QALr, TMP); pequena influência das variáveis representadas pela cor verde claro, laranja claro e rosa (ALF, IncTM, TUN) e ausência da variável de cor cinza (TOV).

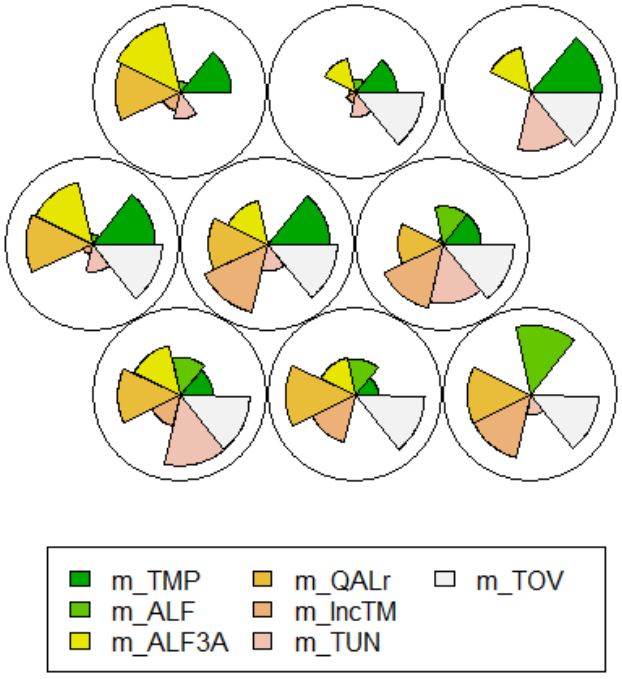


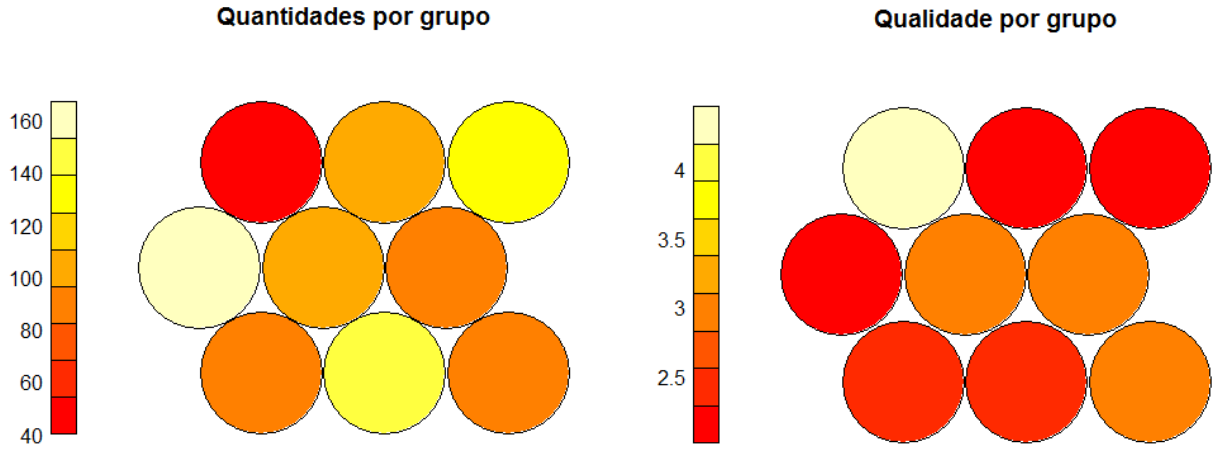
Figura 13 – Grupos de fornos e respectivos níveis de influência.

A análise da Figura 13 foi feita de forma visual, onde cada círculo representa um grupo ou *cluster* e as fatias representam a quantidade de cada variável existente em cada um dos *clusters*. A partir da análise feita para o primeiro grupo, foi decidido usar o valor "Alto" quando o tamanho da fatia é grande e acima do valor médio; valor "Baixo" quando o tamanho é pequeno e abaixo do valor médio e valor "Inexistente" quando a fatia não aparece no gráfico. Após a análise dos resultados, a Tabela 11 foi gerada, como mostrado abaixo:

Tabela 11 – Sumário dos níveis de influência de cada variável nos *clusters* de fornos.

Grupo	TMP	ALF	ALF3A	QALr	IncTM	TUN	TOV
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Inexistente
2	Alto	Baixo	Baixo	Inexistente	Baixo	Baixo	Alto
3	Alto	Inexistente	Alto	Inexistente	Inexistente	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto
5	Alto	Inexistente	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Inexistente	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Inexistente	Alto
9	Inexistente	Alto	Inexistente	Alto	Alto	Baixo	Alto

A quantidade de fornos por grupo e a qualidade do grupo podem ser analisadas através das Figuras 14(a) e 14(b), respectivamente. Em relação às quantidades, quanto mais escuro é o grupo, menos fornos estão contidos nele. Por outro lado, sobre a qualidade, quanto mais escuro é o grupo, melhor é a qualidade do mesmo, já que a distância média entre seus membros é pequena.



(a) (b)

Figura 14 – (a) Quantidade de fornos por *cluster* (b) Qualidade de cada *cluster*.

A Tabela 12 abaixo, mostra a relação entre as quantidades, as qualidades e as características dos grupos. O grupo vermelho possui de forma estimativa valores entre 40 e 60 fornos, o grupo laranja possui entre 80 e 100 fornos, o grupo amarelo possui uma estimativa de 120 a 140 fornos e o grupo bege possui uma faixa de 160 fornos. No que diz respeito às quantidades, é possível conferir que o grupo 1 é o que possui menos fornos; o grupo 4 possui o maior número de fornos se comparado a todos os grupos e é exatamente o grupo que não possui nenhuma característica inexistente; no entanto, apesar de possuir algumas características inexistentes, os grupos 3 e 8 detêm muitos fornos também. Sobre as qualidades, se verifica que o grupo 1, além de possuir menos fornos, é o que tem a pior qualidade e que os grupos 2, 3, 4, 7 e 8 são de alta qualidade, sendo que todos os mais numerosos estão entre os melhores, pois estão mais próximos um do outro e possuem a menor distância de um para o outro.

Tabela 12 – Relação entre quantidades, qualidades e características dos grupos.

QUANTIDADES							
Grupo	**TMP**	**ALF**	**ALF3A**	**QALr**	**IncTM**	**TUN**	**TOV**
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Inexistente
2	Alto	Baixo	Baixo	Inexistente	Baixo	Baixo	Alto
3	Alto	Inexistente	Alto	Inexistente	Inexistente	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto
5	Alto	Inexistente	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Inexistente	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Inexistente	Alto
9	Inexistente	Alto	Inexistente	Alto	Alto	Baixo	Alto
QUALIDADES							
Grupo	**TMP**	**ALF**	**ALF3A**	**QALr**	**IncTM**	**TUN**	**TOV**
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Inexistente
2	Alto	Baixo	Baixo	Inexistente	Baixo	Baixo	Alto
3	Alto	Inexistente	Alto	Inexistente	Inexistente	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto
5	Alto	Inexistente	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Inexistente	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Inexistente	Alto
9	Inexistente	Alto	Inexistente	Alto	Alto	Baixo	Alto

Outras interpretações importantes podem ser encontradas abaixo:

1) **TOV** aparece alto em todos os grupos, exceto no grupo 1, onde aparece inexistente;
2) **QALr** aparece alto em todos os grupos, exceto nos grupos 2 e 3, nos quais aparece inexistente;
3) **ALF** aparece baixo ou inexistente em todos os grupos, exceto no grupo 9;
4) **TMP** aparece inexistente somente no grupo 9;
5) **IncTM** aparece inexistente somente no grupo 3;
6) **TUN** aparece inexistente somente no grupo 8;
7) O grupo com mais variáveis inexistentes é o 3 (variáveis: **ALF**, **QALr**, **IncTM**);
8) Quando **TMP** é alto, **ALF** é baixo ou inexistente;
9) Quando **TMP** é baixo, **ALF** é baixo;
10) Quando **TMP** é inexistente, **ALF** é alto;
11) O grupo 1 é o menor (mais escuro) e possui menor qualidade (mais claro);
12) Os maiores grupos (3, 4, 8) são aqueles que tem, também, maior qualidade.

As regras analisadas acima foram adquiridas através dos dados das técnicas de mineração de dados do algoritmo Kohonen, e as mesmas podem ser utilizadas para explicar parcialmente o funcionamento do forno, e podendo também servir para treinar novos engenheiros, bem como pessoas envolvidas no processo do funcionamento do forno em questão, para que ao invés dos trabalhadores aprenderem a manusear o forno presencialmente, os mesmos podem ter o primeiro contato de forma virtual, através de um programa computacional, no qual sejam feitas várias simulações e testes e assim não aja nenhuma danificação nos fornos.

As Figuras 15 (a) à 15 (f) trazem a análise da qualidade por grupo através de cores diferentes para cada agrupamento através do algoritmo Kohonen, para o experimento com dados filtrados e não filtrados, com medidas de média, mediana e desvio padrão para 2 *clusters*, as Figuras 16 (a) à 16 (f) demonstram a mesma análise para 5 *clusters* e as Figuras 17 (a) à 17 (f) trazem esta análise para 13 *clusters*.

Qualidade por grupo Kohonen (SOM) utilizando 2 *Clusters*:

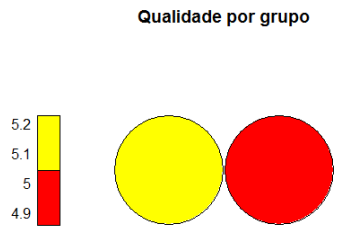


Figura 15 (a) – Gráficos de qualidade por grupo do algoritmo Kohonen (SOM) para 2 *Clusters*. Experimento #1: Média com filtro.

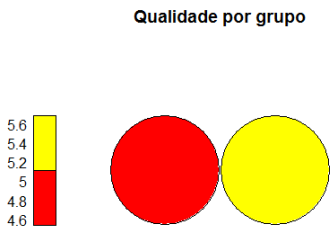


Figura 15 (b) Experimento #2: Média sem filtro.

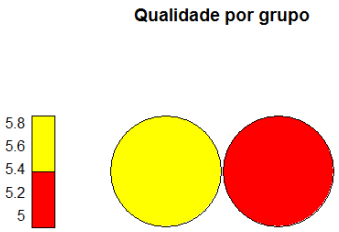


Figura 15 (c) Experimento #3: Mediana com filtro.

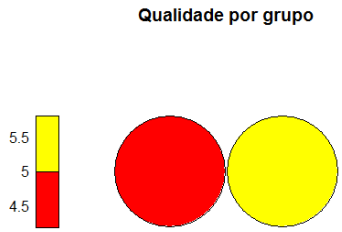


Figura 15 (d) Experimento #4: Mediana sem filtro.

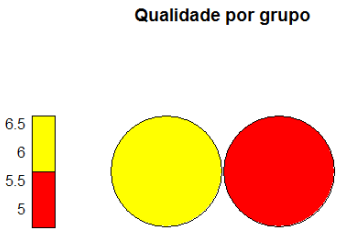


Figura 15 (e) Experimento #5: Desvio Padrão com filtro.

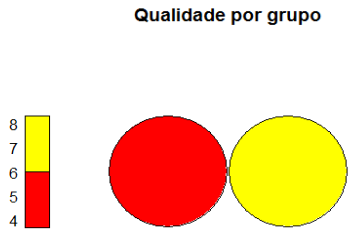


Figura 15 (f) Experimento #6: Desvio Padrão sem filtro.

Qualidade por grupo Kohonen (SOM) utilizando 5 *Clusters*:

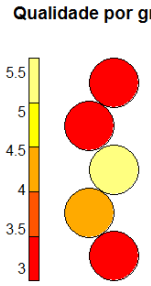


Figura 16 (a) – Gráficos de qualidade por grupo do algoritmo Kohonen (SOM) para 5 *Clusters*. Experimento #1: Média com filtro.

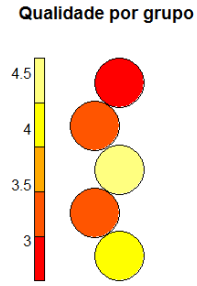


Figura 16 (b) Experimento #2: Média sem filtro.

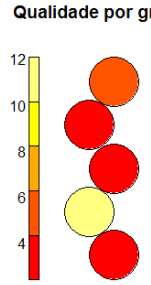


Figura 16 (c) Experimento #3: Mediana com filtro.

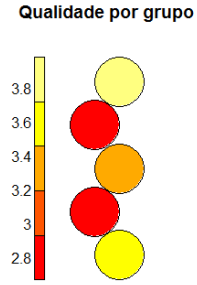


Figura 16 (d) Experimento #4: Mediana sem filtro.

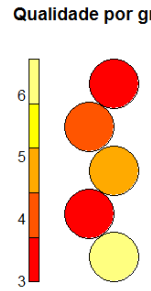


Figura 16 (e) Experimento #5: Desvio Padrão com filtro.

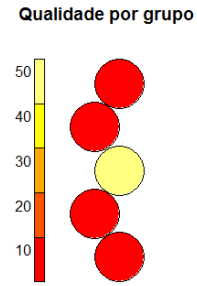


Figura 16 (f) Experimento #6: Desvio Padrão sem filtro.

Qualidade por grupo Kohonen (SOM) utilizando 13 *Clusters*:

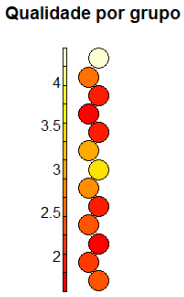


Figura 17 (a) Gráficos de qualidade por grupo do algoritmo Kohonen (SOM) para 13 *Clusters*.

Experimento #1: Média com filtro.

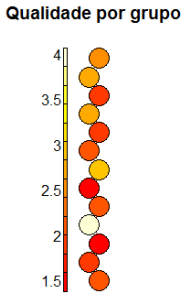


Figura 17 (b) Experimento #2: Média sem filtro.

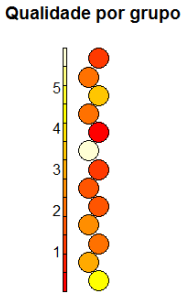


Figura 17 (c) Experimento #3: Mediana com filtro.

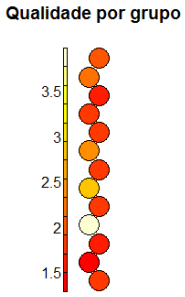


Figura 17 (d) Experimento #4: Mediana sem filtro.

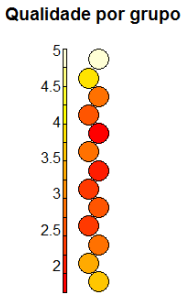


Figura 17 (e) Experimento #5: Desvio Padrão com filtro.

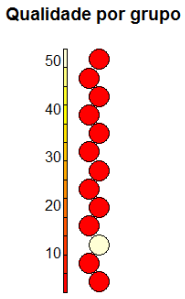


Figura 17 (f) Experimento #6: Desvio Padrão sem filtro.

As Figuras 18 (a) a 20 (f) representam o mapeamento do grupo *versus* localização do algoritmo Kohonen (SOM) para 2, 5 e 13 *clusters*, utilizando média com filtro e sem filtro, mediana com filtro e sem filtro e desvio padrão com filtro e sem filtro. Estas medidas são utilizadas para analisar o comportamento dos agrupamentos e verificar a melhor clusterização.

Mapeamento Kohonen (SOM) com 2 *Clusters*:

Figura 18 (a) Gráficos de mapeamento grupo *versus* localização física do forno conseguida pelo algoritmo Kohonen (SOM) para 2 *Clusters*. Experimento #1: Média com filtro.

Figura 18 (b) Experimento #2: Média sem filtro.

Figura 18 (c) Experimento #3: Mediana com filtro.

Figura 18 (d) Experimento #4: Mediana sem filtro.

Figura 18 (e) Experimento #5: Desvio Padrão com filtro.

Figura 18 (f) Experimento #6: Desvio Padrão sem filtro.

Mapeamento Kohonen (SOM) com 5 *Clusters*:

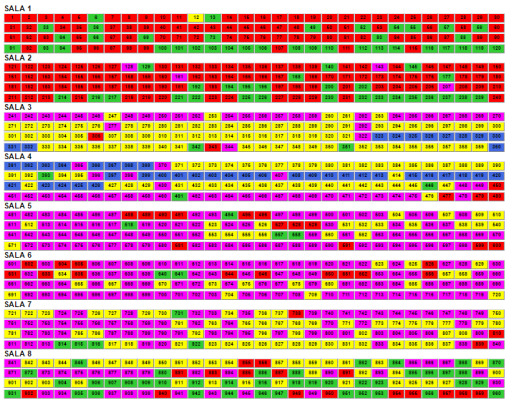


Figura 19 (a) Gráficos de mapeamento grupo *versus* localização física do forno conseguida pelo algoritmo Kohonen (SOM) para 5 *Clusters*. Experimento #1: Média com filtro.

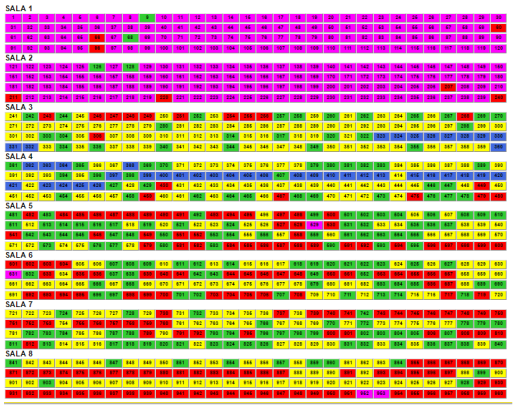


Figura 19 (b) Experimento #2: Média sem filtro.

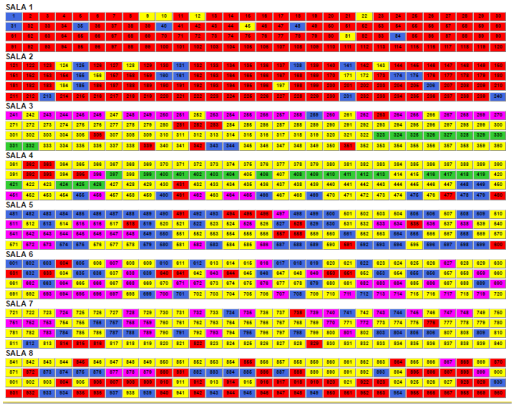


Figura 19 (c) Experimento #3: Mediana com filtro.

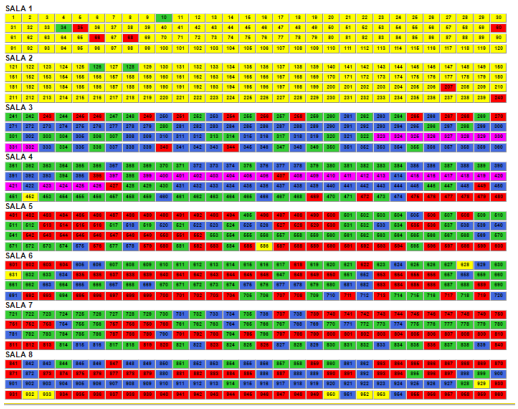


Figura 19 (d) Experimento #4: Mediana sem filtro.

Figura 19 (e) Experimento #5: Desvio Padrão com filtro.

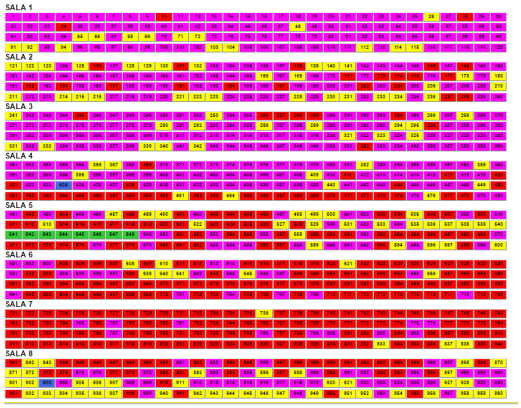


Figura 19 (f) Desvio Padrão sem filtro.

Mapeamento Kohonen (SOM) com 13 *Clusters*:

Figura 20 (a) Gráficos de mapeamento grupo *versus* localização física do forno conseguida pelo algoritmo Kohonen (SOM) para 13 *Clusters*. Experimento #1: Média com filtro.

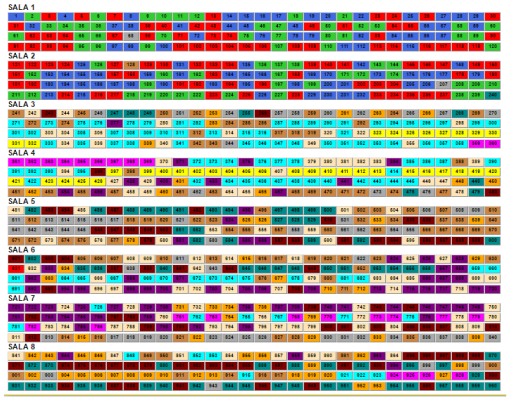


Figura 20 (b) Experimento #2: Média sem filtro.

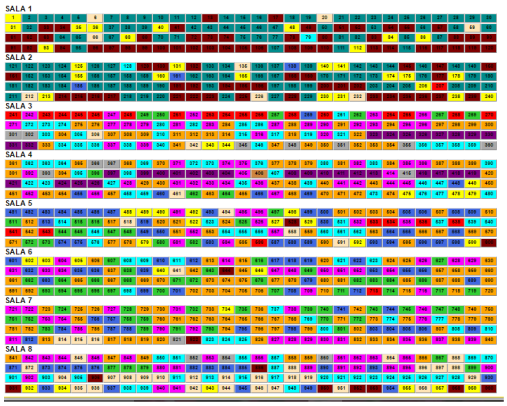


Figura 20 (c) Experimento #3: Mediana com filtro.

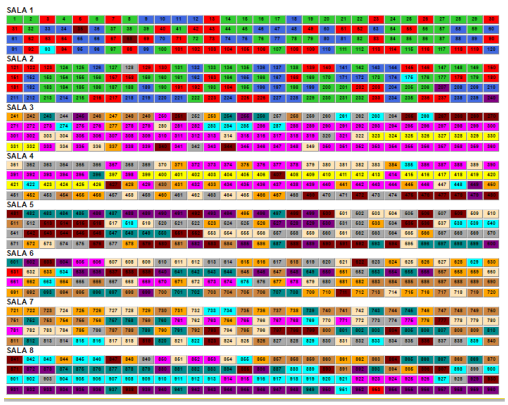


Figura 20 (d) Experimento #4: Mediana sem filtro.

Figura 20 (e) Experimento #5: Desvio Padrão com filtro.

Figura 20 (f) Experimento #6: Desvio Padrão sem filtro.

Após a análise das Figuras 18 (a) à 20 (f) é possível verificar que a melhor medida para o uso com o mapa auto-organizável de Kohonen se deu através da mediana com filtro contendo 5 *clusters*, pois demonstra uma predominância de um único grupo, o grupo vermelho, nas duas primeiras salas; e mediana sem filtro contendo 5 *clusters*, que também demonstra uma predominância de um único grupo nas duas primeiras salas, o grupo amarelo.

6.3. FCM

Depois que o programa foi executado para cada um dos experimentos de acordo com a Tabela 9 mostrada na metodologia, o gráfico gerado pelo programa pode ser visto através das Figuras 21 (a) à 23 (f), as quais mostram a relação de agrupamento entre pares de variáveis para 2, 5 e 13 *clusters*, respectivamente.

As Figuras 21 (a) à 21 (f) trazem a análise par a par por variável de cada agrupamento através do algoritmo *Fuzzy C–Means*, para o experimento com dados filtrados e não filtrados, com medidas de média, mediana e desvio padrão para 2 *clusters*, as Figuras 22 (a) à 22 (f) demonstram a mesma análise para 5 *clusters* e as Figuras 23 (a) à 23 (f) originam a análise para 13 *clusters*. Assim como no algoritmo Kohonen, é possível notar que em todas as imagens há uma separação bem definida na quarta coluna e na quarta linha, a qual representa a variável QALr, que é a quantidade de alumina alimentada.

Agrupamento Par a Par por variável no FCM com 2 *Clusters*:

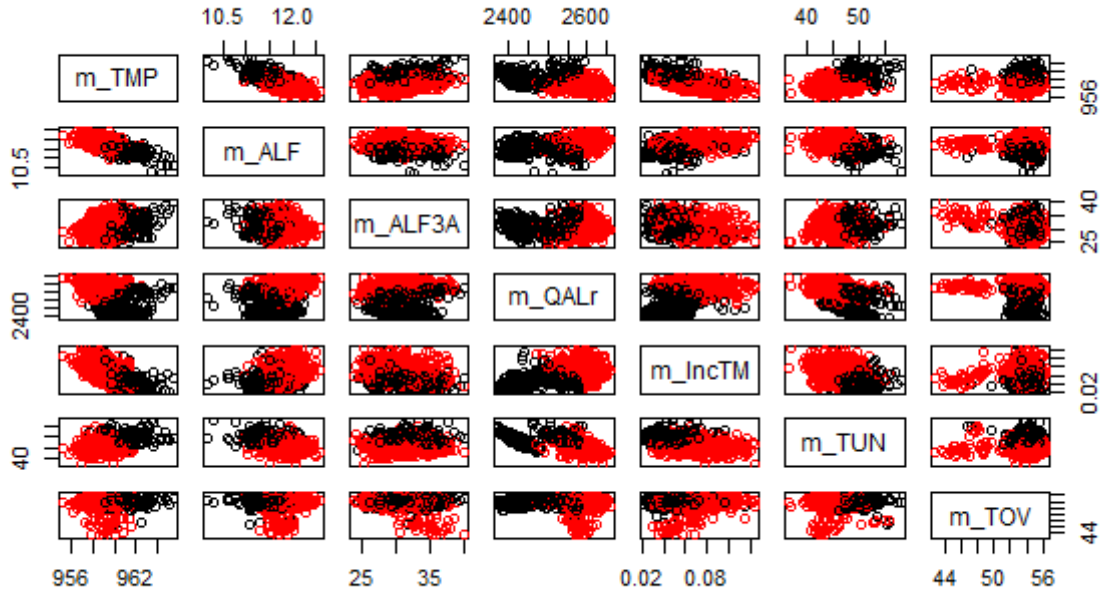


Figura 21 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante para 2 *Clusters*. Experimento #1: Média com filtro.

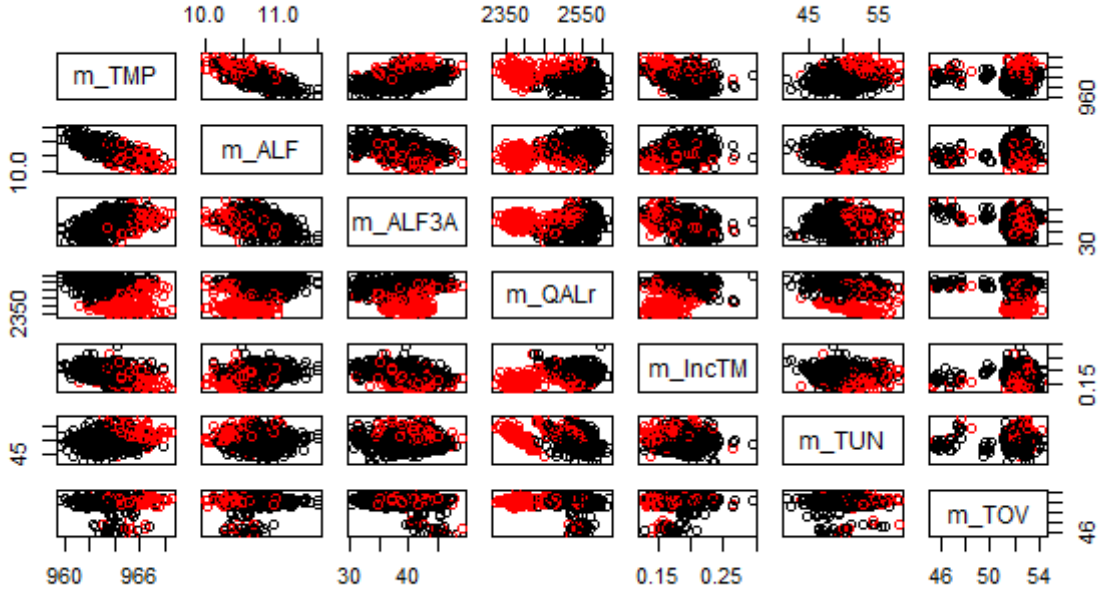


Figura 21 (b) Experimento #2: Média sem filtro.

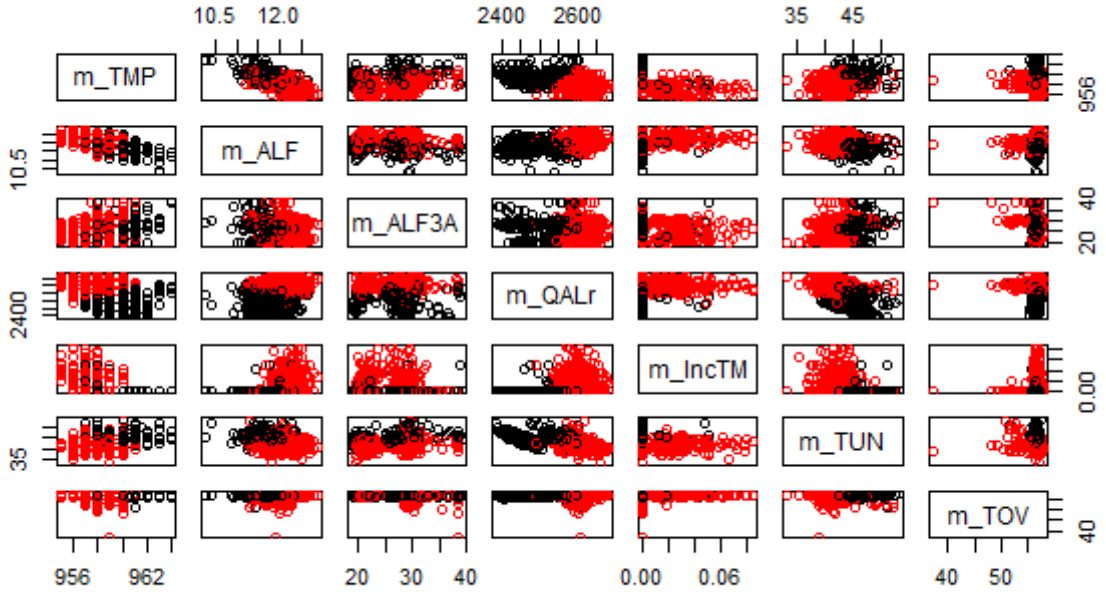


Figura 21 (c) Experimento #3: Mediana com filtro.

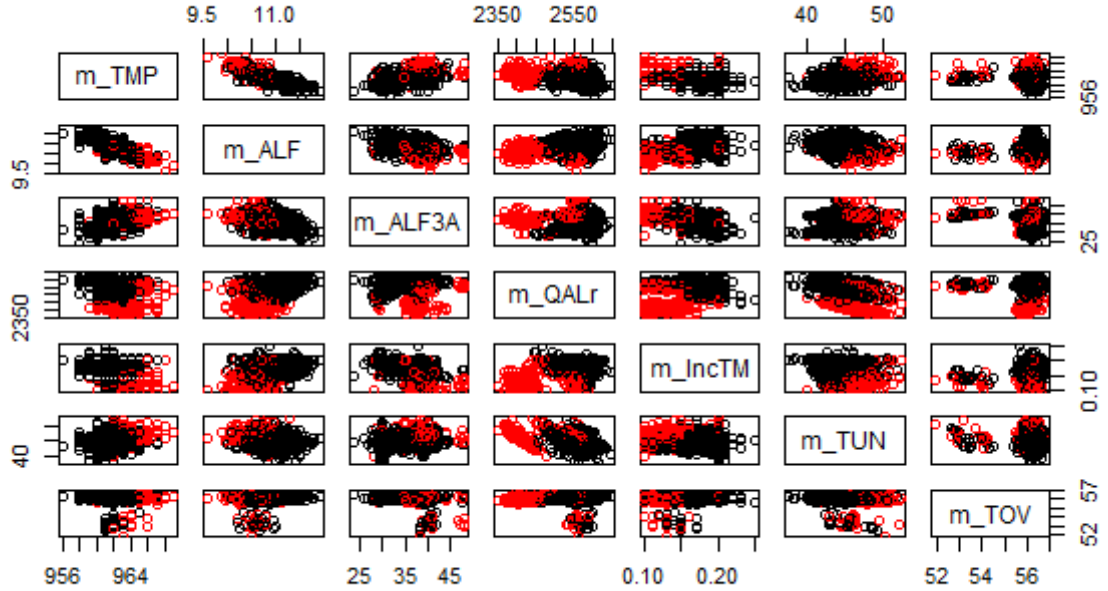


Figura 21 (d) Experimento #4: Mediana sem filtro.

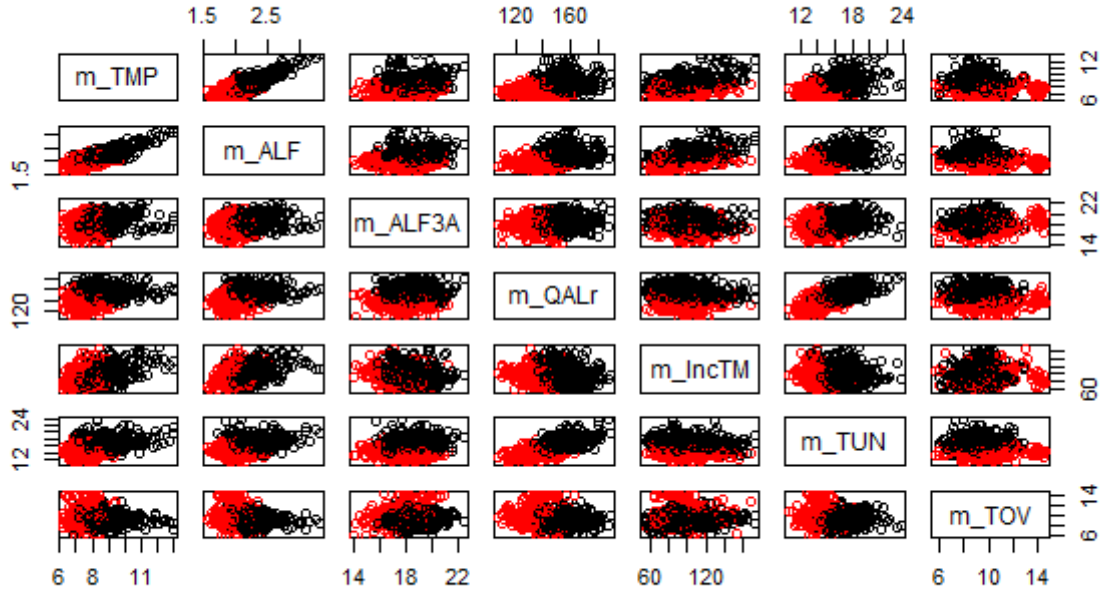


Figura 21 (e) Experimento #5: Desvio Padrão com filtro.

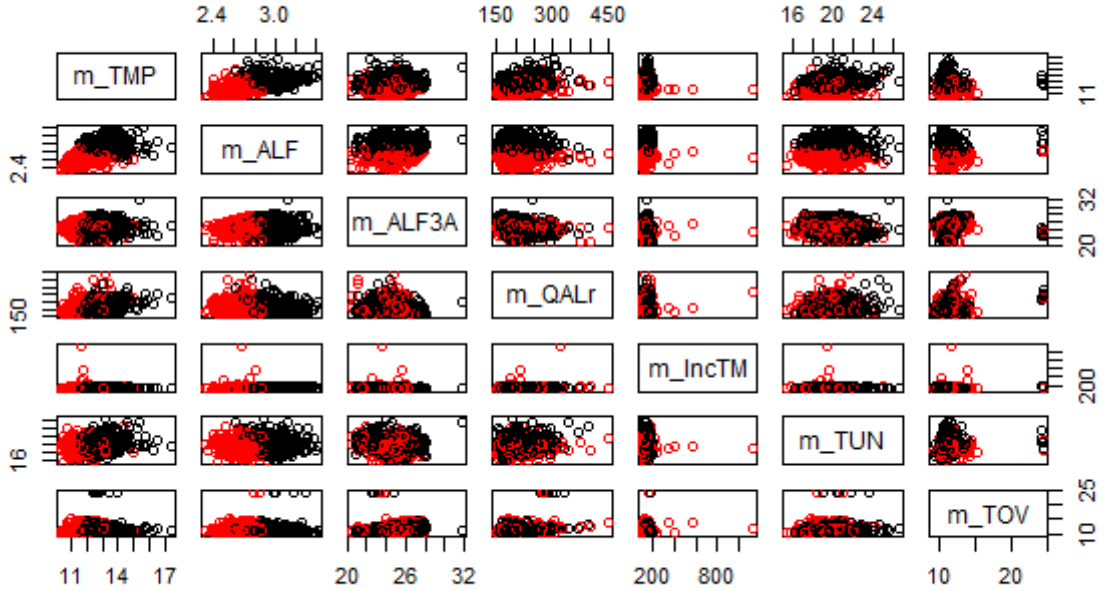


Figura 21 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento Par a Par por variável no FCM com 5 *Clusters*:

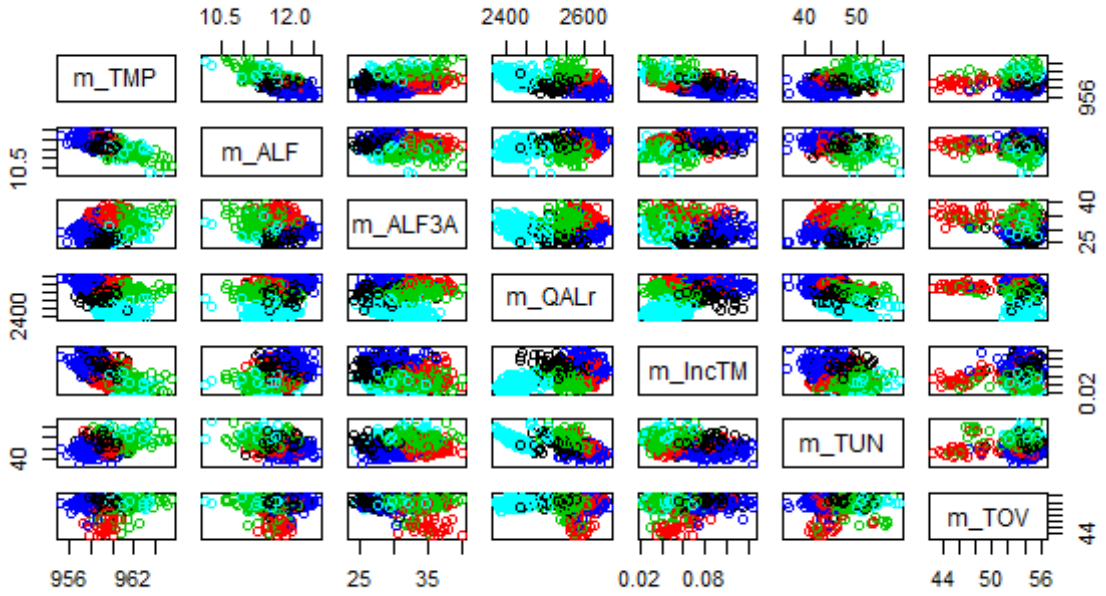


Figura 22 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante para 5 *Clusters*. Experimento #1: Média com filtro.

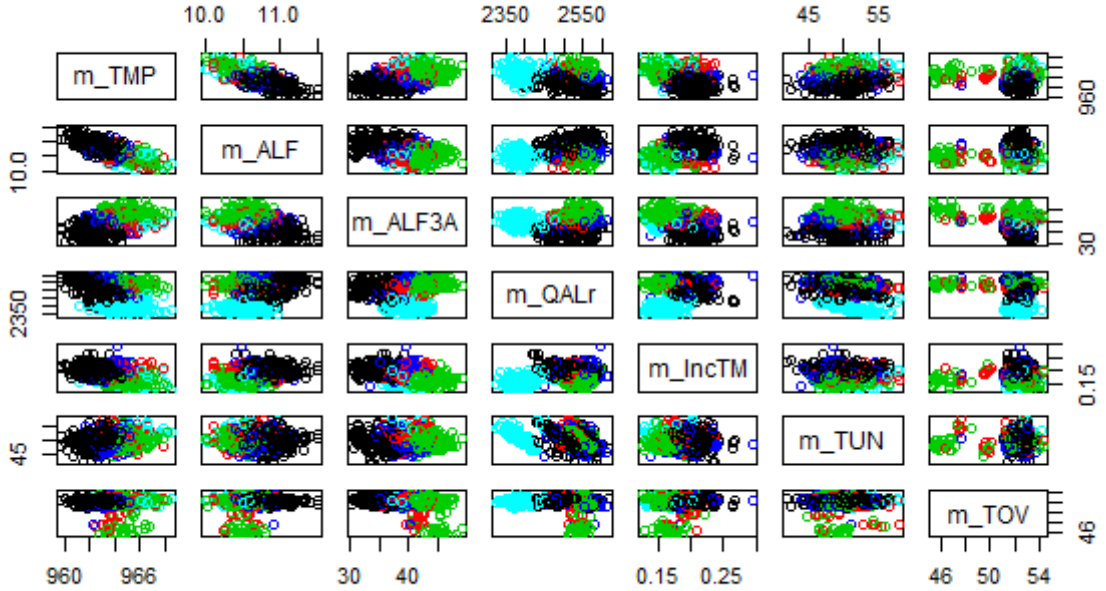


Figura 22 (b) Experimento #2: Média sem filtro.

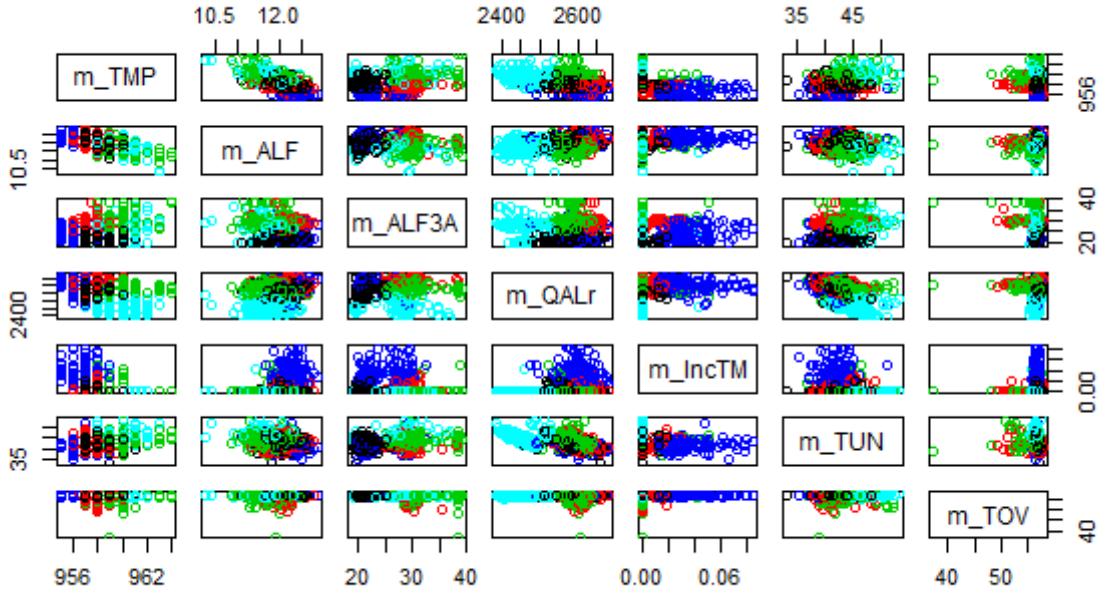


Figura 22 (c) Experimento #3: Mediana com filtro.

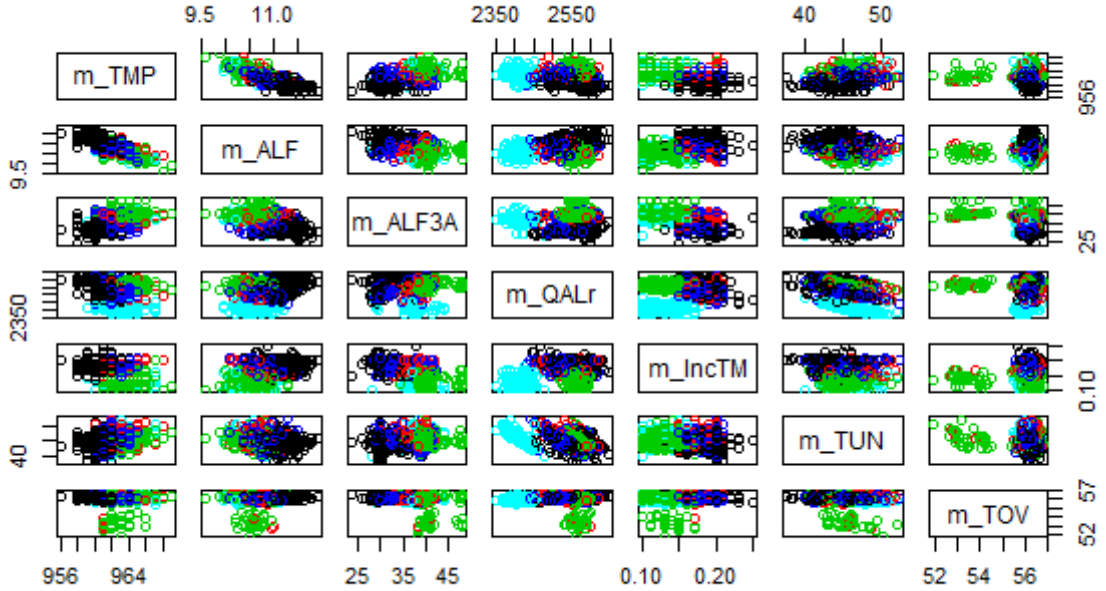


Figura 22 (d) Experimento #4: Mediana sem filtro.

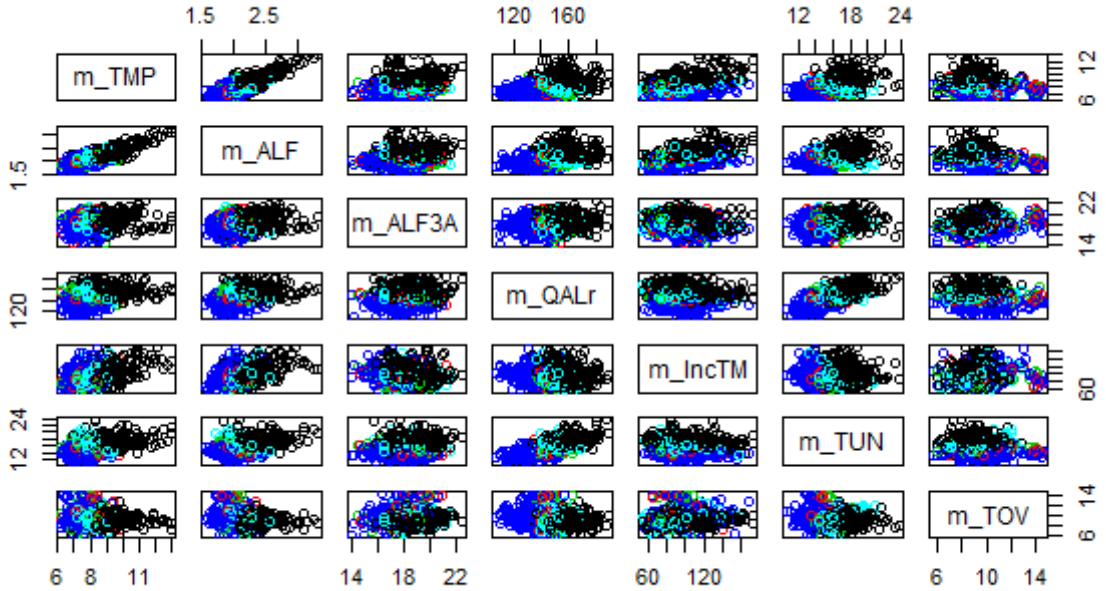


Figura 22 (e) Experimento #5: Desvio Padrão com filtro.

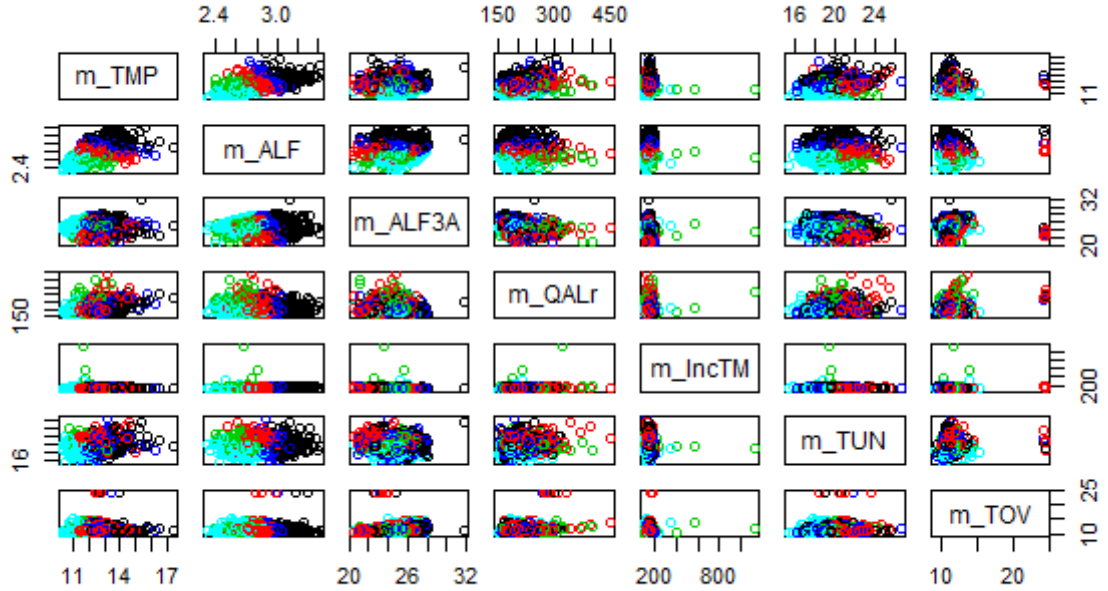


Figura 22 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento Par a Par por variável no FCM com 13 *Clusters*:

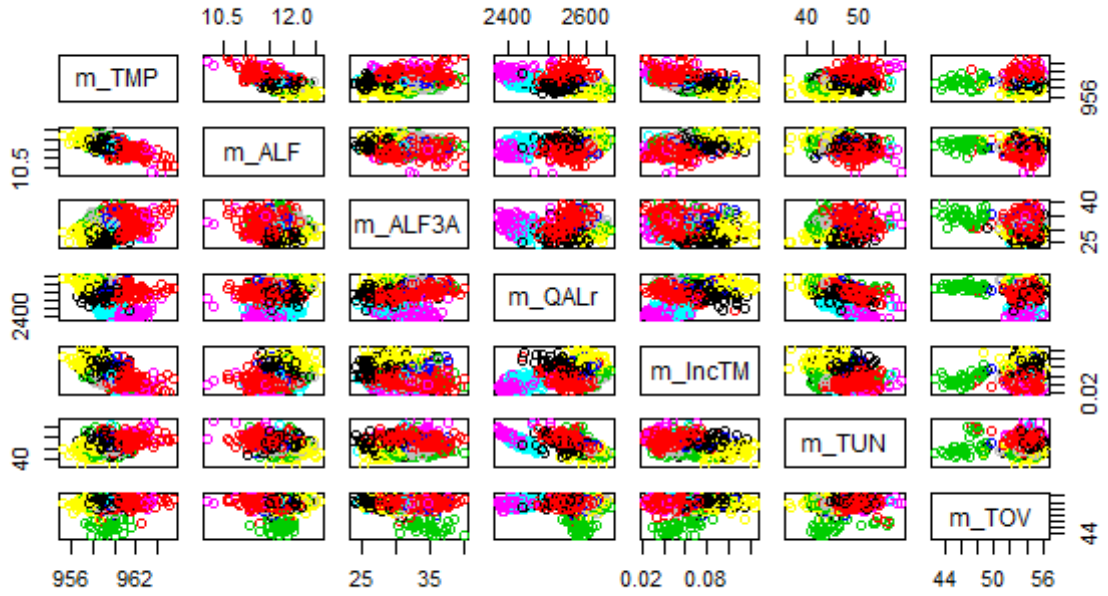


Figura 23 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante para 13 *Clusters*. Experimento #1: Média com filtro.

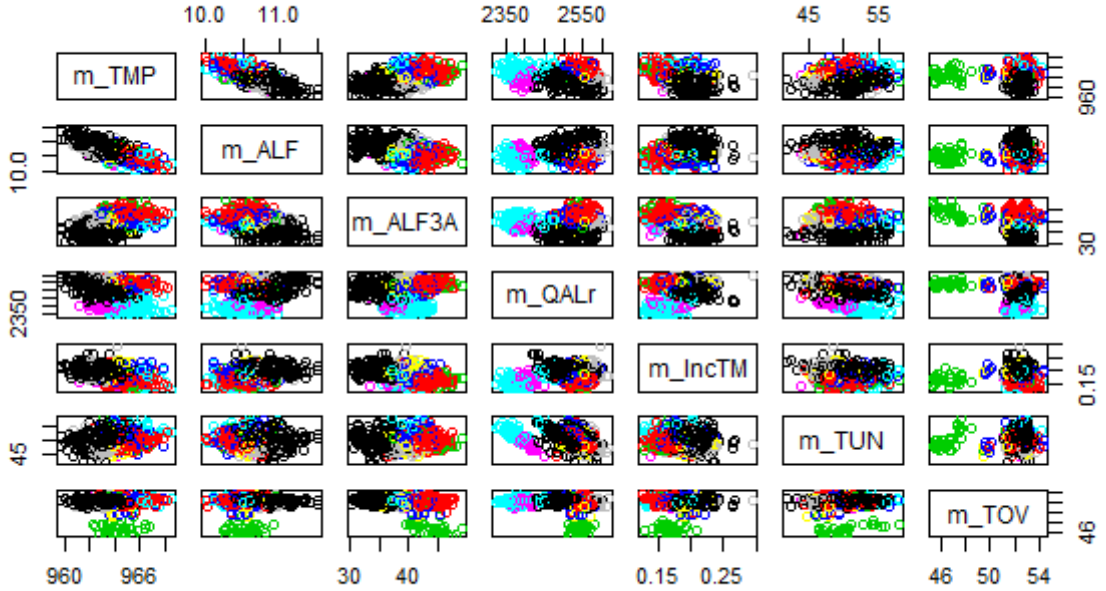


Figura 23 (b) Experimento #2: Média sem filtro.

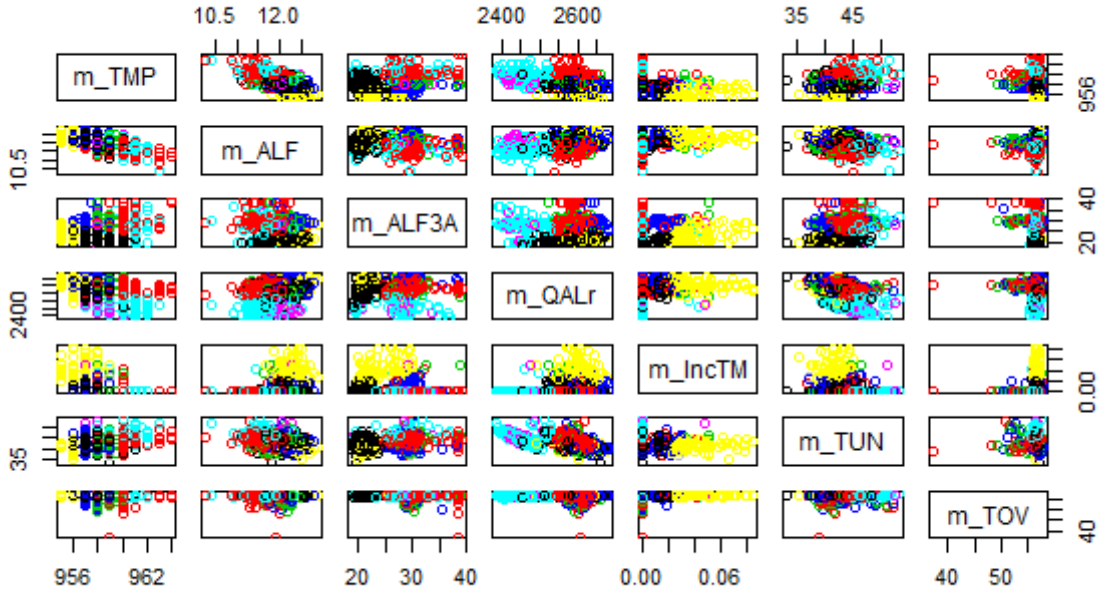


Figura 23 (c) Experimento #3: Mediana com filtro.

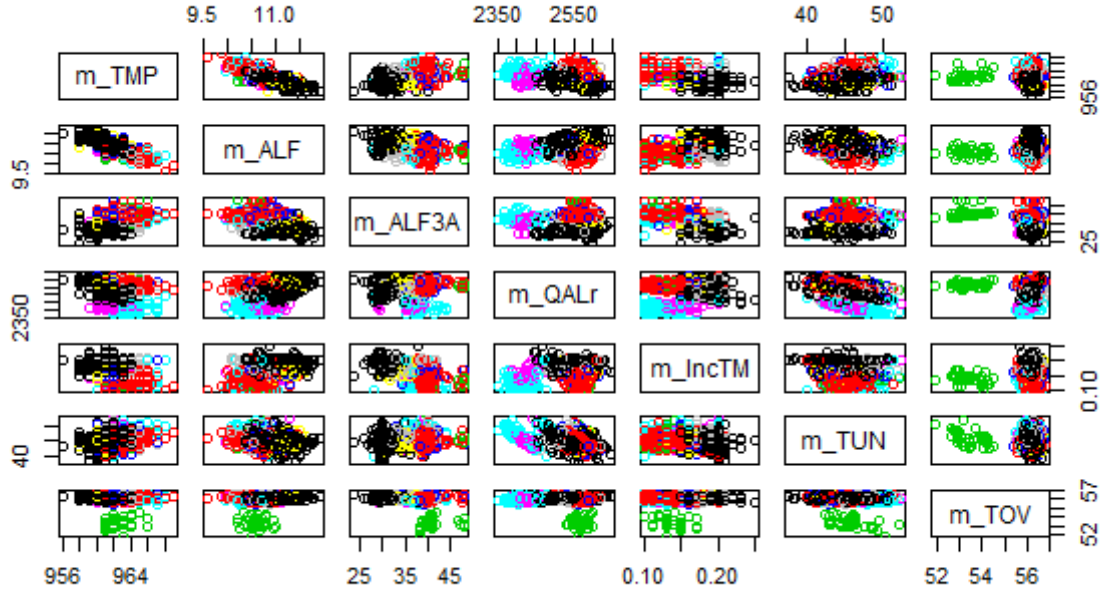


Figura 23 (d) Experimento #4: Mediana sem filtro.

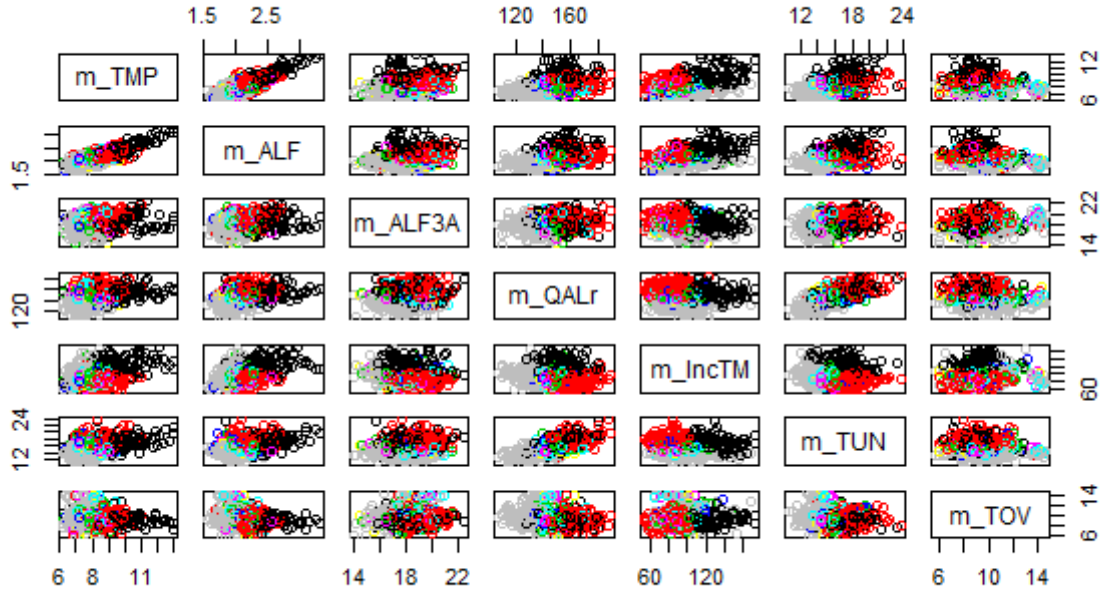


Figura 23 (e) Experimento #5: Desvio Padrão com filtro.

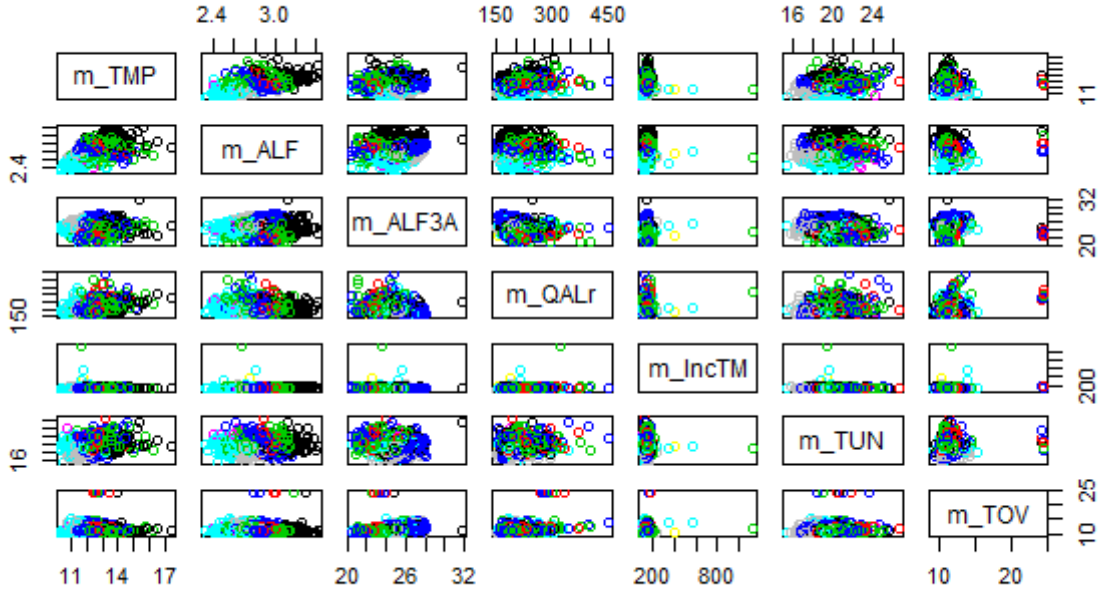


Figura 23 (f) Experimento #6: Desvio Padrão sem filtro.

Para atestar a precisão dos grupos em aglomerar ao redor deles fornos com características semelhantes, gráficos que mostram o valor de pertinência de cada grupo estão disponíveis através das Figuras 24 (a) à 24 (f). Como o gráfico é tridimensional, o experimento só pôde ser realizado com três *clusters*. Verifica–se que as Figuras 24 (a) à 24 (d) exibem valores de pertinência alto para cada um dos grupos. Enquanto as representações demonstradas nas Figuras 24 (e) e 24 (f) possuem valores de pertinência mediano ou baixo para os grupos vermelho e preto.

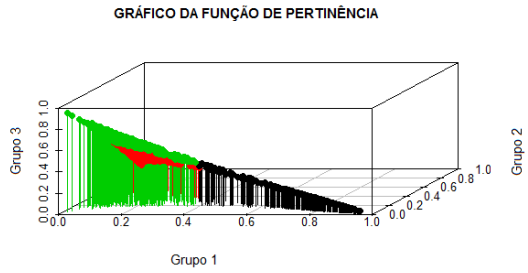


Figura 24 (a) – Gráficos dos experimentos com valores de pertinência para cada grupo encontrado. Experimento #1: Média com filtro.

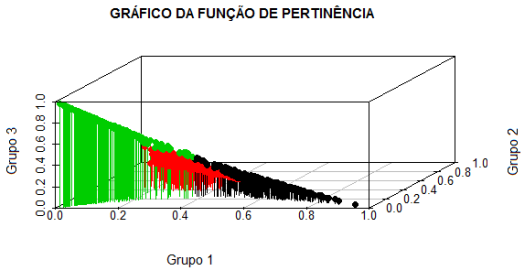


Figura 24 (b Experimento #2: Média sem filtro.

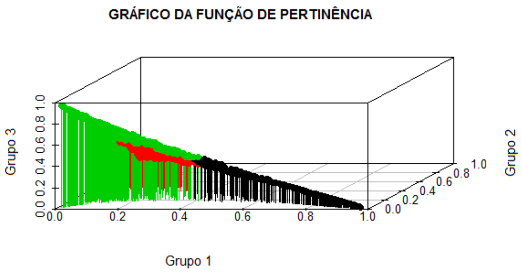


Figura 24 (c) Experimento #3: Mediana com filtro.

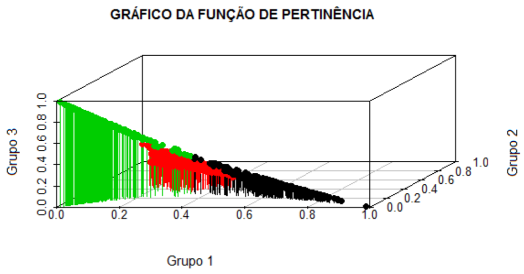


Figura 24 (d) Experimento #4: Mediana sem filtro.

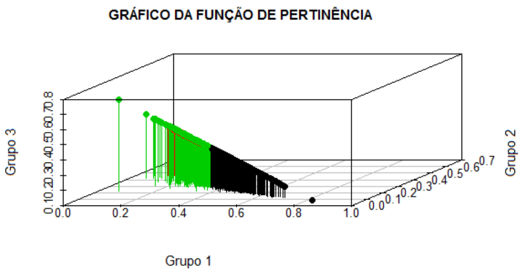


Figura 24 (e) Experimento #5: Desvio Padrão com filtro.

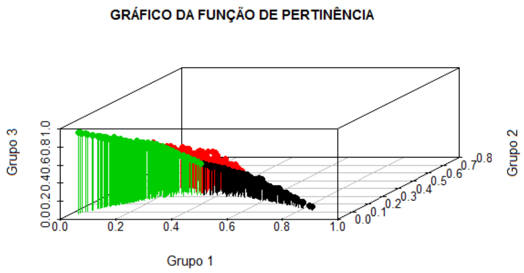


Figura 24 (f) Experimento #6: Desvio Padrão sem filtro.

No final de cada experimento, foi armazenado um vetor com o agrupamento dos fornos resultante. Conforme pode ser visto na Figura 25:

```
Closest hard clustering:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [81] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[161] 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[241] 2 1 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 1 1 2 1 2 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[321] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
[401] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 2 2 1 1 1 2 3 2 2 2 2 2 1 1 2 1 1 1 1 1 2 1 1 2 3 3
[481] 2 2 2 2 2 2 2 3 2 2 3 2 2 3 3 3 2 2 2 2 1 1 1 1 2 2 1 2 2 1 2 1 2 1 2 2 1 1 2 1 1 2 1 1 2 1 2 3 2 2 1 1 1 2 1 2 1 2 1 1 3 2 2 2 2 1 2 2 2 2 1 1 1 1 1 1 1 3 1 1
[561] 2 1 1 2 2 1 1 2 1 1 1 2 2 2 2 1 1 1 2 2 1 2 2 1 1 3 2 2 2 1 3 2 2 2 1 2 2 2 2 3 2 3 2 3 2 1 2 1 1 2 1 2 1 1 1 2 2 2 2 1 1 2 1 1 1 3 2 1 1 2 3 2 3 1 2 2 1 2 2 3
[641] 1 1 2 3 1 2 1 1 2 3 3 1 2 1 2 2 1 1 2 2 1 2 2 2 1 2 1 2 1 1 2 2 1 1 1 2 1 1 2 1 1 2 2 2 1 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 1 1 1 1 1 2 2 1 1 2 2 2 2 1 1 2 2 2 1
[721] 1 1 1 2 1 1 1 2 1 1 1 2 1 2 2 1 1 1 2 2 2 2 2 2 2 1 2 2 1 1 2 2 2 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 2 1 2 1 2 1 2 1 1 2 1 1
[801] 2 1 2 2 2 2 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 2 3 2 2 2 2 2 2 2 1
[881] 1 2 2 2 2 3 1 3 1 2 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 3 1 1 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 3 3 2 2 3 3 2 1 1 3 1 3 2 1 2 3 3 1 2 3 3 3 3 2 1 3 3 3 3 3
```

Figura 25 – Exemplo de *clustering* resultante.

Estes dados foram utilizados para gerar um mapeamento dos grupos de acordo com o layout dos fornos da fábrica de produção de alumínio, exemplificado pela Figura 6. Para isto, um programa desenvolvido em HTML (*HyperText Markup Language*), CSS (*Cascade Style Sheet*) e Javascript foi utilizado para gerar as imagens contidas nas Figuras 26 (a) à 28 (f), com o intuito de auxiliar a análise grupo *versus* localização através de cores diferentes para cada grupo, mapeando o algoritmo *Fuzzy C–Means* para 2, 5 e 13 *clusters*, respectivamente, utilizando a média com e sem filtro, mediana com e sem filtro e desvio padrão com e sem filtro para a análise de agrupamento dos fornos.

Mapeamento FCM com 2 *Clusters*:

Figura 26 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters*. Experimento #1: Média com filtro.

Figura 26 (b) Experimento #2: Média sem filtro.

Figura 26 (c) Experimento #3: Mediana com filtro.

Figura 26 (d) Experimento #4: Mediana sem filtro.

Figura 26 (e) Experimento #5: Desvio Padrão com filtro.

Figura 26 (f) Experimento #6: Desvio Padrão sem filtro.

Mapeamento FCM com 5 *Clusters*:

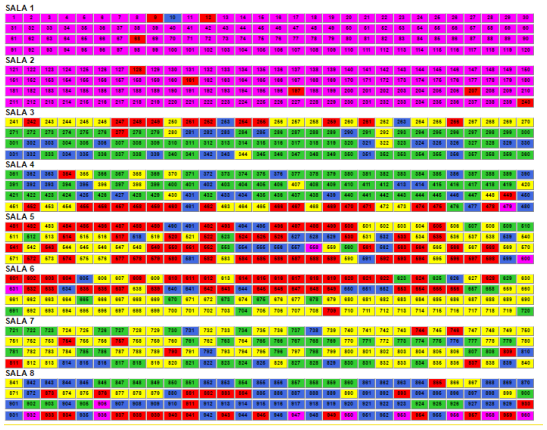


Figura 27 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters*. Experimento #1: Média com filtro.

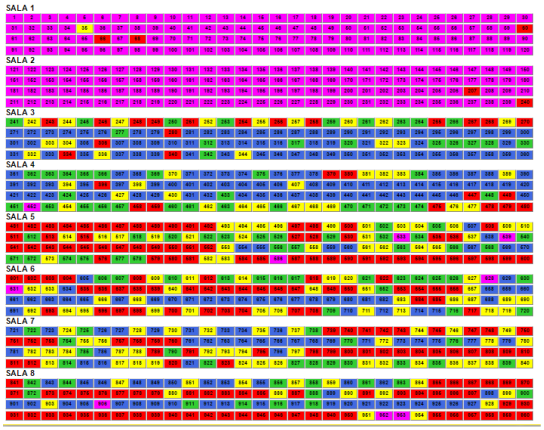


Figura 27 (b) Experimento #2: Média sem filtro.

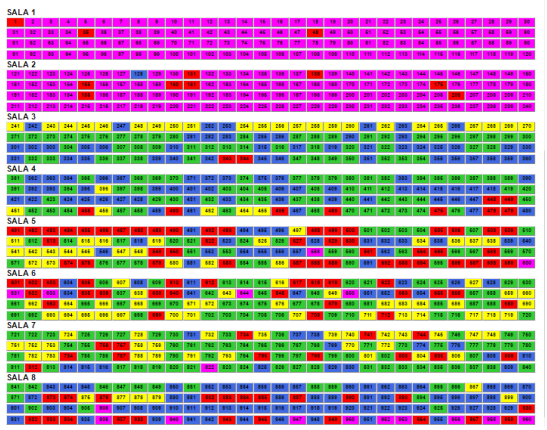


Figura 27 (c) Experimento #3: Mediana com filtro.

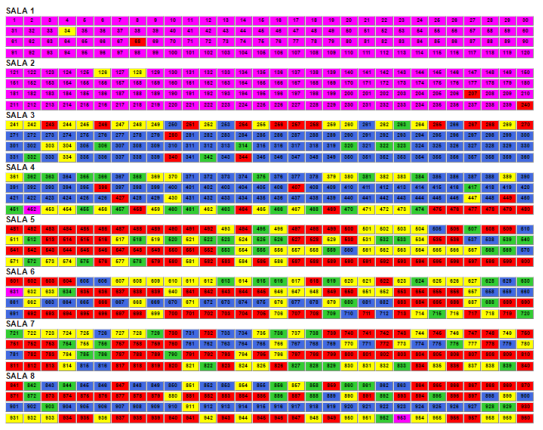


Figura 27 (d) Experimento #4: Mediana sem filtro.

Figura 27 (e) Experimento #5: Desvio Padrão com filtro.

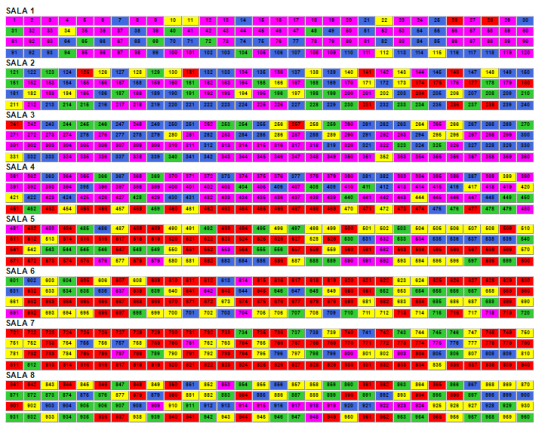


Figura 27 (f) Experimento #6: Desvio Padrão sem filtro.

Mapeamento FCM com 13 *Clusters*:

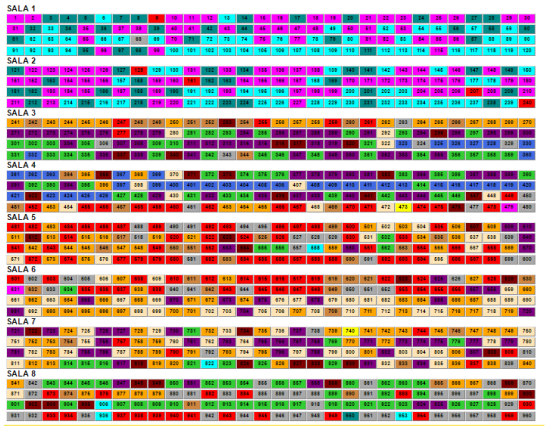


Figura 28 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters*. Experimento #1: Média com filtro.

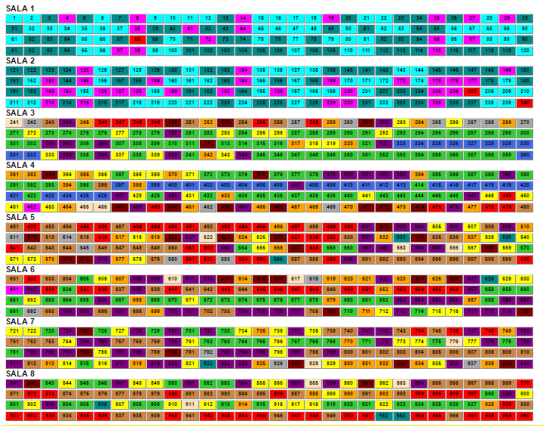


Figura 28 (b) Experimento #2: Média sem filtro.

Figura 28 (c) Experimento #3: Mediana com filtro.

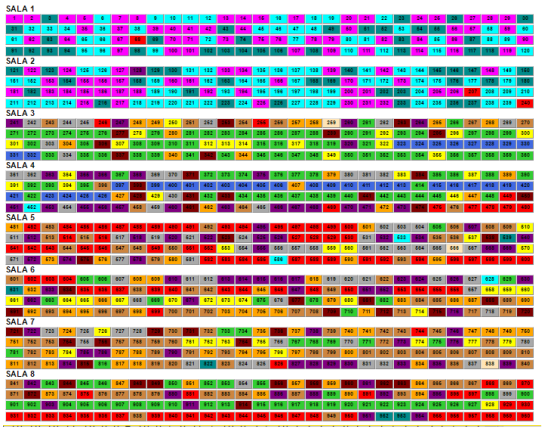


Figura 28 (d) Experimento #4: Mediana sem filtro.

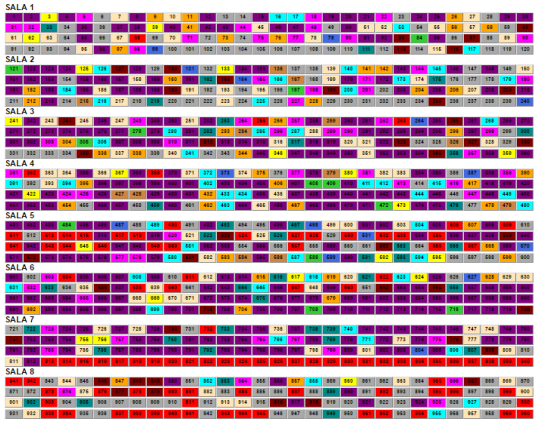


Figura 27 (e) Experimento #5: Desvio Padrão com filtro.

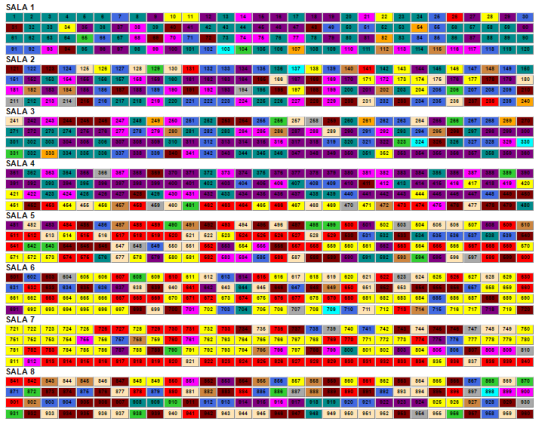


Figura 27 (f) Experimento #6: Desvio Padrão sem filtro.

Através da análise dos mapeamentos do algoritmo *Fuzzy C–Means* para 2, 5 e 13 *Clusters* é possível notar que o agrupamento de 2 a 5 *Clusters* para a média e mediana com filtro e sem filtro possuem suas duas primeiras salas com cores em sua maioria uniformes, o que significa que nas duas primeiras salas há o predomínio de um mesmo agrupamento. Para os agrupamentos 2 e 5 *clusters*, o desvio padrão com filtro e sem filtro não apresentaram uma boa clusterização, já que é notório a mistura de grupos em todas as salas analisadas.

Nos agrupamentos com 13 *clusters* se observa uma mistura de grupos em todas as salas, tanto para a análise de média, mediana e desvio padrão com filtro e sem filtro, o que caracteriza um mau agrupamento.

6.4. *K–MEANS*

O primeiro experimento realizado através do algoritmo *K–Means* resultou em um DENDROGRAMA, o qual corresponde a um gráfico que mostra os agrupamentos realizados através das distâncias calculadas a partir de variáveis da base. Ele ensina que é preciso fazer um corte no dendrograma para achar o número de agrupamentos que será utilizado no algoritmo *K–Means*, ou seja, o valor de K.

O dendrograma resultante dos dados que compreendem a Forno 1 até o Forno 50 é representado pela Figura 29 abaixo. Ao traçar uma reta para achar o *K*, se verifica que ela corta sete galhos. Portanto, para este conjunto de dados, K=7. Observa–se que o valor de *K* poderia ser 6, já que no penúltimo corte, o Forno 35 fica isolado em um grupo.

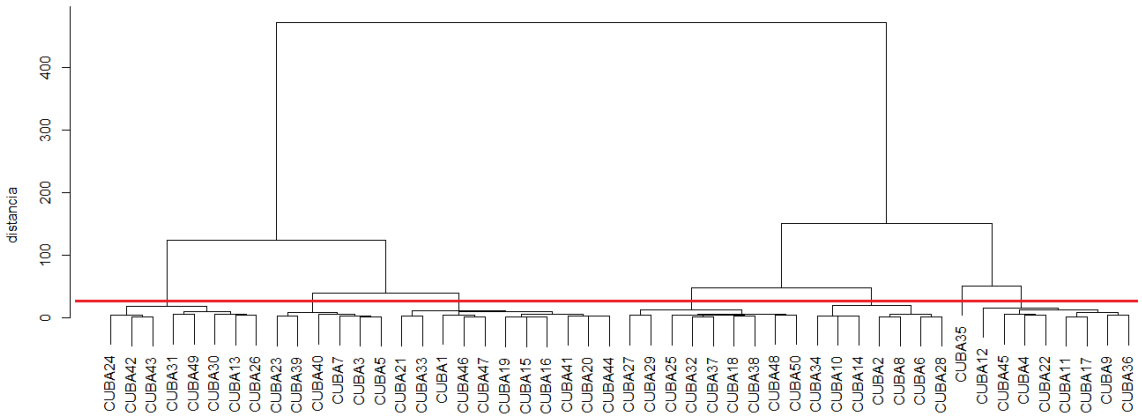


Figura 29 – Dendrograma considerando dados do Forno1 até a Forno 50 (K=7).

Após esta análise, uma contestação surge: Será que ao considerar todos os dados, ou seja, da Forno 1 até a Forno 960, terá K=6 ou K=7?

Para responder a esta pergunta, o programa desenvolvido em R foi executado novamente, agora considerando dados de todos os Fornos, e um novo dendrograma foi encontrado. Por intermédio da Figura 30, é possível verificar que a linha vermelha cortou oito galhos. Sendo assim, de acordo com esta técnica, para agrupar todos os 960 Fornos eletrolíticos, é necessário que K seja igual a 8.

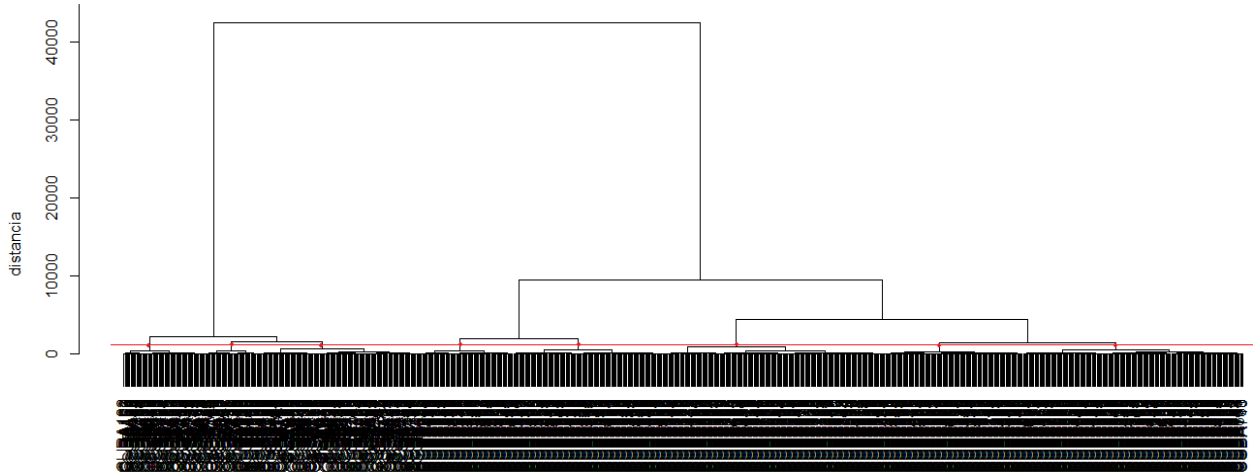


Figura 30 – Dendrograma considerando dados de todos os Fornos (K=8).

Ressalta–se que outros valores de *K* podem ser utilizados. Para analisar os agrupamentos resultantes, se estudou um tutorial (Overflow, 2012), onde vários pesquisadores da área de estatística discutem uma forma de visualizar melhor os resultados do *K–Means* através do programa RStudio, também conhecido como R. Um deles cita uma rotina que leva em consideração uma biblioteca externa do R, intitulada "FPC". Esta biblioteca pode ser facilmente instalada pelo próprio ambiente do R, acessando o menu "Pacotes".

A rotina utilizada é mostrada na Tabela 13 a seguir:

Tabela 13 – Código–Fonte programado em R para agrupamento usando *K–Means*.

```
library(cluster)
library(FPC)

Fornos <– read.csv2(file='planilhaMédias_parte1.csv');

# Kmeans cluster analysis
clus <– kmeansruns(Fornos, center=6, runs=100)

# Fig 01
plotcluster(Fornos, clus$cluster)
print(clus)
```

Sabe–se que as salas da fábrica de alumínio estudada possuem o total de 960 fornos eletrolíticos, no entanto, para o primeiro experimento, foram utilizados 50 fornos, com o intuito de verificar o comportamento inicial dos agrupamentos, para então realizar os testes com os 960 fornos. A rotina utilizada apresenta os seguintes resultados expostos na Tabela 14:

Tabela 14 – Resultado do agrupamento usando *K–Means*.

K–Means clustering with 6 *Clusters* of sizes 12, 11, 7, 7, 12, 1

Cluster means:

	m_TMP	m_ALF	m_ALF3A	m_QALr	m_IncTM	m_TUN	m_TOV
1	964.0431	10.48209	39.79966	2387.839	0.1520937	50.19658	52.49008
2	963.8373	10.51272	40.03450	2418.779	0.1598167	48.01081	52.47951
3	965.0033	10.51246	41.83794	2379.353	0.1477008	52.24146	52.47031
4	965.2150	10.38872	41.00629	2367.298	0.1589267	53.41787	52.39285
5	964.3490	10.51247	40.24452	2401.363	0.1512953	49.58048	52.37546
6	964.1203	10.58809	33.35525	2451.767	0.1387892	43.89210	52.41602

Clustering vector:

FORNO1 FORNO2 FORNO3 FORNO4 FORNO5 FORNO6 FORNO7 FORNO8 FORNO9 FORNO10 FORNO11

1 5 3 2 3 5 3 5 2 2 2

FORNO12 FORNO13 FORNO14 FORNO15 FORNO16 FORNO17 FORNO18 FORNO19 FORNO20 FORNO21 FORNO22

2 4 2 1 1 2 5 1 1 1 2

FORNO23 FORNO24 FORNO25 FORNO26 FORNO27 FORNO28 FORNO29 FORNO30 FORNO31 FORNO32 FORNO33

3 3 5 4 1 5 5 4 4 5 1

FORNO34 FORNO35 FORNO36 FORNO37 FORNO38 FORNO39 FORNO40 FORNO41 FORNO42 FORNO43 FORNO44

2 6 2 5 5 3 3 1 4 4 1

FORNO45 FORNO46 FORNO47 FORNO48 FORNO49 FORNO50

2 1 1 5 4 5

Within *cluster* sum of squares by *cluster*:

[1] 138.16127 388.36429 79.41077 139.92600 176.74973 0.00000

(between_SS / total_SS = 95.1 %)

A Tabela 14 demonstra a quantidade de *clusters* resultantes através do experimento realizado com o algoritmo *K-Means*, onde se obteve 6 grupos, tendo o 1° grupo 12 fornos, o 2° grupo 11 fornos, o 3° grupo 7 fornos, o 4° grupo 7 fornos, o 5° grupo 12 fornos e o 6° grupo apenas 1 forno, sendo o forno 1 disposto no grupo 1, o forno 2 disposto no grupo 5, o forno 3 disposto no grupo 3, o forno 4 disposto no grupo 2, e assim sucessivamente, adquirindo neste experimento um grau de acurácia (precisão) de 95,1%.

Como dito anteriormente, após a análise com 50 fornos, foram realizados experimentos com os 960 fornos, com o intuito de verificar se o comportamento do agrupamento.

As Figuras 31 (a) à 33 (f) demonstram o agrupamento do algoritmo *K–Means* para 2, 5 e 13 *clusters*, respectivamente, de acordo com a similaridade de cada forno ao seu destinado grupo, onde dc 1 e dc 2 significam discriminante canônica, que é uma técnica da estatística multivariada que permite a redução da dimensionalidade de dados, e é semelhante a componentes principais e correlações canônicas. Essa técnica é especialmente empregada em análises discriminantes realizadas a partir de amostras com observações repetidas. A análise discriminante canônica também pode ser utilizada para representar várias populações em um subespaço de menor dimensão. A análise procura, com base em um grande número de características originais correlacionadas, obter combinações lineares dessas características denominadas variáveis canônicas de tal forma que a correlação entre essas variáveis seja nula (Khattree e Naik, 2000). A utilização dessa técnica permite capturar o efeito simultâneo de características originais e com isso pode capturar variações não percebidas quando do uso de características originais isoladamente.

Assim como para o experimento com 50 fornos, também foi criado uma rotina afim de uma melhor visualização dos resultados do *K–Means* através do programa RStudio. Essa rotina também levou em consideração uma biblioteca externa do R, intitulada “FPC”, sendo instalada pelo próprio ambiente do R, acessando o menu “Pacotes”.

Utilizou–se esta rotina, definindo o número total de cubas como sendo 960 e variando o valor de K para 2, 5 e 13, respectivamente, tendo como parâmetro a média com filtro e sem filtro, mediana com filtro e sem filtro e desvio padrão com filtro e sem filtro, conforme os resultados das Figuras 31 (a) à 33 (f) a seguir:

Agrupamento *K–Means* com 2 *Clusters*:

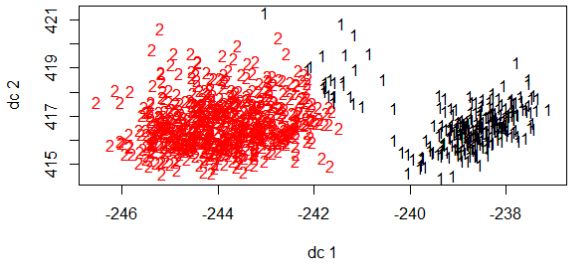


Figura 31 (a) Gráfico de cada experimento para agrupar os *Clusters* de acordo com suas similaridades no algoritmo *K–Means* para 2 *Clusters*. Experimento #1: Média com filtro.

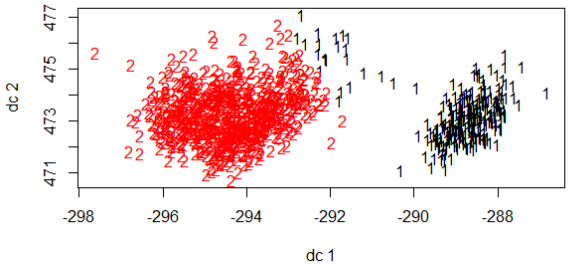


Figura 31 (b) Experimento #2: Média sem filtro.

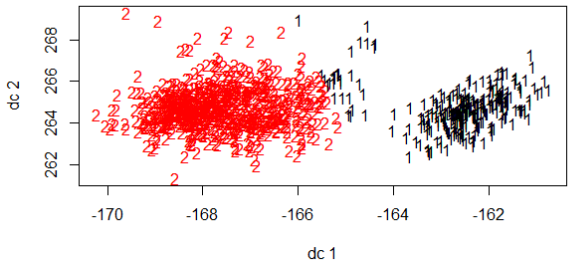


Figura 31 (c) Experimento #3: Mediana com filtro.

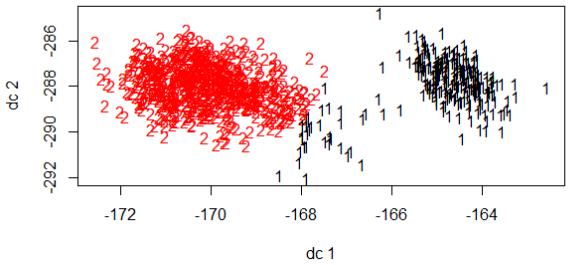


Figura 31 (d) Experimento #4: Mediana sem filtro.

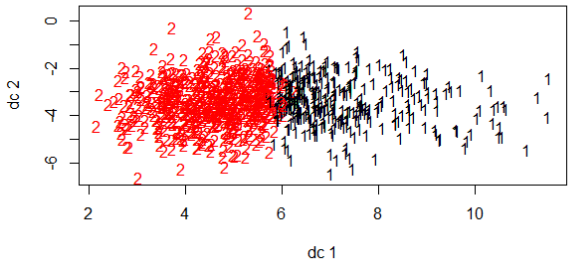


Figura 31 (e) Experimento #5: Desvio Padrão com filtro.

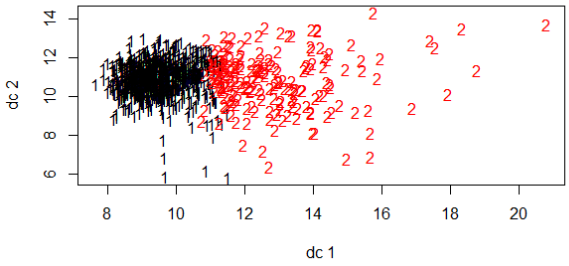


Figura 31 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento *K–Means* com 5 *Clusters*:

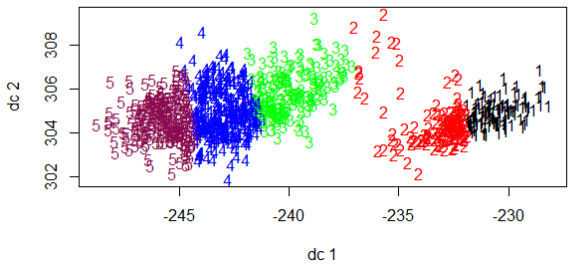


Figura 32 (a) Gráfico de cada experimento para agrupar os *Clusters* de acordo com suas similaridades no algoritmo *K–Means* para 5 *Clusters*. Experimento #1: Média com filtro.

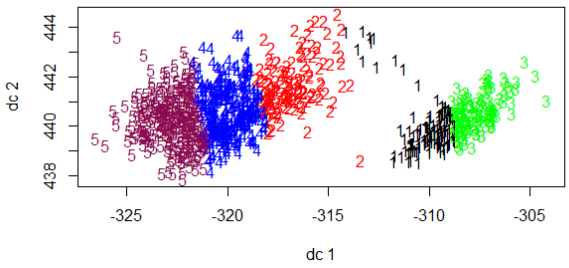


Figura 32 (b) Experimento #2: Média sem filtro.

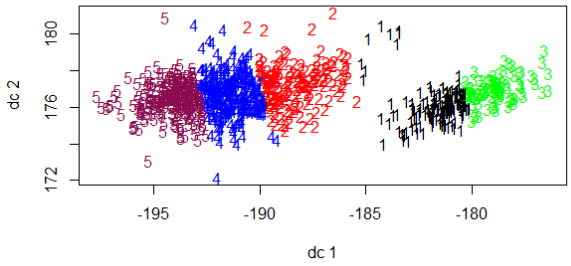


Figura 32 (c) Experimento #3: Mediana com filtro.

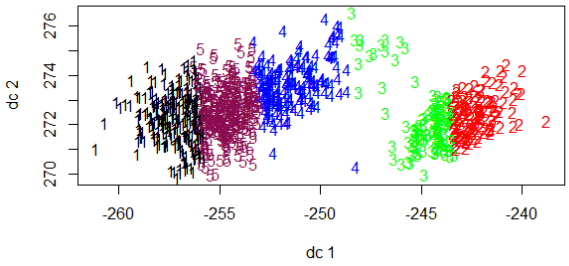


Figura 32 (d) Experimento #4: Mediana sem filtro.

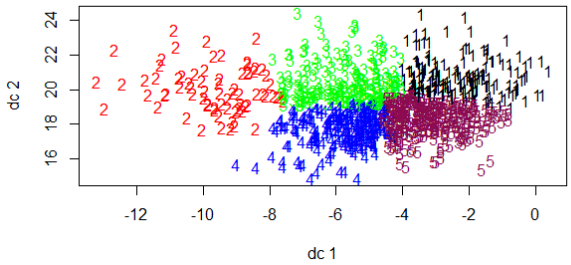


Figura 32 (e) Experimento #5: Desvio Padrão com filtro.

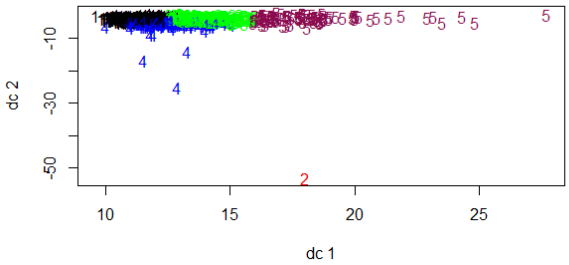


Figura 32 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento *K–Means* com 13 *Clusters*:

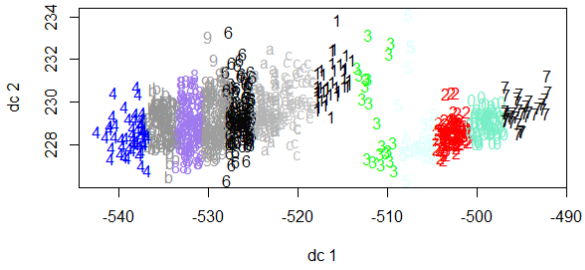


Figura 33 (a) Gráfico de cada experimento para agrupar os *Clusters* de acordo com suas similaridades no algoritmo *K–Means* para 13 *Clusters*. Experimento #1: Média com filtro.

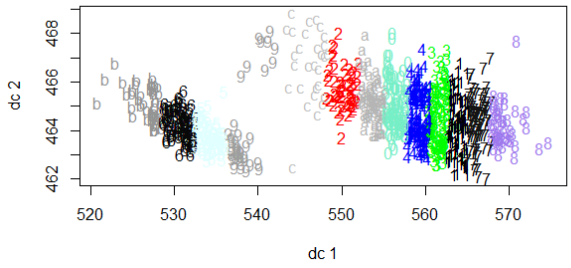


Figura 33 (b) Experimento #2: Média sem filtro.

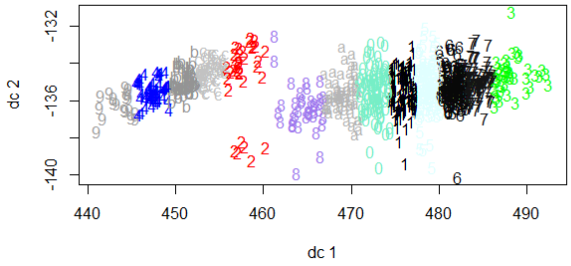


Figura 33 (c) Experimento #3: Mediana com filtro.

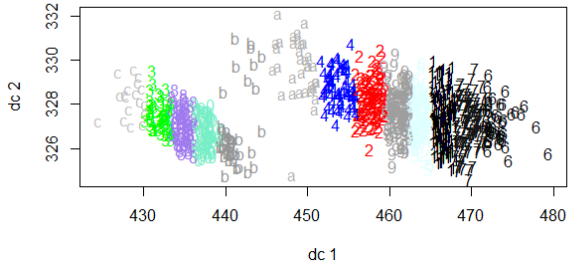


Figura 33 (d) Experimento #4: Mediana sem filtro.

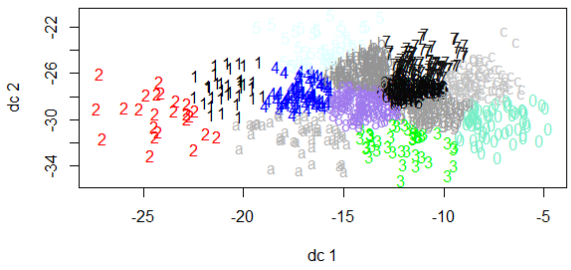


Figura 33 (e) Experimento #5: Desvio Padrão com filtro.

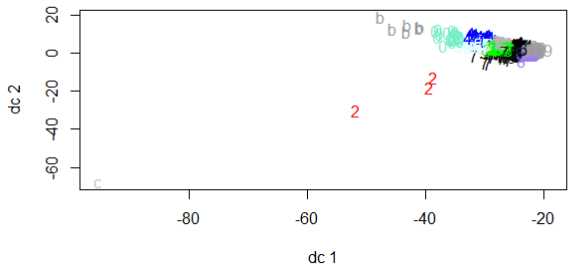


Figura 33 (f) Experimento #6: Desvio Padrão sem filtro.

As Figuras 34 (a) à 36 (f) mostram os resultados da relação de agrupamento entre pares de variáveis do algoritmo K–Means para 2, 5 e 13 clusters, respectivamente, de acordo com a similaridade de cada forno ao seu destinado grupo.

Assim como no FCM, é possível notar que em todas as imagens há uma separação bem definida na quarta coluna e na quarta linha. Como a quarta variável é a QALr, então, na maioria dos casos, quando se compara outra variável do processo com esta, o agrupamento é mais preciso em relação às outras comparações entre variáveis.

Agrupamento Par a Par por variável no *K–Means* com 2 *Clusters*:

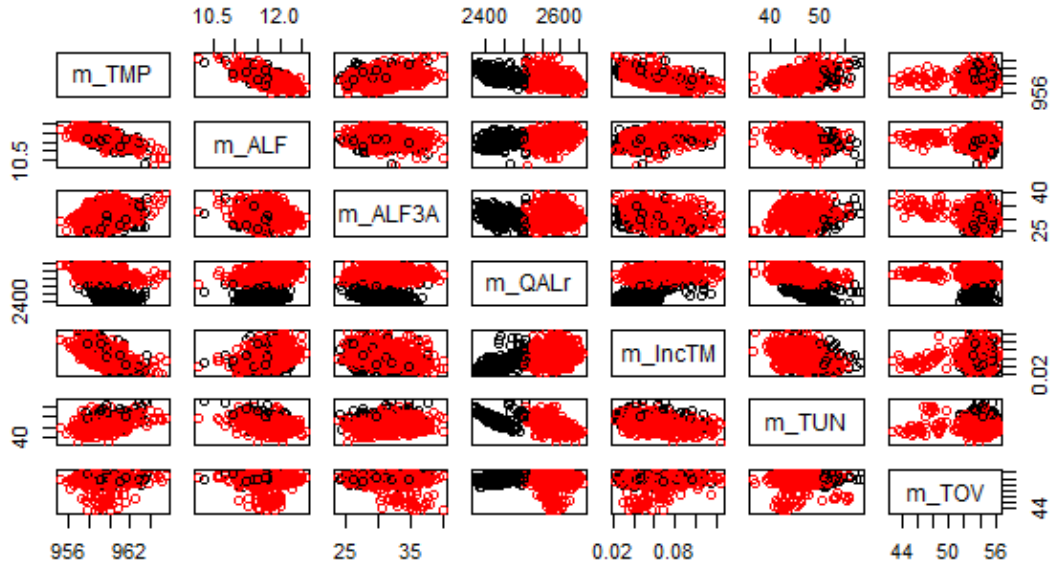


Figura 34 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante com 2 *Clusters*. Experimento #1: Média com filtro.

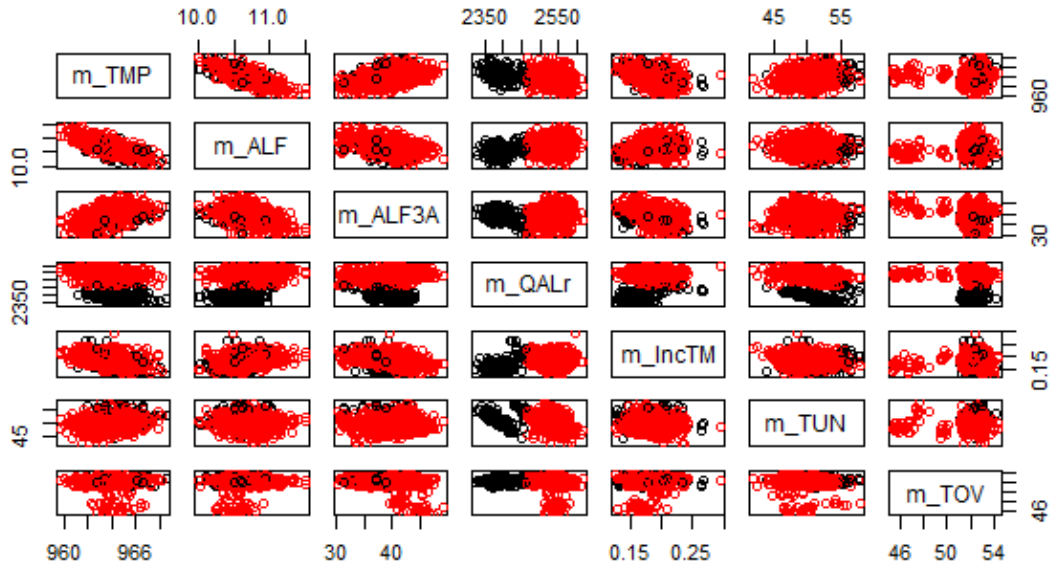


Figura 34 (b) Experimento #2: Média sem filtro.

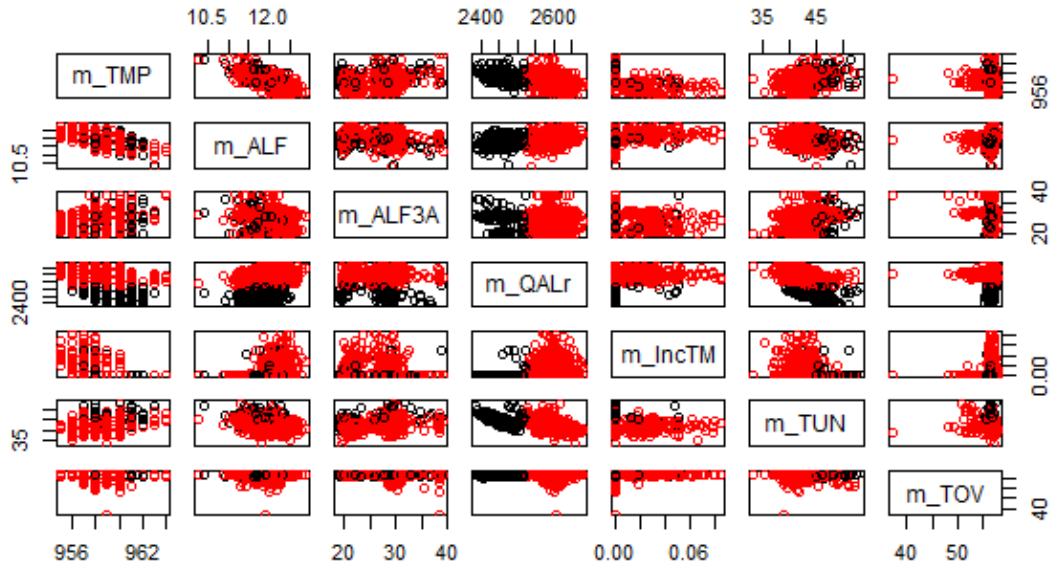


Figura 34 (c) Experimento #3: Mediana com filtro.

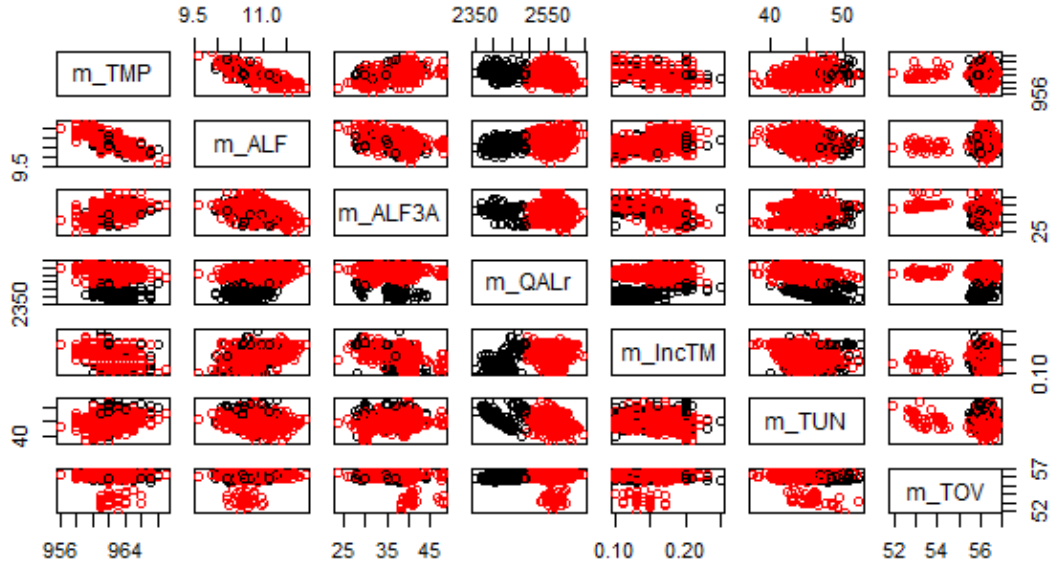


Figura 34 (d) Experimento #4: Mediana sem filtro.

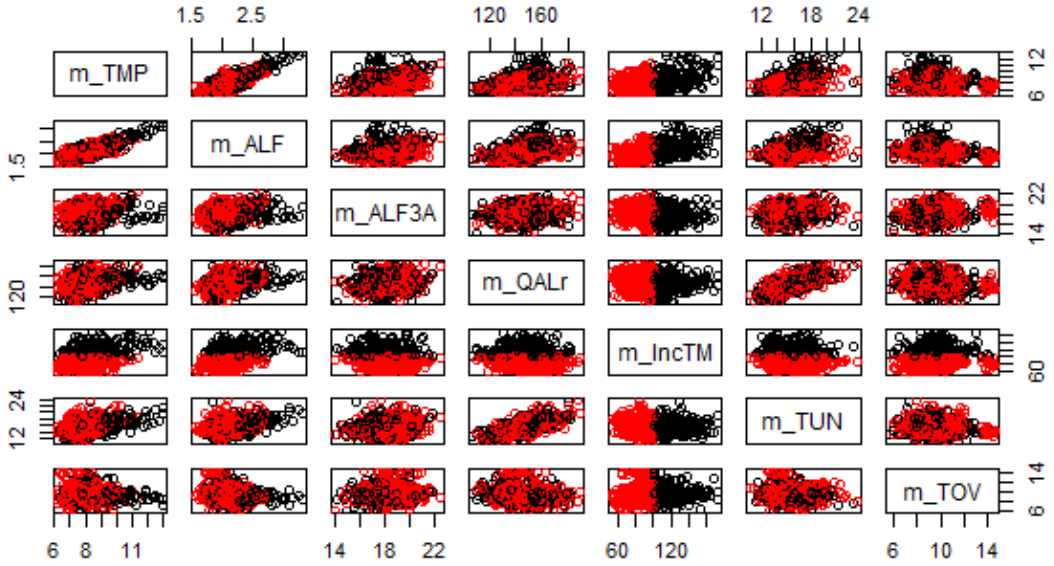


Figura 34 (e) Experimento #5: Desvio Padrão com filtro.

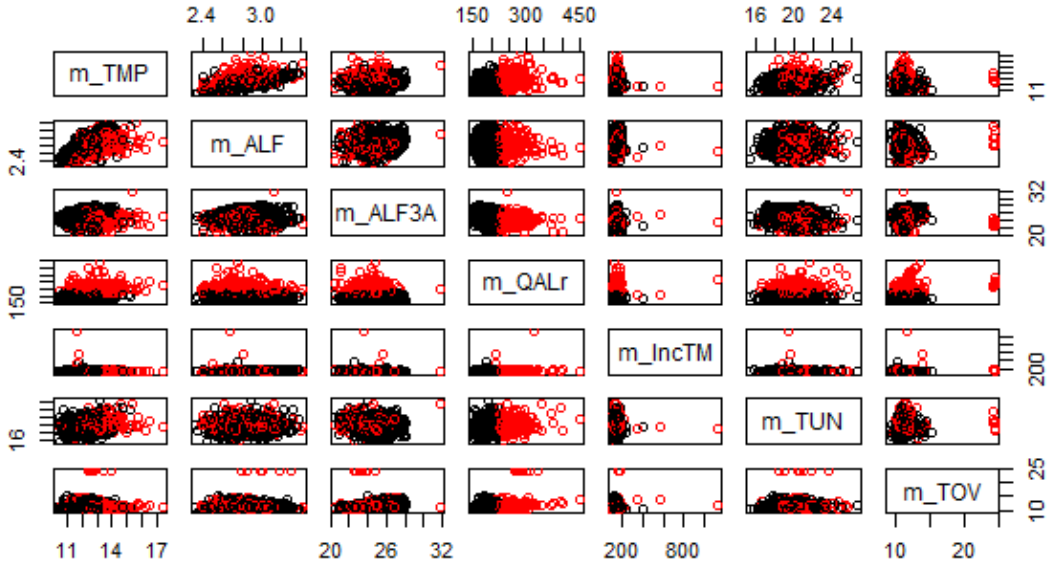


Figura 34 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento Par a Par por variável no *K–Means* com 5 *Clusters*:

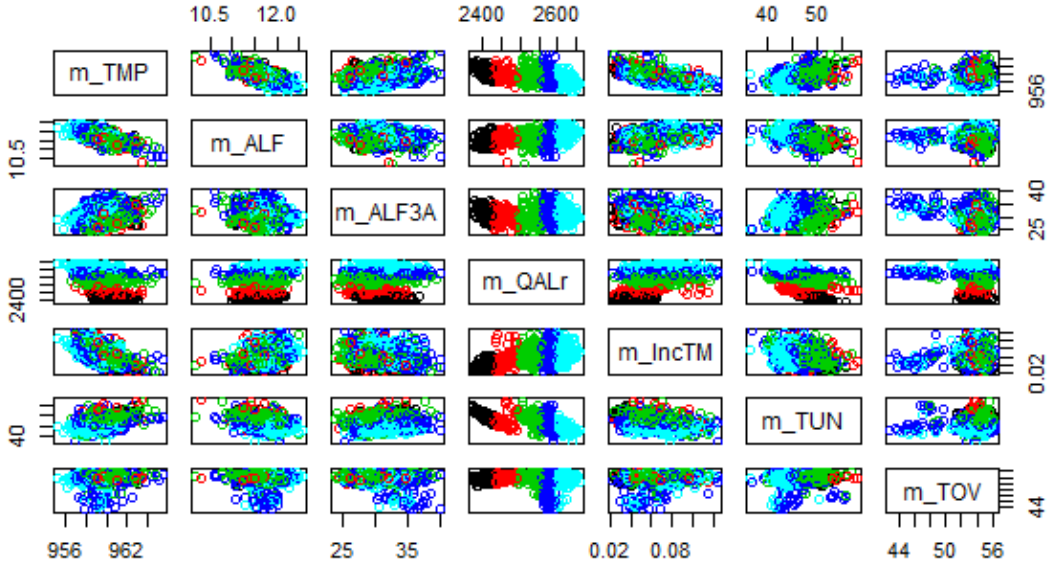


Figura 35 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante com 5 *Clusters*. Experimento #1: Média com filtro.

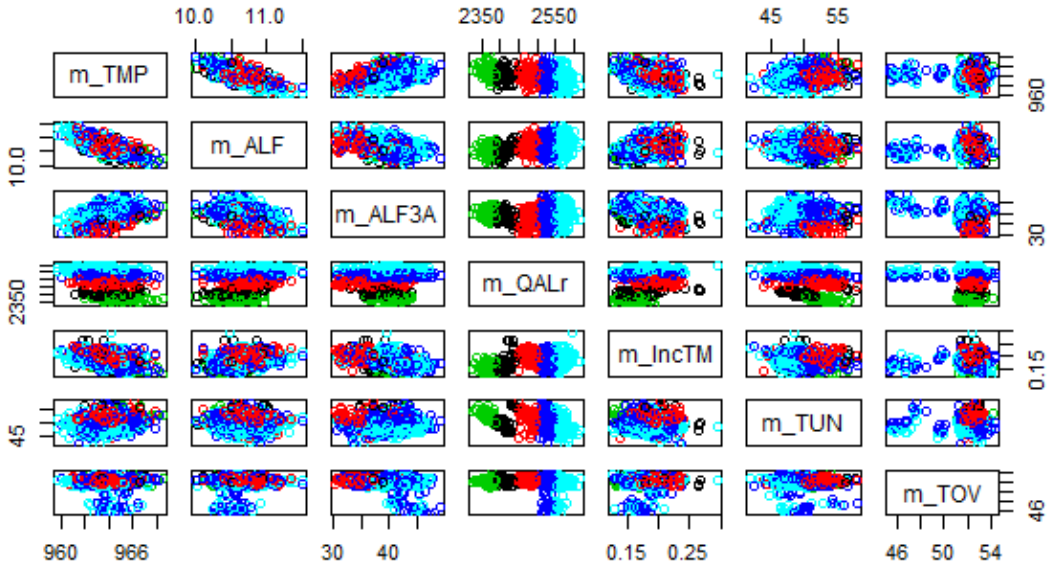


Figura 35 (b) Experimento #2: Média sem filtro.

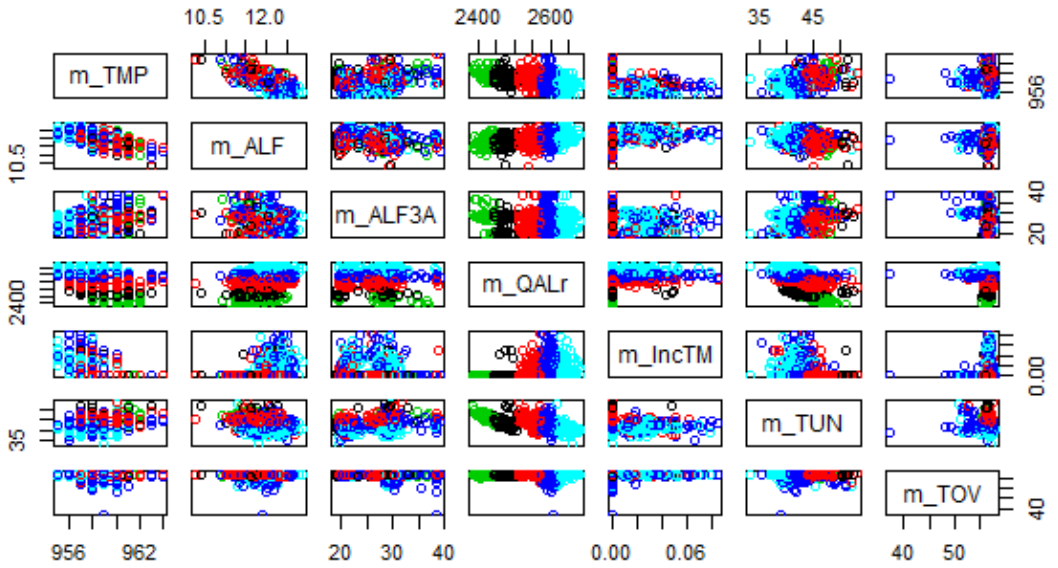


Figura 35 (c) Experimento #3: Mediana com filtro.

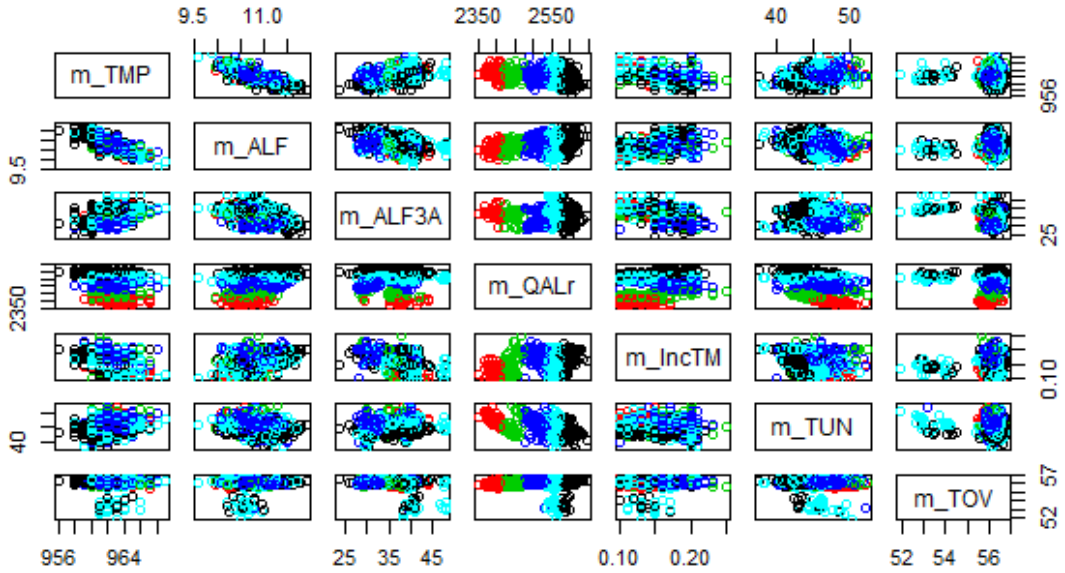


Figura 35 (d) Experimento #4: Mediana sem filtro.

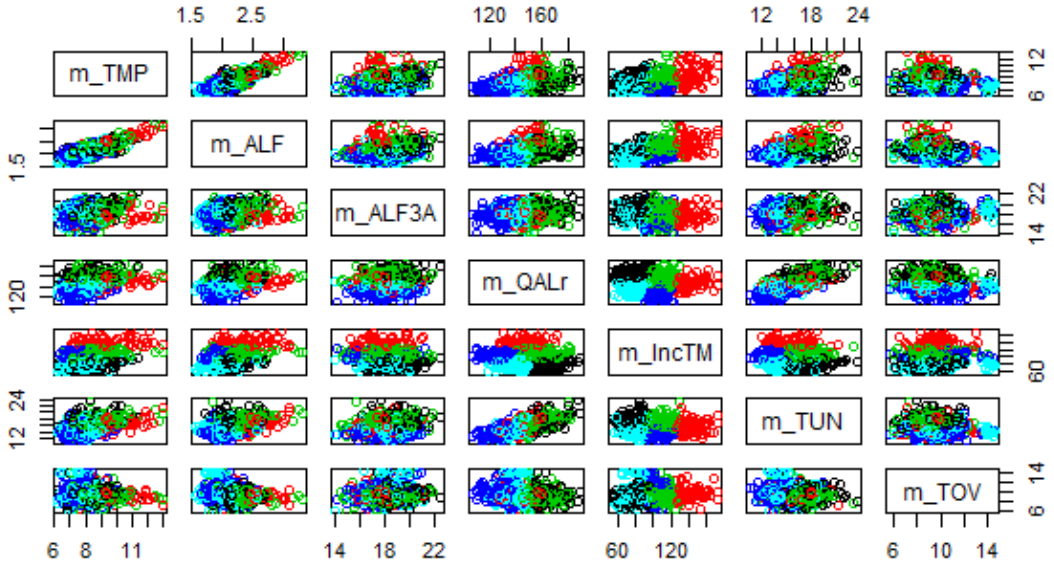


Figura 35 (e) Experimento #5: Desvio Padrão com filtro.

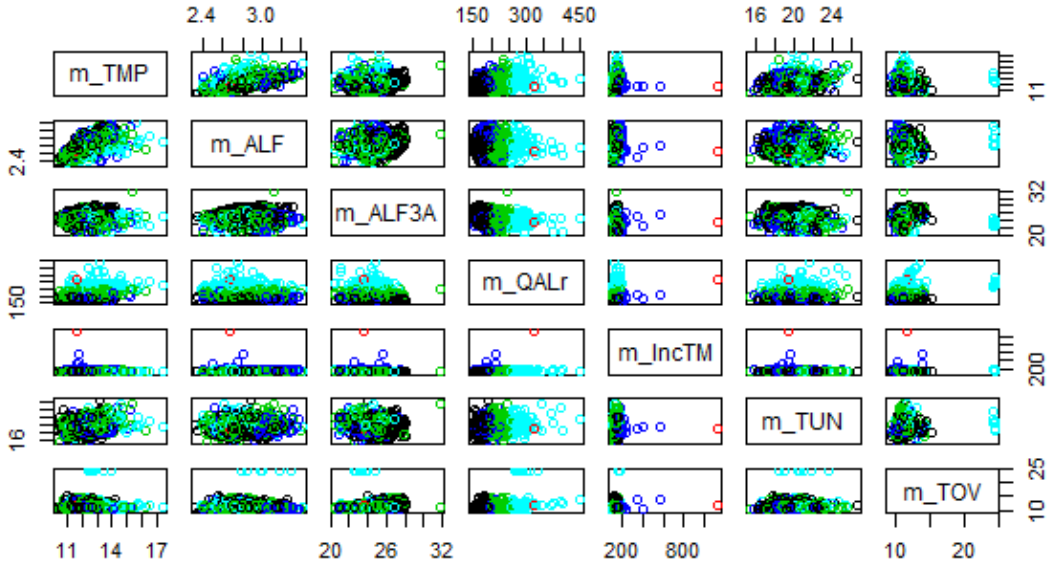


Figura 35 (f) Experimento #6: Desvio Padrão sem filtro.

Agrupamento Par a Par por variável no *K–Means* com 13 *Clusters*:

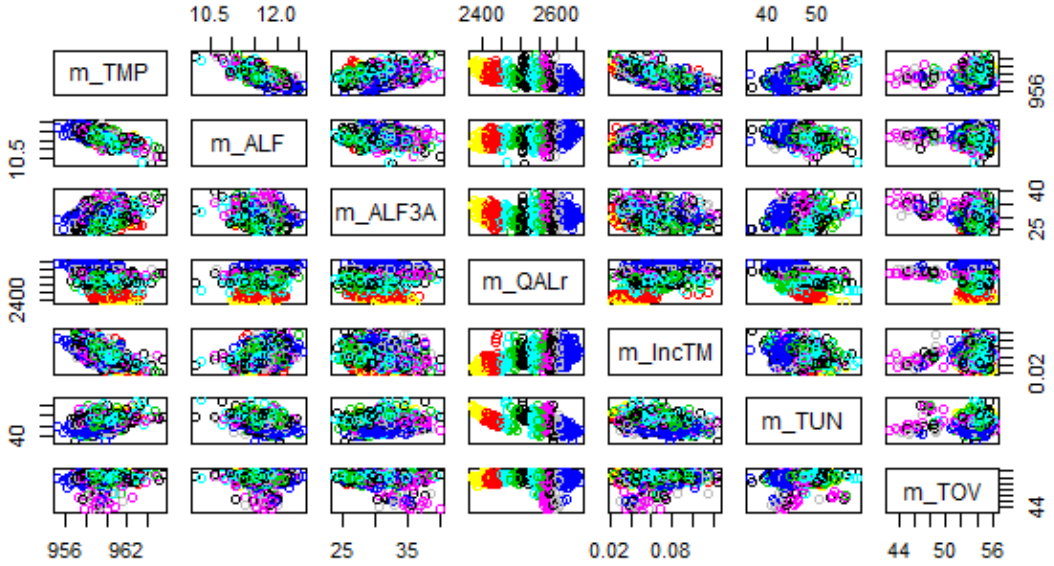


Figura 36 (a) Gráficos de cada experimento para comparar variáveis par a par de acordo com o agrupamento resultante com 13 *Clusters*. Experimento #1: Média com filtro.

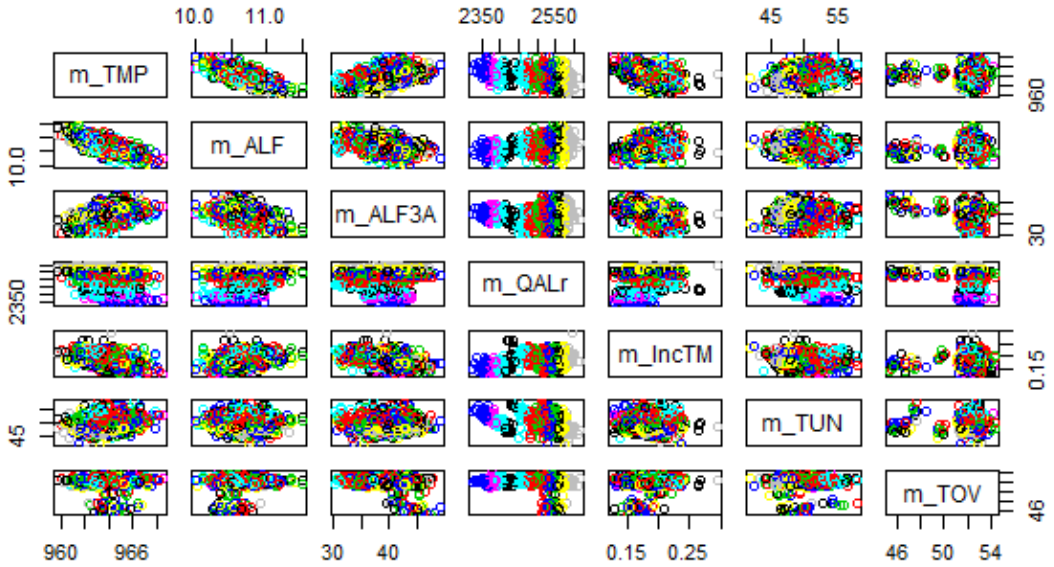


Figura 36 (b) Experimento #2: Média sem filtro.

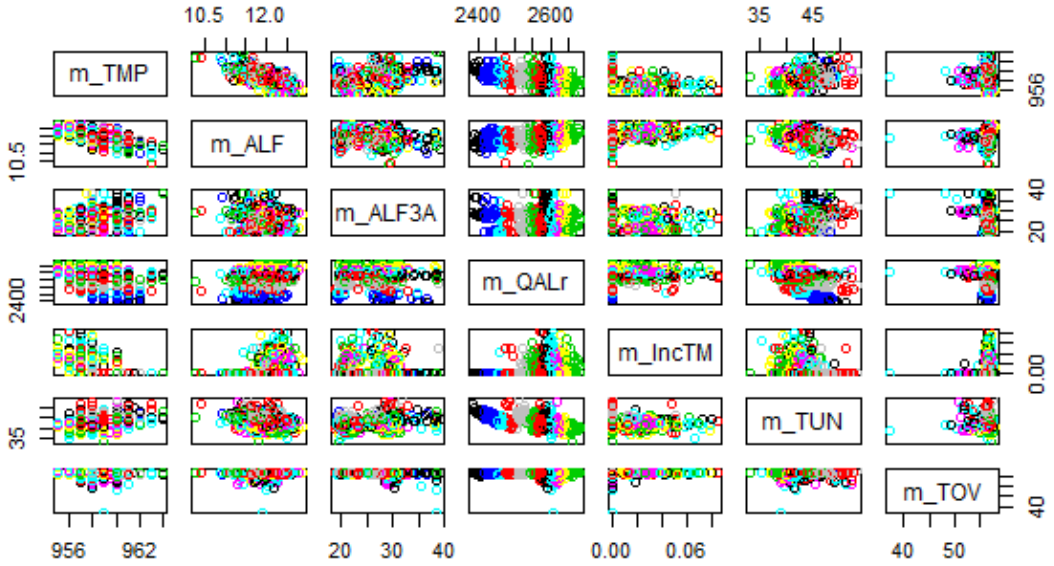


Figura 36 (c) Experimento #3: Mediana com filtro.

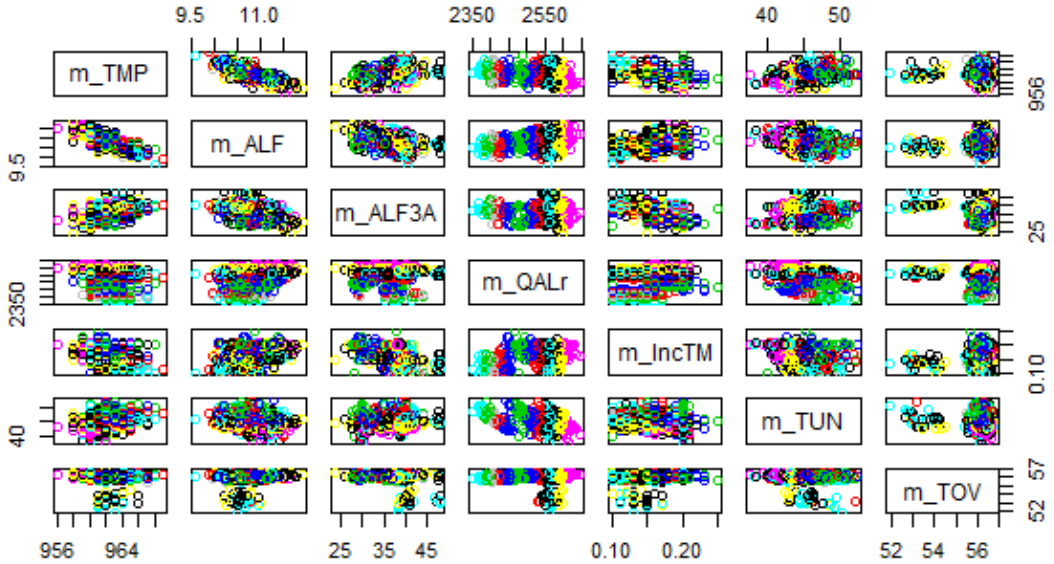


Figura 36 (d) Experimento #4: Mediana sem filtro.

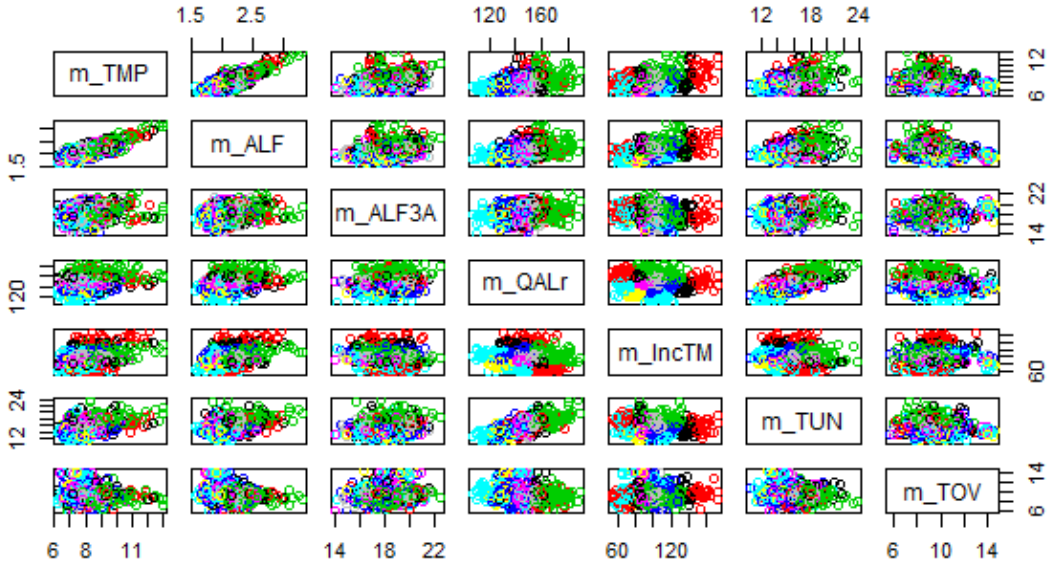


Figura 36 (e) Experimento #5: Desvio Padrão com filtro.

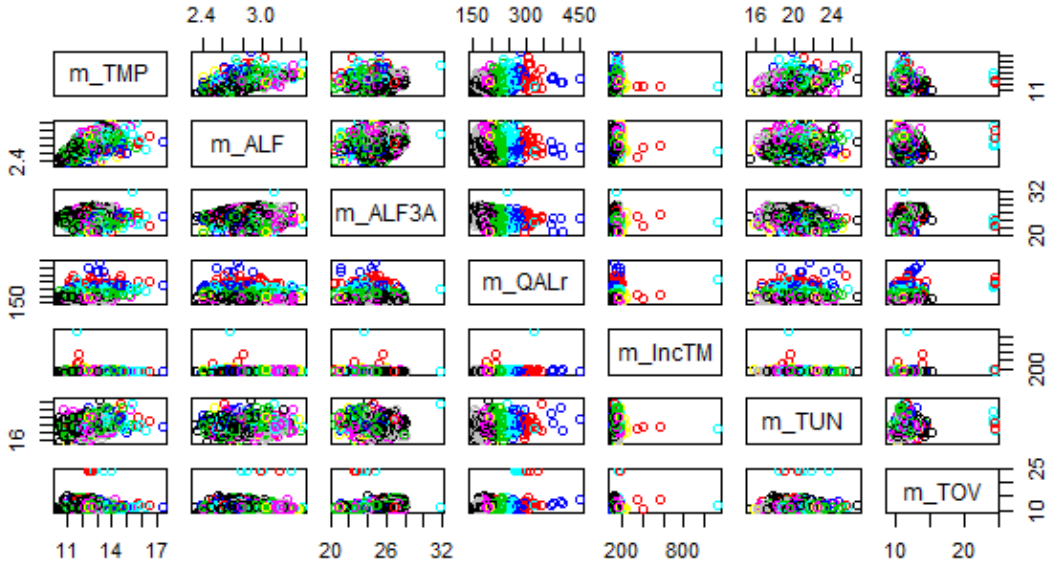


Figura 36 (f) Experimento #6: Desvio Padrão sem filtro.

As Figuras 37 (a) à 39 (f) trazem a análise grupo versus localização através de cores diferentes para cada agrupamento, através do algoritmo *K–Means*, para o experimento com medidas aritméticas de média, mediana e desvio padrão com filtro e sem filtro. Foram realizados agrupamentos para 2, 5 e 13 *clusters*, respectivamente.

Mapeamento *K–Means* com 2 *Clusters*:

Figura 37 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters*. Experimento #1: Média com filtro.

Figura 37 (b) Experimento #2: Média sem filtro.

Figura 37 (c) Experimento #3: Mediana com filtro.

Figura 37 (d) Experimento #4: Mediana sem filtro.

Figura 37 (e) Experimento #5: Desvio Padrão com filtro.

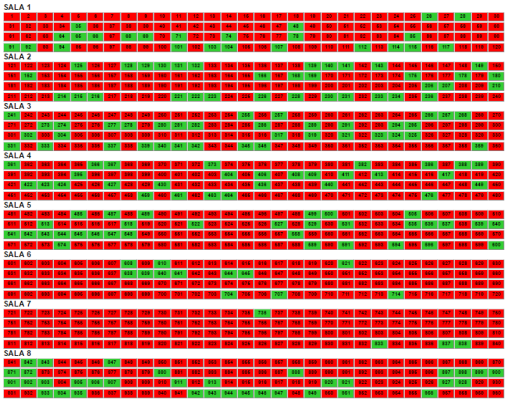


Figura 37 (f) Experimento #6: Desvio Padrão sem filtro.

Mapeamento *K–Means* com 5 *Clusters*:

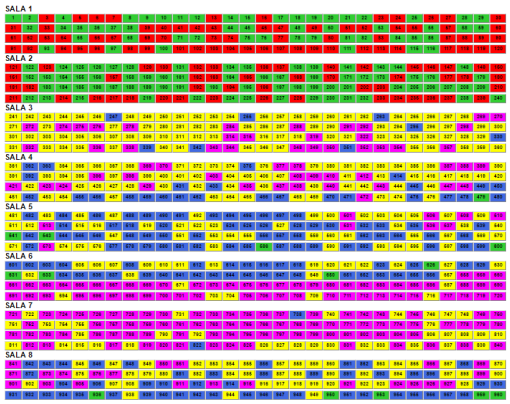


Figura 38 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters*. Experimento #1: Média com filtro.

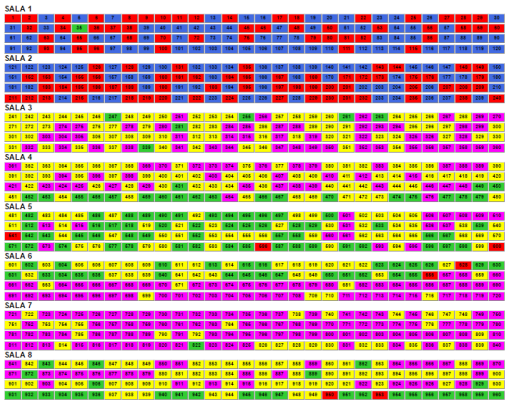


Figura 38 (b) Experimento #2: Média sem filtro.

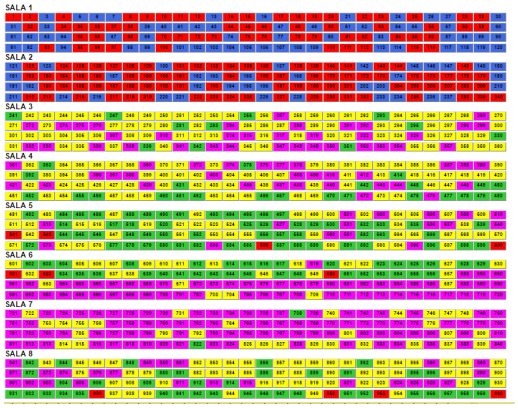


Figura 38 (c) Experimento #3: Mediana com filtro.

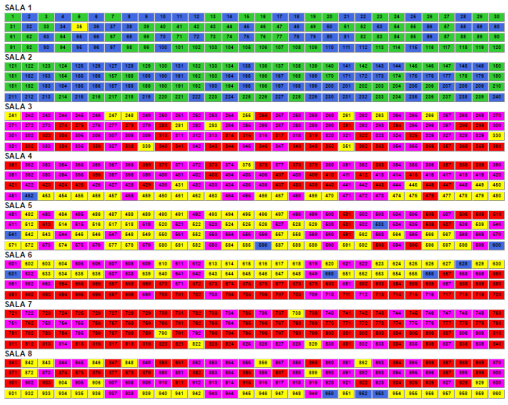


Figura 38 (d) Experimento #4: Mediana sem filtro.

Figura 38 (e) Experimento #5: Desvio Padrão com filtro.

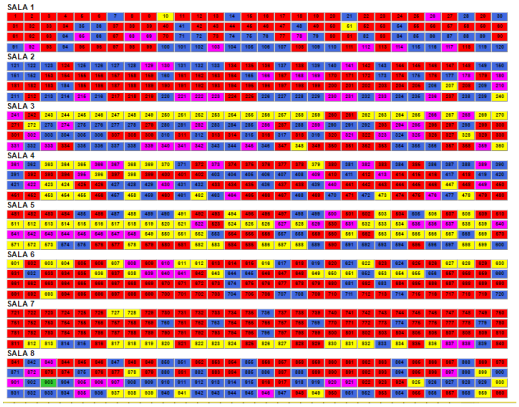


Figura 38 (f) Experimento #6: Desvio Padrão sem filtro.

Mapeamento *K–Means* com 13 *Clusters*:

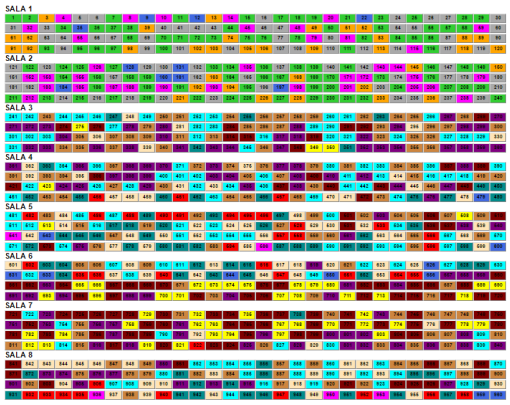


Figura 39 (a) Gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters*. Experimento #1: Média com filtro.

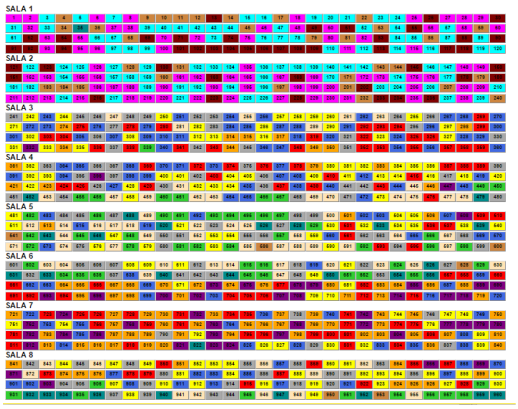


Figura 39 (b) Experimento #2: Média sem filtro.

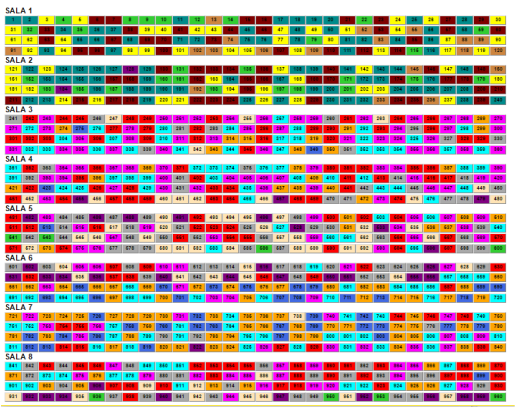


Figura 39 (c) Experimento #3: Mediana com filtro.

Figura 39 (d) Experimento #4: Mediana sem filtro.

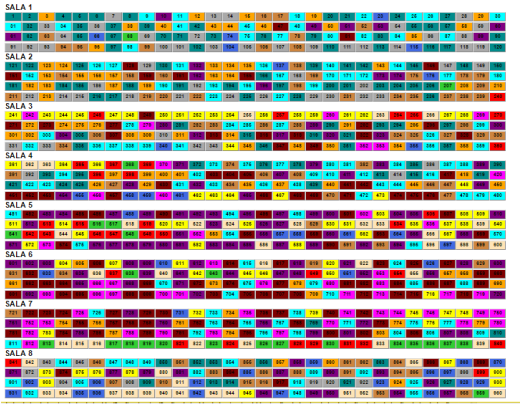


Figura 39 (e) Experimento #5: Desvio Padrão com filtro.

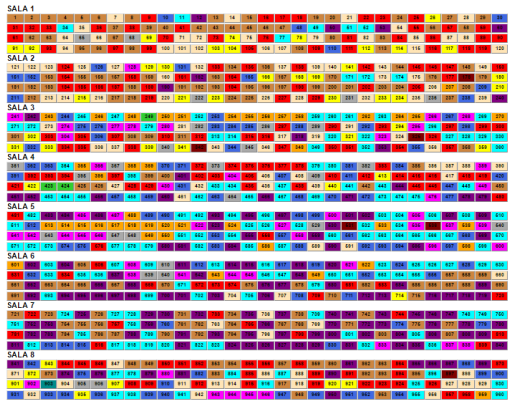


Figura 39 (f) Experimento #6: Desvio Padrão sem filtro.

Após a análise das Figuras 37 (a) à 39 (f) é possível verificar que a melhor medida para o uso com o K-Means se deu através das medidas de média e mediana com filtro e sem filtro com 5 *clusters*, pois em ambas ocorre a predominância de dois grupos nas duas primeiras salas dos experimentos, sendo que na média com filtro a predominância se deu através dos grupos vermelho e verde, na média sem filtro a predominância se deu através dos grupos azul e vermelho, na mediana com filtro a predominância se deu também através dos grupos vermelho e azul e por fim, na mediana sem filtro a predominância se deu através dos grupos verde e azul.

6.5.COMPARAÇÃO ENTRE OS RESULTADOS

Através dos resultados apresentados, foi possível notar que os métodos de agrupamentos utilizados podem ser considerados promissores para o estudo e a interpretação de bancos de dados, trazendo informações que a olho nu seria de difícil compreensão. O importante é conhecer suas propriedades, qualidades e deficiências, pois essas características irão ajudar na escolha daquele que melhor responde ao objetivo que se pretende alcançar.

A seguir se tem a comparação entre as técnicas de acordo com o número de *clusters*.

As Figuras 40 (a) à 40 (f) reportam os agrupamentos realizados pelo algoritmo *Affinity Propagation*, tendo que ser analisado de forma separada dos outros algoritmos por apresentar números de *clusters* diferentes em cada experimento realizado. As medidas aritméticas de média com filtro e sem filtro resultaram em um agrupamento com 6 *clusters*, representados pelas Figuras 40 (a) e 40 (b), respectivamente; já a mediana com filtro e sem filtro resultou em uma clusterização contendo 5 *clusters*, representados pelas Figuras 40 (c) e 40 (d), respectivamente; o desvio padrão com filtro obteve 13 *clusters*, representado pela Figura 40 (e); e o desvio padrão sem filtro obteve 19 *clusters*, representado pela Figura 40 (f).

As Figuras 41 (a) à 49 (f) trazem a comparação de todos os mapeamentos dos grupos versus sua localização, com cores diferentes para cada agrupamento, através dos algoritmos Kohonen (SOM), *Fuzzy C–Means* e *K–Means*, para experimentos com dados filtrados e não filtrados, bem como a utilização de propriedades aritméticas de média, mediana e desvio padrão com filtro e sem filtro. Foram comparados os agrupamentos para 2, 5 e 13 *clusters*, respectivamente.

Agrupamento utilizando a técnica do *Affinity Propagation*

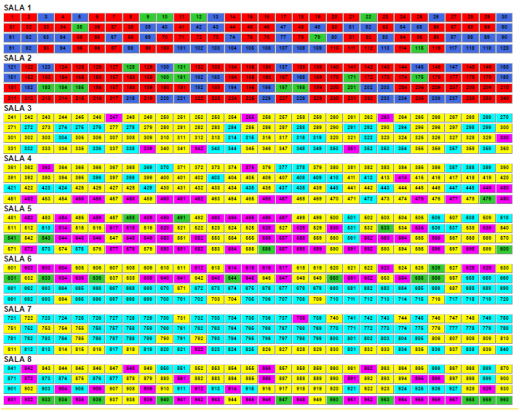


Figura 40 (a) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 6 *clusters*. Experimento #1: Média com filtro.

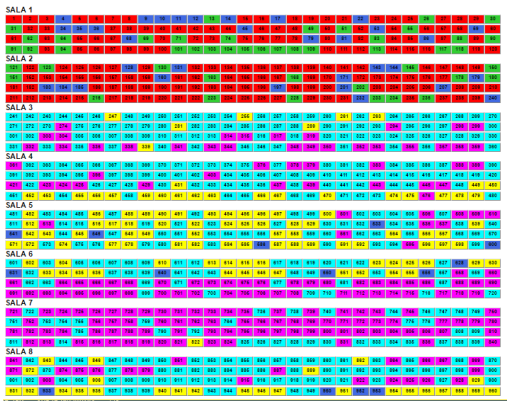


Figura 40 (b) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 6 *clusters*. Experimento #2: Média sem filtro.

Figura 40 (c) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 5 *clusters*. Experimento #3: Mediana com filtro.

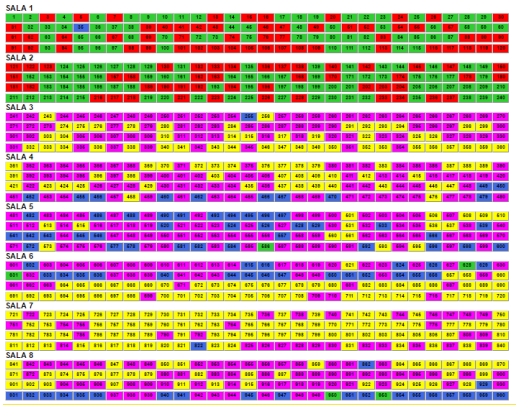


Figura 40 (d) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 5 *clusters*. Experimento #4: Mediana sem filtro.

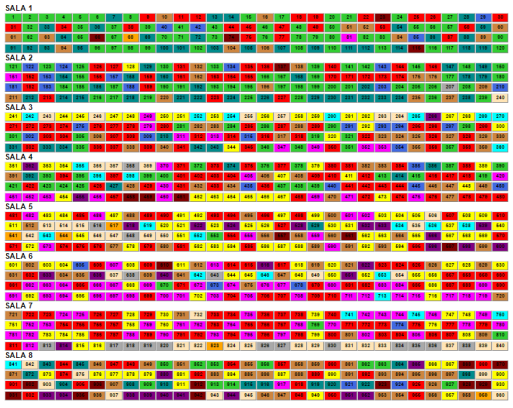


Figura 40 (e) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 13 *clusters*. Experimento #5: Desvio Padrão com filtro.

Figura 40 (f) – Gráficos de mapeamento grupo *versus* localização conseguida pelo algoritmo *Affinity Propagation*, resultando em 19 *clusters*. Experimento #6: Desvio Padrão sem filtro.

Agrupamento com 2 *Clusters*

Média com Filtro

Figura 41 (a) Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* – Média com Filtro.

Média sem Filtro

Figura 41 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* – Média sem Filtro.

Figura 41 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* – Média com Filtro.

Figura 41 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* – Média sem Filtro.

Figura 41 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters* – Média com Filtro.

Figura 41 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters* – Média sem Filtro.

Mediana com Filtro

Figura 42 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* – Mediana com Filtro.

Mediana sem Filtro

Figura 42 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* Mediana sem Filtro.

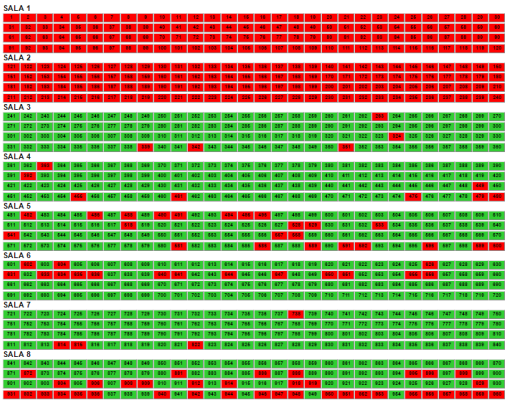


Figura 42 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* Mediana com Filtro.

Figura 42 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* Mediana sem Filtro.

Figura 42 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters* Mediana com Filtro.

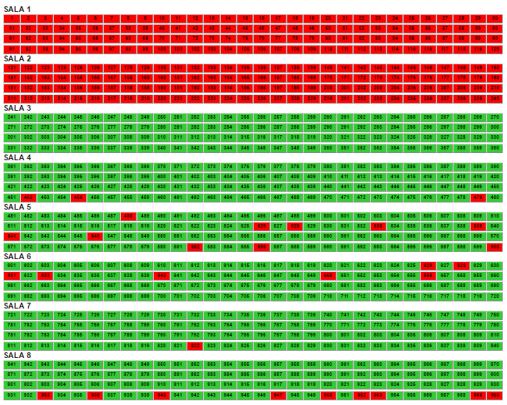


Figura 42 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters* Mediana sem Filtro.

Desvio Padrão com Filtro

Figura 43 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* – Desvio Padrão com filtro.

Desvio Padrão sem Filtro

Figura 43 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 2 *Clusters* – Desvio Padrão sem filtro.

Figura 43 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* – Desvio Padrão com filtro.

Figura 43 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 2 *Clusters* – Desvio Padrão sem filtro.

Figura 43 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters* – Desvio Padrão com filtro.

Figura 43 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 2 *Clusters*– Desvio Padrão sem filtro.

Agrupamento com 5 *Clusters*

Média com Filtro

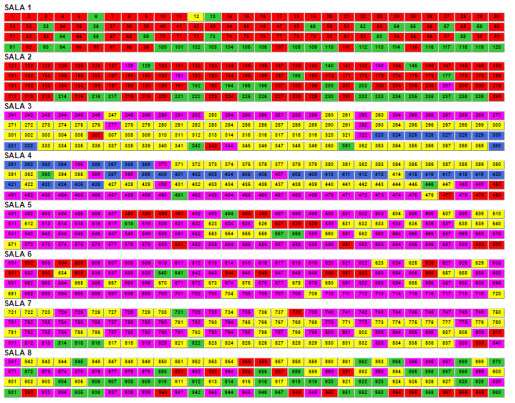


Figura 44 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters*– Média com Filtro.

Média sem Filtro

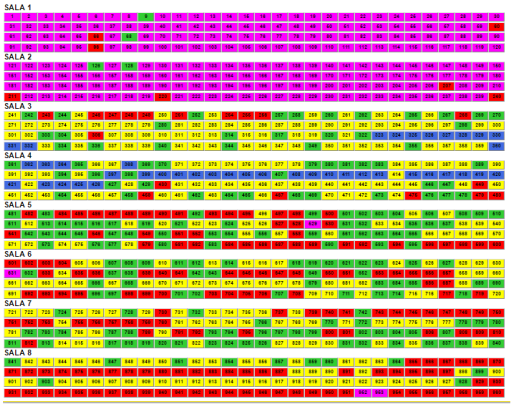


Figura 44 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters* – Média sem Filtro.

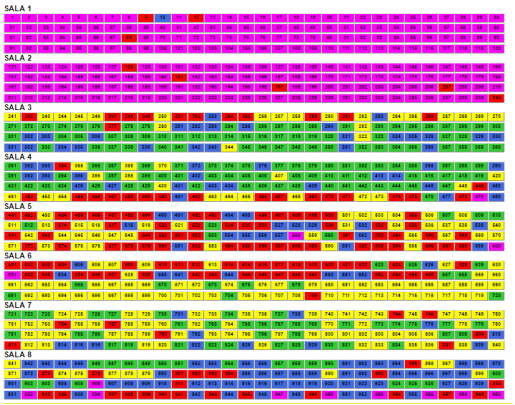


Figura 44 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Média com Filtro.

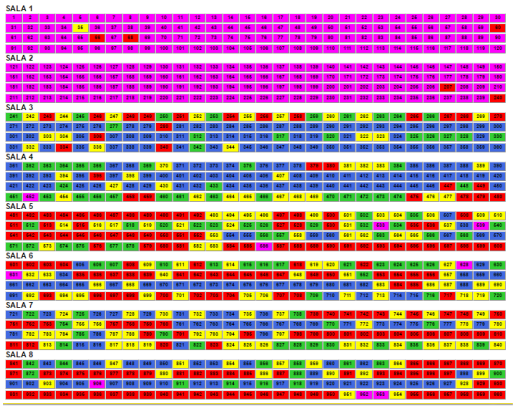


Figura 44 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Média sem Filtro.

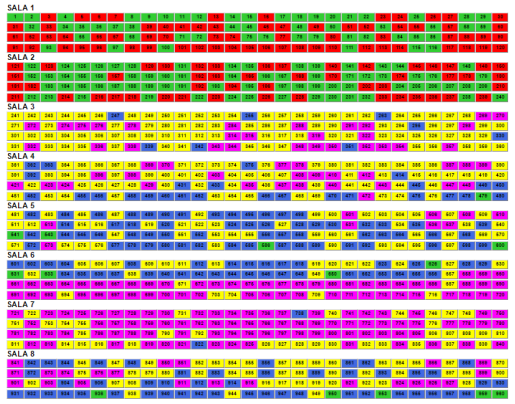


Figura 44 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters* – Média com Filtro.

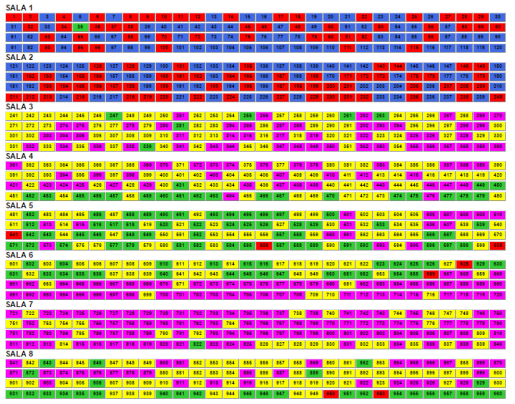


Figura 44 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters* – Média sem Filtro.

Mediana com Filtro

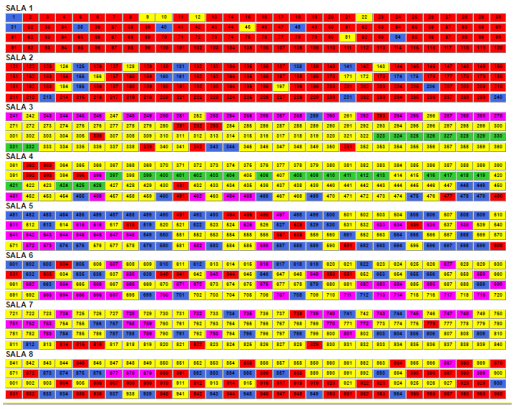


Figura 45 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters* – Mediana com Filtro.

Mediana sem Filtro

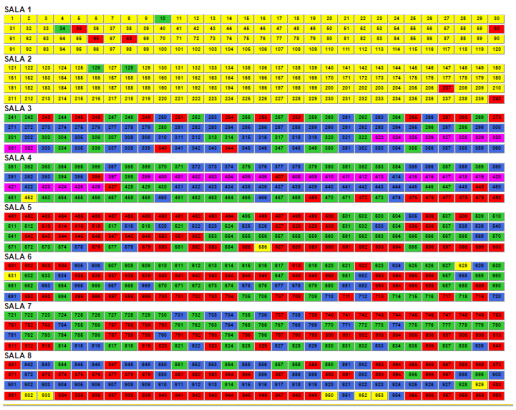


Figura 45 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters* – Mediana sem Filtro.

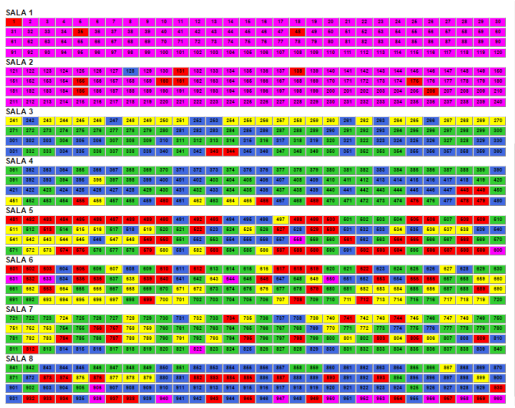


Figura 45 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Mediana com Filtro.

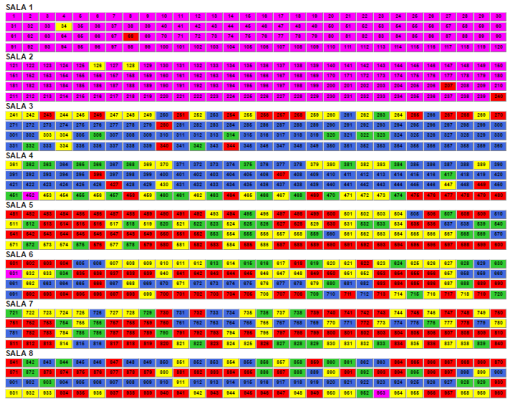


Figura 45 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Mediana sem Filtro.

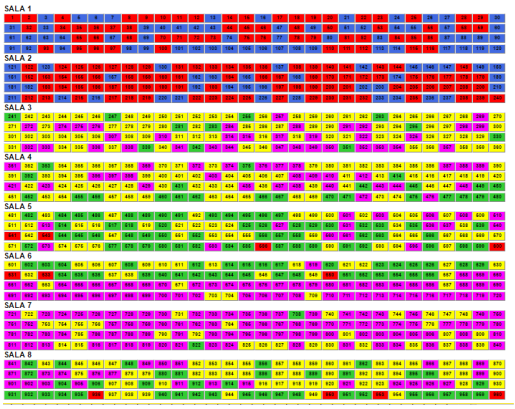


Figura 45 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters*– Mediana com Filtro.

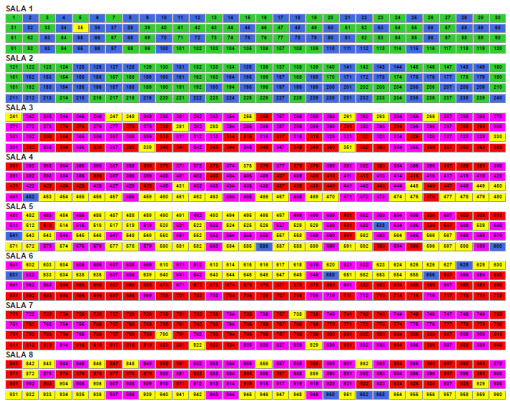


Figura 45 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters*– Mediana sem Filtro.

Desvio Padrão com Filtro

Figura 46 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters* – Desvio Padrão com filtro.

Desvio Padrão sem Filtro

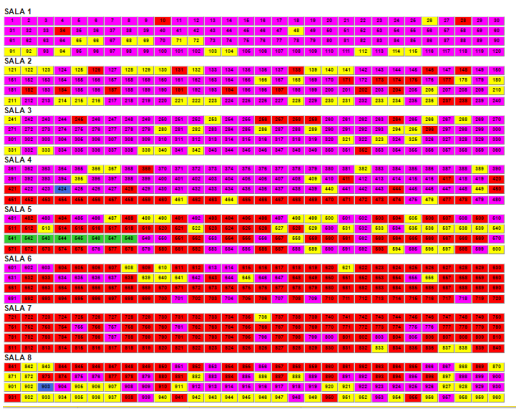


Figura 46 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 5 *Clusters* – Desvio Padrão sem filtro.

Figura 46 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Desvio Padrão com filtro.

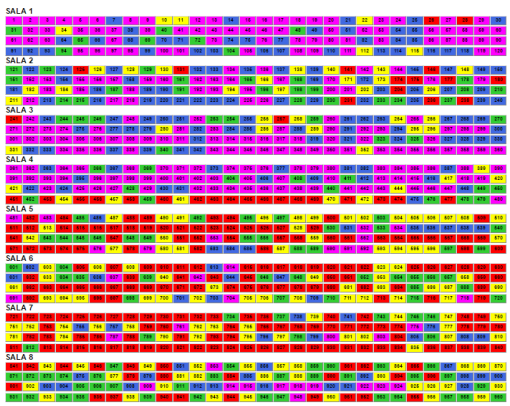


Figura 46 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 5 *Clusters* – Desvio Padrão sem filtro.

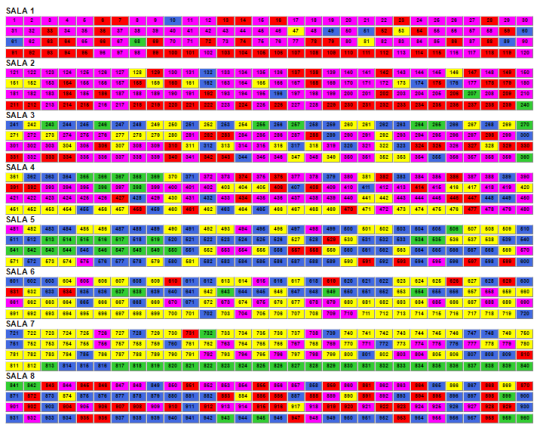


Figura 46 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters* – Desvio Padrão com filtro.

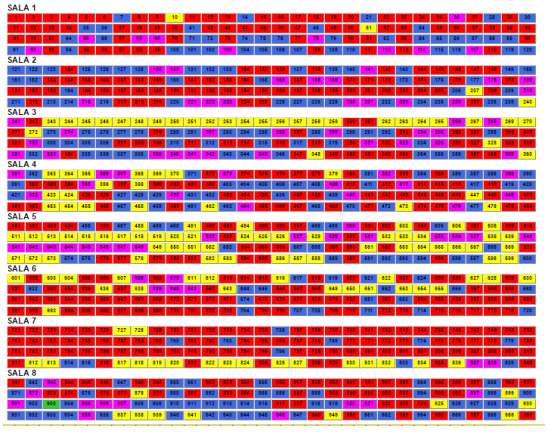


Figura 46 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 5 *Clusters* – Desvio Padrão sem filtro.

Agrupamento com 13 *Clusters*

Média com Filtro

Figura 47 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters*– Média com Filtro.

Média sem Filtro

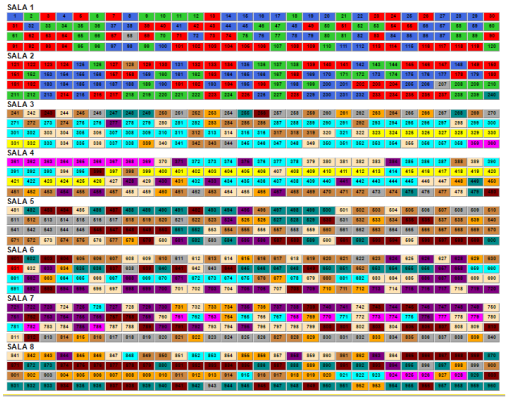


Figura 47 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters*– Média sem Filtro.

Figura 47 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters*– Média com Filtro.

Figura 47 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters*– Média sem Filtro.

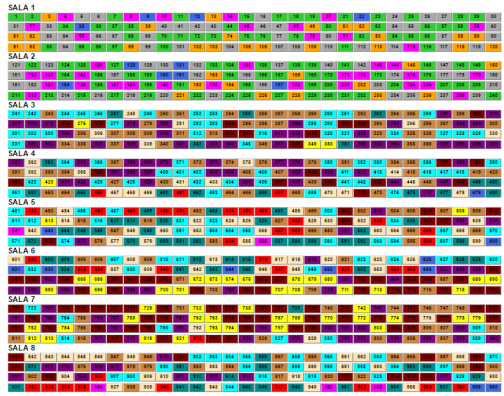


Figura 47 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Média com Filtro.

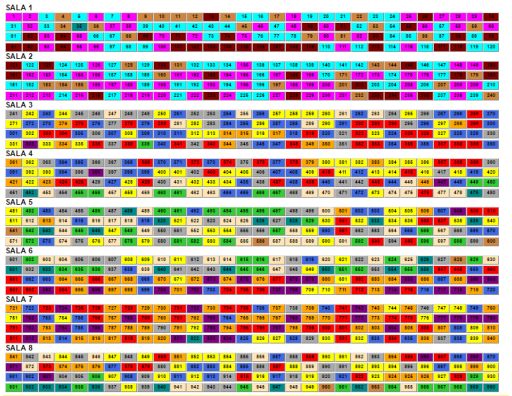


Figura 47 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Média sem Filtro.

Mediana com Filtro

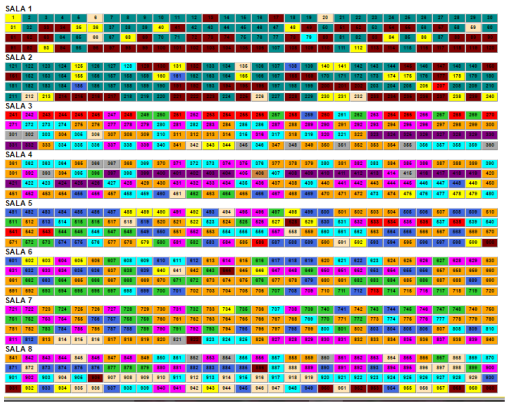


Figura 48 (a) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters* – Mediana com Filtro.

Mediana sem Filtro

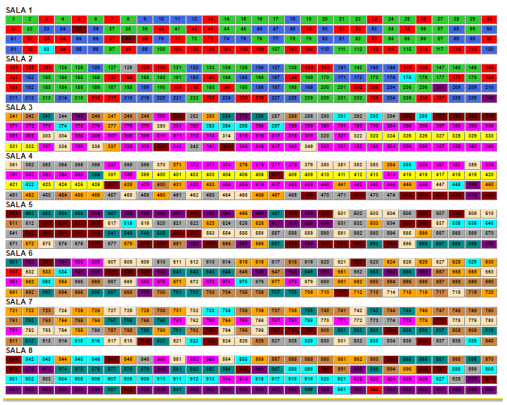


Figura 48 (b) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters* – Mediana sem Filtro.

Figura 48 (c) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters* – Mediana com Filtro.

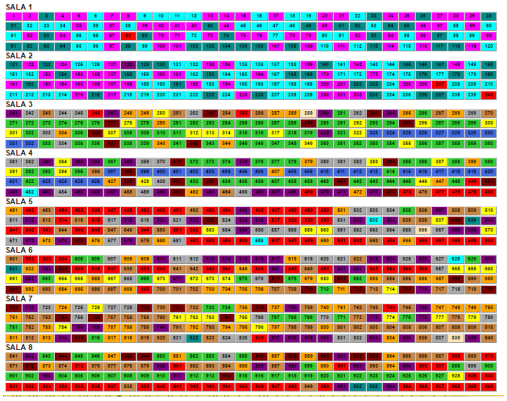


Figura 48 (d) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters* – Mediana sem Filtro.

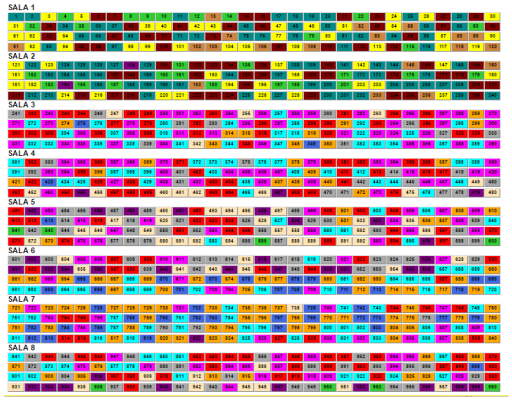


Figura 48 (e) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Mediana com Filtro.

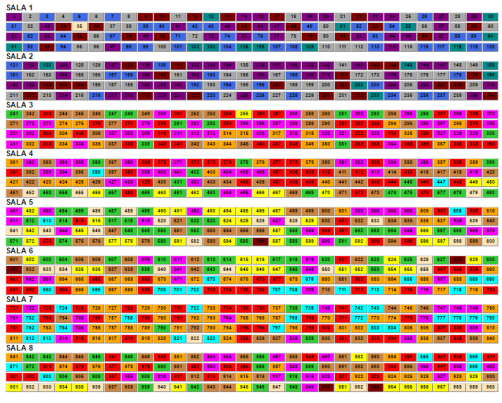


Figura 48 (f) – Comparação entre os gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Mediana sem Filtro.

Desvio Padrão com Filtro

Figura 49 (a) – Gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters* – Desvio Padrão com Filtro.

Desvio Padrão sem Filtro

Figura 49 (b) – Gráficos de mapeamento grupo *versus* localização do algoritmo SOM para 13 *Clusters* – Desvio Padrão sem Filtro.

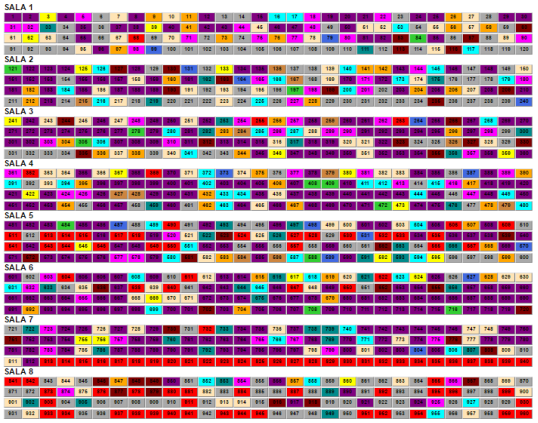


Figura 49 (c) – Gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters* – Desvio Padrão com Filtro.

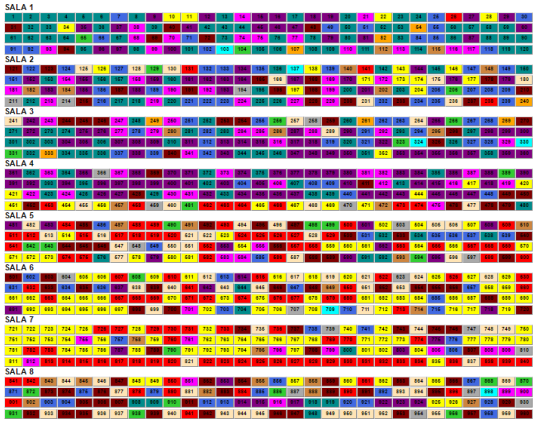


Figura 49 (d) – Gráficos de mapeamento grupo *versus* localização do algoritmo FCM para 13 *Clusters* – Desvio Padrão sem Filtro.

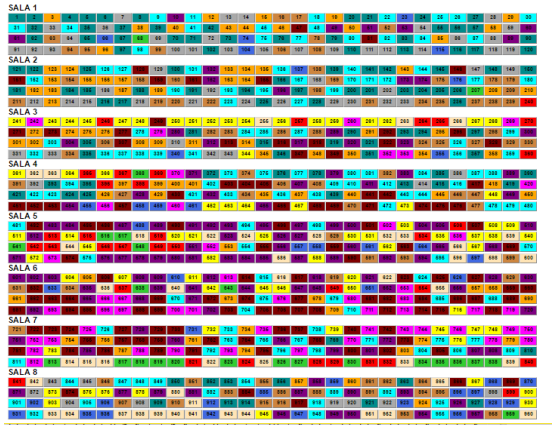


Figura 49 (e) – Gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Desvio Padrão com Filtro.

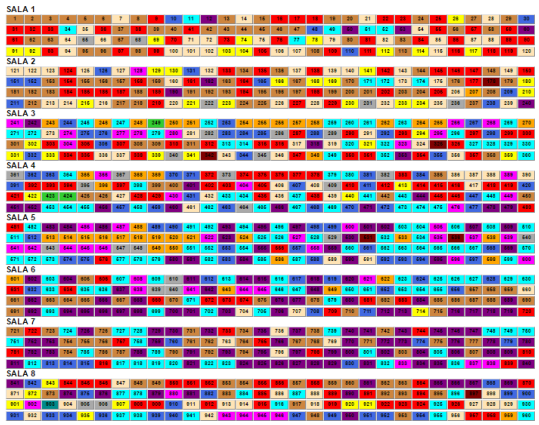


Figura 49 (f) – Gráficos de mapeamento grupo *versus* localização do algoritmo *K–Means* para 13 *Clusters* – Desvio Padrão sem Filtro.

A partir dos resultados obtidos é possível notar que no algoritmo *Affinity Propagation*, as Figuras 10 (a) a 10 (d) resultaram em uma predominância de dois grupos nas duas primeiras salas, enquanto que as Figuras 10 (e) e 10 (f) as quais representam o desvio padrão resultaram em uma mistura de grupos em todas as salas.

No algoritmo Kohonen (SOM), as Figuras 18 (a), 18 (b), 18 (c) e 18 (d), as quais representam média com filtro, média sem filtro, mediana com filtro e mediana sem filtro com 2 *clusters*, respectivamente, demonstram uma predominância de um único grupo nas duas primeiras salas, sendo que nas Figuras 18 (a) e 18 (c) (média e mediana com filtro) a predominância se deu com o grupo vermelho, e nas Figuras 18 (b) e 18 (d) (média e mediana sem filtro) a predominância se deu com o grupo verde. As Figuras 18 (e) e 18 (f) que representam o desvio padrão com filtro e sem filtro houve uma mistura de grupos. As Figuras 19 (a) a 19 (f) representaram o agrupamento do algoritmo Kohonen para 5 *clusters*, onde as Figuras 19 (a) a 19 (d) apresentaram as duas primeiras salas com um grupo predominante, enquanto que as Figuras 19 (e) e 19 (f) apresentaram uma mistura de grupos em todas as salas analisadas. Para o agrupamento do algoritmo Kohonen com 13 *clusters* todas as salas das Figuras 20 (a) até 20 (f) apresentaram uma mistura de grupos, mostrando que ao contrário do algoritmo *Affinity Propagation*, o melhor agrupamento do algoritmo Kohonen não se deu com 13 *clusters* e sim de 2 a 6 *clusters*.

No algoritmo *Fuzzy C–Means*, as Figuras 26 (a) a 26 (d) e 27 (a) a 27 (d), que representam a média e mediana com filtro e sem filtro com 2 e 5 *clusters*, respectivamente, apresentaram as duas primeiras salas com a predominância de um grupo, enquanto que as Figuras 26 (e) e 26 (f) e 27 (e) e 27 (f), as quais representam o desvio padrão com filtro e sem filtro com 2 e 5 *clusters*, respectivamente, apresentaram uma mistura de grupos em todas suas salas.

Assim como no algoritmo Kohonen, no algoritmo FCM o agrupamento com 13 *clusters* todas as salas das Figuras 28 (a) até 28 (f) apresentaram uma mistura de grupos, mostrando que também não tiveram sua melhor clusterização com 13 *clusters*, sendo seus melhores agrupamentos de 2 a 5 *clusters*.

No algoritmo *K–Means*, as Figuras 37 (a) a 37 (d) apresentam suas duas primeiras salas com a predominância de um grupo, representado pela cor vermelha, e as quatro salas restantes com a predominância do outro grupo, representado pela cor verde. Já a Figura 37 (e) possui as três primeiras salas com a predominância do grupo verde, e por fim, a Figura 37 (f) apresenta uma predominância do grupo vermelho em todas suas salas. Para o agrupamento

com 5 *clusters* no algoritmo *K–Means*, as Figuras 38 (a) a 38 (d) apresentaram a predominância de dois grupos nas duas primeiras salas, e as Figuras 38 (e) e 38 (f) apresentaram uma mistura de grupos em todas as salas.

Para o agrupamento com 13 *clusters* no *K–Means* todas as salas das Figuras 39 (a) até 39 (f) apresentaram uma mistura de grupos, mostrando que assim como no Kohonen e no *Fuzzy C–Means*, o algoritmo *K–Means* também não apresentou seu melhor agrupamento com 13 *clusters*, e sim com 2 a 4 *clusters*.

No agrupamento do *Affinity Propagation* expostos nas Figuras 40 (a) à 40 (f) é possível notar um bom agrupamento na medida de mediana com filtro disposta na Figura 40 (c), agrupando 5 *clusters* e tendo uma predominância do grupo verde nas duas primeiras salas.

Através das Figuras 41 (a) à 43 (f) é possível observar que para o agrupamento com 2 *clusters*, os melhores agrupamentos dos algoritmos SOM, FCM e *K–Means* se dão através da média e mediana, por apresentarem as duas primeiras salas com uma predominância de grupos, enquanto que no desvio padrão há uma mistura de grupos entres todas as salas.

Para os agrupamentos com 5 e 13 *clusters*, as duas primeiras salas dos agrupamentos utilizando a média e mediana com filtro e sem filtro começam a se dividir em dois ou mais *clusters*, e o desvio padrão permanece com o comportamento de mistura de grupos em todas as salas.

No agrupamento com o algoritmo *Affinity Propagation*, o melhor resultado se deu através da mediana com filtro, contendo 5 *clusters*. Já no algoritmo do mapa auto-organizável de Kohonen, os melhores agrupamentos se deram através da medida aritmética de média e mediana com filtro e sem filtro para 2 e 5 *clusters*, pois nota-se o predomínio de um grupo nas duas primeiras salas desses experimentos. No experimento do algoritmo *Fuzzy C-Means* percebe-se que os melhores resultados também foram com a média e mediana com filtro e sem filtro para 2 e 5 *clusters*, onde também há a predominação de um grupo nas duas primeiras salas dos experimentos. E no agrupamento dos fornos através da técnica do *K-Means*, observa-se que os melhores resultados vieram da média e mediana com filtro e sem filtro na análise com 5 *clusters*, onde houve a dominância de dois grupos nas duas primeiras salas dos fornos estudados.

Através das análises realizadas percebe-se que a medida do desvio padrão não consegue descrever o desempenho dos fornos estudados, já que as características do processo não têm muita variância, onde todas possuem o desvio padrão bem parecidos, dificultando assim o processo de agrupamento.

7. CONCLUSÕES E PROPOSTAS PARA TRABALHOS FUTUROS

Técnicas de agrupamento têm sido utilizadas com sucesso em várias áreas, especialmente naquelas em que não há conhecimento prévio sobre a organização dos dados.

A dissertação apresentada contribuiu com o desenvolvimento de soluções de qualidade para o agrupamento dos fornos de redução de uma fábrica de alumínio, através da qual foi possível mostrar o desempenho dos algoritmos *Affinity Propagation*, do Mapa Auto–Organizável de Kohonen (SOM), do *Fuzzy C–Means* (FCM) e do *K–Means* na tarefa de realizar agrupamentos com dados reais, de acordo com as características de cada técnica utilizada, e com a finalidade de encontrar o melhor número de *clusters*, realizando experimentos com média, mediana e desvio padrão com filtro e sem filtro de 2 a 13 *clusters*, respectivamente.

A análise estatística dos dados reais foi realizada através de um pré-processamento de dados do período de 2006 até 2012, disponibilizados por uma indústria de alumínio, que contém 8 salas com 120 fornos cada uma, totalizando 960 fornos de redução, onde através da filtragem dos dados destes fornos foi possível selecionar a melhor faixa de registros das variáveis de entrada e saída, eliminando assim os dados considerados espúrios ou *outliers*, que são os dados que apresentam características distintas das demais do grupo.

Através dos experimentos realizados é factível notar que as duas primeiras salas da fábrica de alumínio possuem seus fornos com características semelhantes, pois nota-se que em todas as técnicas utilizadas as melhores clusterizações houve a predominância de um grupo nas duas primeiras salas.

A qualificação dos agrupamentos foi realizada através de critérios de validação, onde envolveram índices que afirmam a qualidade dos dados distribuídos entre os grupos. Esta validação do algoritmo de agrupamento (*clustering*) foi estimada por um critério objetivo para determinar quão boa é a partição gerada pelo algoritmo, e estes critérios são importantes porque permitem comparar os resultados de diversos algoritmos e consentem em determinar o melhor número de *clusters*.

Além do agrupamento, foi possível identificar importantes regras que podem também servir para treinar novos engenheiros, bem como pessoas envolvidas no processo do funcionamento do forno em questão, para que ao invés dos trabalhadores aprenderem a manusear o forno presencialmente, os mesmos possam ter o primeiro contato de forma virtual, através de um programa computacional, no qual sejam feitas várias simulações e testes,

evitando que haja danos aos fornos e que acidentes não ocorram com os operadores dos mesmos.

A partir da análise dos experimentos realizados se conclui que não é possível descrever qual técnica possui o melhor agrupamento, já que os algoritmos do *Affinity Propagation*, do Mapa Auto–Organizável de Kohonen, do *Fuzzy C–Means* e do *K–Means* possuem métodos diferentes de agrupamentos, com suas características e peculiaridades diferenciadas uma das outras. O que se pode inferir através dos experimentos realizados é que a medida que o número de *clusters* aumenta, a clusterização começa a ficar mais dividida, principalmente nas duas primeiras salas dos fornos, onde até 4 *clusters* possuem uma predominância de um único grupo, e a medida que essa clusterização aumenta para até 13 *clusters* as duas primeiras salas passam a ter mais de um grupo predominante.

Com base em todos os experimentos realizados foi possível notar que a medida do desvio padrão tanto com filtro quanto sem filtro possui uma conduta com mistura de grupos em todas as salas, não conseguindo discriminar o comportamento dos fornos de redução, ou seja, as variáveis do processo não possuem muitas mudanças, tendo todas o desvio padrão semelhante, o que dificulta o agrupamento. De modo geral, isso ocorre devido as variáveis do forno serem bem controladas, não indo muito distante dos valores de média. Quando se usa a medida de desvio-padrão, todos os dados ficam muito próximos, ou seja, sem diferenças significativas, logo essa medida não serve para dizer quão um forno é diferente do outro.

Através dos experimentos realizados foi possível notar que quando se utiliza o RStudio para gerar o experimento da rede Kohonen, o mesmo gera o gráfico da Figura 14, o qual representa os grupos de fornos e respectivos níveis de influência, onde o resultado é a saída do processo da construção de uma rede neural, e ao final é necessário que aja uma interpretação de forma visual dos resultados obtidos. Já o *Affinity Propagation*, o FCM e o *K-Means* resultam outros tipos de gráficos que não permitem fazer a interpretação e análise das regras, mesmo que visualmente, pois não geram gráficos de saída como os gerados no experimento da rede Kohonen, sendo proposta para trabalhos futuros o estudo e interpretação da forma com que se constitui as regras dos algoritmos *Affinity Propagation*, *Fuzzy C-Means* e *K-Means*.

Também como proposta para trabalhos futuros tem–se o estudo de outras variáveis para a verificação de seu comportamento dentro da clusterização dos algoritmos estudados, bem como a inserção de outros algoritmos de agrupamento e a variação do número de *clusters*.

REFERÊNCIAS BIBLIOGRÁFICAS

AAKER, D. A.; KUMAR, V.; DAY, G. S. **Pesquisa de marketing**. São Paulo: Atlas, página 745, 2001.

AGRESTI, A. **Categorical data analysis**. 3° Edição. Nova Jersey: John Wiley & Sons Inc, 2012.

ARGENTINA. **Sector de la Industria del Aluminio y sus manufacturas**. *Dirección de Oferta Exportable Dirección General de Estrategias de Comercio Exterior Subsecretaría de Comercio Internacional*. Ministério de Relações Exteriores, Comércio Internacional e Cultural, páginas 4 – 9, 2010.

_____, **Manual de Fundamentos do Processo de Produção de Alumínio Parte II**, 1999.

BANERJEE, Anoopam; GRUPTA, Preeti. **Extension to Alpha Algorithm for Process Mining**. *International Journal of Engineering & Computer Science*, 2015.

BARROSO, L. P.; ARTES, R. **Análise de Multivariada**. Lavras: UFLA, 157p., 2003.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. *Kluwer Academic Publishers*, Norwell, MA, USA, 1981.

BODENHOFER, U.; PALME, J. ; MELKONIAN, C.; KOTHMEIER, A. **APCluster – An R Package for *Affinity Propagation* Clustering**. Versão 1.4.4, 1 de Julho de 2017. Disponível em <http://www.bioinf.jku.at/software/ap*cluster*/AP*Cluster*–Manual.pdf>. Último acesso Julho/2017.

BRAMER, M. **Undergraduate Topics in Computer Science – Principles of *Data mining***. Springer, 2007.

BRENDAN, J. F.; DUECK, D. **Clustering by passing messages between data points**. Science. 315 (5814):972–976. doi: 10.1126/science.1136800. PMID 17218491, 2007.

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. (2008). **clValid: Validation of Clustering Results**. Disponível em <http://www.louisville.edu/~g0broc01/research>. Último acesso Dezembro/2016.

BRUZOS, T.; BRUZOS, D. Sabelotodo.org. **Produción del alumínio**. Disponível em: <http://www.sabelotodo.org/metalurgia/produccionalumínio.html>. Último acesso Janeiro/2017.

BUSSAB, W. DE O; MIAZAKI, E. S; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, página 105, 1990.

CAMBRIA, E. et. al. **New avenues in opinion mining and sentimento analysis**. *IEEE Intelligent Systems*, v. 28, n. 2, páginas 15–21, 2013.

CAMILO, C. O.; SILVA, J. C. DA. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Universidade Federal de Goiás. Instituto de Informática, 2009.

CARDOSO, J. M. M.; MADEIRA, J. A. G. **Uso de técnicas de geração de agrupamentos para determinação no número de classes dos portos que movimentam granéis sólidos**. In: XLII SBPO, 2010, Bento Gonçalves. Anais... Bento Gonçalves, 2010.

CARNERO, M.; HERNÁNDEZ, J.; SÁNCHEZ, M. **A new metaheuristic based approach for the design of sensor networks**. *Computers and Chemical Engineering*; vol. 55, páginas 83–96, 2013.

CARVALHO, L. A. V. DE. **Datamining: A Mineração de Dados No Marketing, Medicina, Economia, Engenharia e Administração**. Rio de Janeiro: Editora Ciência Moderna, 2005.

CASTRO, L. N. DE; FERRARI, D.G. **Introdução à Mineração de Dados: Conceitos básicos, algoritmo e aplicações**. 1° Edição. São Paulo: Editora Saraiva, 2016.

CÉSAR, Perez Lópes. **Minería de datos: Técnicas y herramientas**. España: Thomson Editores, 2007.

CLIFFORD, H. T.; STEPHENSON, W. **An introduction to numerical taxonomy**. London: *Academic Press*, página 229, 1975.

CVG VENALUM (2008). **Desarrollo endogeno y responsabilidad social**. Disponível em: <http:// www.venalum.com.ve/Aluminio_ID/ desarrollo_endogeno/Aluminio__Desarrollo_ Endogeno__Responsabilidad_Social.pdf>. Último acesso Julho/2017.

DÍAZ, B. D.; MORILLAS, A. R. **Minería de Datos y Lógica Difusa. Una aplicación al studio de la rentabilidad económica de las empresas agroalimentarias en Andalucía**. *Estadística Española*. Vol. 46, Núm. 157, páginas 409–430, 2004.

DONI, M. V. **Análise de *Cluster*: métodos hierárquicos e de partição**. São Paulo: Mackenzie: 2004. 93f. Monografia (Pós–graduação) – Universidade Presbiteriana Mackenzie, 2004.

DUARTE, M. C.; SANTOS, J. B.; MELO, L. C. **Comparison of similarity coefficients based on RAPD markers in the common bean**. *Genetics and Molecular Biology*, v.22, n.3, páginas 427–432, 1999.

DUDA, R. O.; HART, P. E.; STORK, D. **Pattern Classification.** *Wiley series in Probabilistic and Statistic, John Wiley and Sons*, 2° Edição, 2001.

EVERITT, B. **Cluster analysis**. London: *Heinemann Educational Books*, página 136, 1974.

EVERITT, B.; S, LANDAU, S.; LEESE, M. **Cluster Analysis**. 4º ed. London: Arnold, página 207, 2001.

FROST, F.; KARRI, V. **Performance Comparison of BP and GRNN Models of Neural Networks Paradigm Using a Pratical Industrial Application**. *IEEE Transactions*, páginas 1069 – 1074, Jun/1999.

GADA, V.; DHAKRAS, A.; DEULKAR, K. **Forecasting of Aluminium Prices using Data Mining Techniques**. *International Journal of Innovations & Advancement in Computer Science* – IJIACS, ISSN 2347 – 8616. Volume 4, Issue 9, 2015.

GEORGE, J. K.; YUAN, B. **Fuzzy Sets and Fuzzy Logic theory and applications**. *New Jersey: Prentice Hall*, páginas 357–362, 1995.

GOLDSCHMIDT, R.; PASSOS, E. ***Data mining*: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações**. São Paulo: Elsevier; 2005.

GOMES, L. F. A. M. **Teoria da Decisão**. São Paulo: Editora Thompson. Página 116. (Coleção Debates em Administração), páginas 1–20. ISBN 85–221–0529–4, 2007.

GOWER, J. C.; LEGENDRE, P. **Metric and euclidean properties of dissimilarity coefficients**. *Journal of Classification*, v. 3, páginas 5 – 48, 1986.

GRJOTHEIM, K.; KVANDE, H. **Introduction to Aluminium Electrolysis Understanding the Hall–Héroult Process**. *Aluminium*–Verlag, 2ª edição, 1993.

HAN, J.; KAMBER, M.; PEI, J. ***Data mining*: Concepts and techniques**. 3° Edição. São Francisco: *Morgan Kaufmann*, 2011.

HARDING, J. A. (2008). **A *Data mining* integrated architecture for shop floor control.** *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 222 (5), páginas 605 – 624.

HAUPIN, W.; KVANDE, H. **Mathematical Model of fluoride evolution from Hall Heroult Cells**. *TMS Light Metals*, páginas 257–263, 1993.

HEATON, J. **Introduction to the math of neural networks**. *St. Louis: Heaton Research Inc.*, 2012.

IAI. 2010. **International Aluminium Institute**. Disponível em: <http://www.world-aluminium.org/ 191 prospectiva de la industria del aluminio en Venezuela y su rol en la construcción de futuro sostenible. pp. 175-191 3(1)2011 / ENERO-DICIEMBRE About+Aluminium/Story+of > Último acesso Julho/2017.

JACKSON, A. A.; SOMERS, K. M.; HARVERY, H. H. **Similarity coefficients: measures for co–occurrence and association or simply measures of occurrence?** *American Naturalist*, v.133, páginas 436–453, 1989.

JANTZEN, J. **Neurofuzzy Modellin**. *Technical University of Denmark, Department of Automation Aerospace Corp*., Los Angeles, CA, Tech. report no 98–H–874 (nfmod), 30 Oct 1998.

JAVIER, M. DE P. A. F. **Optimización Mediante Técnicas de Minería de Datos del Ciclo de recocido de una línea de galvanizado**. Tese de Doutorado, *Universidad de la Rioja*, 2003.

JOHN, Y.; REZA, L. **Fuzzy Logic intelligence control and information**. *New Jersey: Prentice Hall*, páginas 351–362, 1999.

JÚNIOR, N. L. C. **Clusterização baseada em algoritmos *fuzzy***. Dissertação de Mestrado – Universidade Federal de Pernambuco – Recife–Pe, 2006.

KAUFMANN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. *New York: Jonh Wiley*, página 342, 1990.

KHATTREE, R.; NAIK, D.N**. Multivariate data reduction and discrimination with SAS software**. *Cary, NC, USA: SAS Institute Inc*., página 558, 2000.

KOHONEN, T. **Self-organized formation of topologically correct feature maps**. Jornal Biological Cybernetics, Volume 43, Issue 1, páginas 59–69, 1982.

KOHONEN, T. **Self–Organization and Associative Memory**. 2° Edição, *Springer–Verlag, Berlin*, 1987.

KVANDE, H. **O Processo de Fundição do Alumínio**. Copyright Lippincott Williams & Wilkins. Texto traduzido, sob autorização, de artigo publicado no *Journal of Occupational and Environmental Medicine* (JOEM), Número 5S. Maio de 2014.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to *Data mining***. *John Wiley and Sons, Inc*., 2005.

LARRAÑAGA, P.; LOZANO, J.A. **A review on EDAs, in Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation**. *Norwell, MA: Kluwer*, capítulo 3, 2002.

LEE, J. (2017). **Aplicações e Tendências de mineração de dados**. Disponível em: <http:// www.w3ii.com/pt/data_mining/dm_applications_trends.html>. Último acesso Julho/2017.

LEE, S.; HAYES, M.H. **Properties of the singular value decomposition for efficient data clustering**. *IEEE Signal Processing Letters*, 11(11):862–866, Nov. 2004.

LI, J.; YANG, B.; SONG, W. **A New *Data mining* Process Model for Aluminum Electrolysis.** *Proceedings of the International Symposium on Intelligent Information Systems and Applications* (IISA'09). Qingdao, P. R. China, páginas 193–195, 28 a 30 de Outubro de 2009.

LINDEN, R. **Técnicas de Agrupamento**. Revista de Sistemas de Informação da FSMA, n. 4 , pp. 18-36. 2009.

MATOS, D. (2015). **Conceitos Fundamentais de Machine Learnig**. Disponível em: < http: //www.cienciaedados.com/conceitos-fundamentais-de-machine- learning/>. Último acesso Julho/2017.

MCFADDEN, F. S.; BEARNE, G. P.; AUSTIN, P. C.; WELCH, B.J. **Application of advanced Process Control to Aluminium Reductions Cells – A review**. *TMS Light Metals, Light Metals 2001 – Proceedings of the technical Sessions, 130rd Technical TMS Annual Meeting, February* 11 – 15, páginas 1233–1242, New Orleans, LA, USA, 2001.

MCKINSEY Global Institute. **Big Data: The next Frontier for Innovation, Competition, and Productivity**, 2011.

MERELO, J. J. (2004). **Mapa autoorganizativo de Kohonen, Tutorial**. Universidad de Granada. Disponível em: <http://geneura.ugr.es/~jmerelo/tutoriales/bioinfo/Kohonen. html>. Útimo acesso:Abril/2017.

NALDI, M. C. **Técnicas de combinação para agrupamento centralizado e distribuído de dados**. Tese de Doutorado, ICMC–USP. São Paulo, página 245, 2011.

NAMIKKA, E.; GIBBON, G. J. **Identification of data mining techniques for industrial process analysis and control**. *School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg*, 2002.

NATH, B. et. al. **Incremental association rule mining: a survey**. *Data mining and Knowlwdge Discovery*, v. 3, n. 3, página 157 – 169, 2013.

NEVES, L. T. **Análise de *Cluster* no R**. Disponível em <http://lincolntneves. weebly.com/blog/anlise–de–*cluster*–no–r>. Último acesso Julho/2016.

OVERFLOW, Stack (2012). **How to produce a pretty plot of the results of *K–Means* cluster analysis?** Disponível em <http://stats.stackexchange.com/questions/ 31083/how–to–produce–a–pretty–plot–of–the–results–of–*K*–*Means*–*cluster*–analysis>. Último acesso Janeiro/2017.

PASSOS, A. C. **Avaliação Multicritério de Material de Emprego Militar**. 79f. Dissertação (Mestrado em Administração de Empresas) – Faculdades Ibmec, Rio de Janeiro, 2002.

ROBERT, P.; CORONEL, C. **Sistemas de banco de dados – projeto, implementação e administração**. 8° Edição. São Paulo: *Cengage Learning*, 2011.

ROCHA, T.; PERES, S. M.; BÍSCARO, H. H.; MADEO, R. C. B.; BOSCARIOLI, C.; **Tutorial sobre *Fuzzy*–c–Means e *Fuzzy* Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamento e Classificação**. Volume 19. Número 1. 2012.

ROGERS, S.; GIROLAMI, M. **A first course in machine learning**. Boca Raton: CRC Press, 2011.

ROZENBERG, G.; BACK, T.; KOK, J.N. **Handbook of natural computing**. *Nova York: Springer*, 2012.

SEO, Y. K; JAE, W. L.; SUNG, B. J. **Effect of data normalization on fuzzy clustering of DNA microarray data**. BMC *informatics*. 7:134, 2006.

SILVA, A. J. **Modelagem Paramétrica de Fornos Eletrolíticas para Predição do Efeito Anódico**. Dissertação de Mestrado – Programa de Pós–graduação em Engenharia de Eletricidade – Universidade Federal do Maranhão, 2009.

SOARES, F. M. **Aplicação de Sensores Virtuais na Inferência da temperatura de banho no processo de fabricação de alumínio primário**. Dissertação de Mestrado – Universidade Federal do Pará – Belém–Pa, 2009.

SOUZA, F.; ALBUQUERQUE, C. **Linear wireless mesh network planning**. In 9th International Information and Telecommunication Technologies Symposium (I2TS 2010), 2010.

SOUZA, A. M. F. **Estimação da porcentagem de Flúor em alumina fluoretada proveniente de uma planta de tratamento de gases por meio de um sensor virtual neural**. Dissertação de Mestrado – Universidade Federal do Pará – Belém–Pa, 2011.

SOUZA, A. M. F; AFFONSO, C. M.; SOARES, F. M.; LIMÃO, R. C.; **Sensor Virtual Neural para Estimação de Flúor em Alumina no Processo de Fabricação de Alumínio Primário**. *Learning and Nonlinear Models* (L&NLM) – *Journal of the Brazilian Neural Network Society*, Vol. 9, Iss.3, páginas 157–167. Sociedade Brasileira de Redes Neurais (SBRN), 2011.

STROMMEN, S. O.; BJORNSTAD, E.; WEDDE, G. **$S0_2$ Emission Control in the Aluminium Industry**. *TMS Light Metals*, páginas 962–967, 2000.

SUDARSHAN, S.; SILBERCHATZ, A.; KORTH, H.F. Sistema de banco de dados. 6° Edição. Rio de Janeiro: Elsevier – Campus, 2013.

SUMATHI, S., SIVANANDAM, S. N. **Introduction to *Data mining* and its Applications**. *Springer*, 2006.

TAN, J. **Different types of association rules mining review**. *Applied Mechanics and materials*, páginas 1589 – 1592, 2012.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. *London : Elsevier*, 2009.

TRIOLA, M.F. **Introdução à estatística**. 11° Edição. São Paulo: LTC Livros Técnicos e Científicos Editora S.A., 2013.

TSAI, H. T.; LANE, J. W.; LIN, C.S. **Modern Control Techniques for The Processing Industries**, *Marcel Dekker Inc.*, 1986.

WEBB, A. **Statistical Pattern Recognition**. 2° Edição, *Weat Sussex: John Wiley & Sons, Inc.*, 2002.

WEBER, R. ***Data mining*** **en la Empresa Y en las Finanzas Utilizando Tecnologías Inteligentes**. Revista Ingeniería de Sistema, *Departamento de Ingeniería Industrial, Universidad de Chile*, vol 14, Núm. 1, 2000.

WEHRENS, R.; BUYDENS, L.M.C. **Self– and Super–organizing Maps in R: The Kohonen Package**. *Journal of Statistical Software*, V.21 (5), 2007.

WITTEN, I.H.; FRANK, E.; HALL, M.A. **Data Mining: practical machine learning tools and techniques**. 3° Edição. *Massachusetts: Morgan Kaufmann*, 2011.

XIANGTAO, C. et. al. Edited by Travis J. Galloway. **The development and application of data warehouse and data mining in aluminium electrolysis control systems**. *TMS The Minerals, Metals and Materials Society*, páginas 515–518, USA, 2006.

ZHA, W.; WEI–YIP, C. **Objective Speech Quality Measurement Using Statistical *Data mining*.** *EURASIP Journal on Applied Signal Processing*, páginas 1410–1424, 2005.

ZELNIK-MANOR, L.; PERONA, P. **Self-tuning spectral clustering**. In Advances in Neural Information Processing Systems 17, páginas 1601–1608. MIT Press, 2004.

Zhuo, C. et. al. **A New Model for the Industrial Process Control based on Data Mining**. *IEEE Advancing Technology for Humanity*. *Chinese Control and Decision Conference* (CCDC 2008), páginas 1368 – 1370, 2008.