Abordagens Inteligentes para a Identificação e Tratamento de Outliers em Dados de Demanda de Energia Elétrica

Tadeu de Carvalho Machado
Bacharelado em Ciência da Computação
Universidade Federal de Alfenas – UNIFAL-MG
Alfenas – MG, Brasil
a08035@bcc.unifal-mg.edu.br

Ricardo Menezes Salgado
Docente do Bacharelado em Ciência da Computação
Universidade Federal de Alfenas – UNIFAL-MG
Alfenas – MG, Brasil
ricardo@ bcc.unifal-mg.edu.br

***Resumo* — O setor elétrico brasileiro é predominantemente composto de hidrelétricas. Estas dependem de uma determinada taxa de demanda mínima de carga elétrica para que a produção possa fornecer eficientemente o seu serviço. Para tanto, sistemas de previsão de carga são cada vez mais utilizados, e neste contexto, tais sistemas dependem da qualidade dos dados manipulados para que se tenham resultados satisfatórios. Sendo assim, a análise desses dados quanto à existência de dados corrompidos torna-se indispensável. No entanto, ao utilizar apenas um determinado modelo de filtragem de elementos, corre-se o risco de se obter resultados mascarados, com baixo grau de detecção, alto grau de erros, ou apresentar bons resultados apenas para determinados tipos de séries. Para solucionar este problema, neste trabalho foi desenvolvido um modelo de combinação dos resultados obtidos por diversos modelos de detecção deste tipo de dado, o *outlier*.**

***Abstract* — The brazilian electric sector is predominantly made up of hydropowers. These depend on a certain rate of minimum demand charge for electricity production efficiently to provide this service. For this, systems of load forecasting are increasingly used, and in this context, such systems depend on the quality of the data manipulated to get satisfactory results. So the analysis of these data as to the existence of outliers becomes indispensable. However, when using only one particular model of filtering of elements, take the risks of getting masked results, low degree of detection, high degree of errors, or get good results only for certain types of series. To solve this problem, in this paper was developed a model combining the results from different models of detection of this type of data, the *outlier*.**

Palavras-chave* — *Detecção de Outliers, Ensemble, Chave Inglesa, Conjunto de Dados de Energia Elétrica.

I. INTRODUÇÃO

A qualidade dos dados no setor elétrico é de extrema importância, uma vez que estes dados contêm informações que refletem o estado operacional do sistema elétrico em questão. Contudo, dentre os elementos que compõem este conjunto de dados, há um tipo especial de elemento que se deve ter a preocupação de se avaliar, o *outlier*.

O *outlier* é um elemento de uma série de dados que, por algum motivo, se distancia de forma característica à normalidade da amostra em questão, parecendo assim ser inconsistentes de acordo com o restante desse conjunto de dados [1]. Anscombe [2] define um *outlier* como uma observação com grandes anomalias residuais. Outros termos para *outlier* seriam "dissonante", "anormal" ou "aberrante".

Outliers são geralmente comuns, no entanto, sua verdadeira causa de ocorrência é geralmente desconhecida para os usuários de dados e/ou analistas [3]. Algumas das causas que podem ser atribuídas aos seus aparecimentos numa amostra são principalmente, anomalias de medição inerentes a falhas nos sensores, erros inerentes à execução, ou ainda a fatores ligados a própria variabilidade de seus elementos.

Em séries temporais, esta que designa um conjunto de dados que corresponde a um período delimitado de tempo, associando cada elemento a uma métrica unitária de tempo, como horas ou minutos, tem-se uma preocupação ascendente com a consistência dos dados.

Mesmo com ferramentas gráficas altamente sofisticadas, muitas vezes se torna impossível analisar estes dados utilizando processos de visualização e identificação manual [4]. Fato que se atribui a grande quantidade de informações contidas neste espaço amostral, o qual representa uma série temporal.

Tratando-se de séries de carga elétrica, os níveis de precisão de uma análise detalhada em uma amostra de dados estão diretamente ligados à concepção idônea dos mesmos. Muitas vezes *outliers* contêm informações úteis sobre o comportamento anormal do processo, sendo assim, é exigida uma investigação mais aprofundada, analisando cada tipo de padrão de ocorrência [5]. Logo, se a amostra não for tratada, detectando e normalizando ou mesmo descartando os dados corrompidos, qualquer afirmação sobre a mesma não estará corretamente embasada e, portanto não traduzirá a realidade. Com isso, a previsão de carga alcança erros que podem comprometer a eficiência do sistema como um todo.

Contudo, deve-se ficar atento ao que é detectado como um *outlier*, pois é possível que ele não pertença de fato a essa classe, dados que são denominados falso-positivos. Caso tais elementos sejam normalizados, o conjunto de dados perde as suas características originais, o que caracteriza o mascaramento da série temporal.

Deve-se ainda ressaltar se ele será tratado ou descartado. Tais decisões devem ser determinadas analisando cada caso individualmente, de acordo com a natureza da amostra.

II. Revisão de Literatura

Os principais métodos de detecção de *outliers* são baseados em densidade, e a detecção dos mesmos depende da relação dos *outliers* para o restante do conjunto de dados. No entanto, em um amplo espaço dimensional, como numa série temporal, os dados se tornam cada vez mais independentes e a noção de proximidade e densidade perde o seu significado, perdendo seu poder de separação é minimizado o bastante para se tornar ineficaz [5]. Desta forma, os cientistas passaram a concentrar seus esforços na concepção de métodos baseados em modelos com regressões robustas, devido à presença de *outliers* de origem não sistêmica desconhecidos, inerentes a natureza da amostra, tem-se conseguido um avanço mais significativo.

Laurikkala, Juhola e Kentala [6] fizeram identificação de *outliers* em dados reais médicos. Nos testes, foram utilizados dados de incontinência urinária. Os resultados experimentais sugerem que a detecção dos dados corrompidos pode ser utilizada para a redução de dados. A vantagem obtida após a exclusão de *outliers* é dependente do conjunto de dados, pois a remoção de um *outlier* ajuda a análise descritiva, mas a análise preditiva, ou seja, a classificação dos casos invisíveis pode piorar. Por isso, a utilidade do método depende também do objetivo final da análise. Nos estudos foi constatado que o comportamento de análise preditiva precisa ser focado ainda mais, pois o aplicativo de aprendizagem muitas vezes é usado para classificar os dados novos. A principal limitação deste trabalho é a utilização informal do método *BoxPlot*. Infelizmente, o método faz muitas suposições que devem ser observadas, estas condenam os testes antes de ser aplicável. Há também duas limitações multivariadas na identificação de *outlier* com a distância de Mahalanobis, pois esta funciona melhor com métodos quantitativos normais do que com dados distribuídos, e ainda vale ressaltar que devem ser tratados os valores ausentes antes do cálculo de distância. Finalizando o estudo, é observado que algum tipo de função heterogênea de distância poderia resolver esses problemas, mas infelizmente torna o método menos eficiente.

Segundo Last & Kandel [3], é possível a concepção de bons resultados através de modelos computacionais com percepção humana. Basicamente trata-se de uma ferramenta gráfica que se encarrega de apresentar o conjunto de dados, encarregando a responsabilidade pela detecção dos valores excepcionais a uma pessoa especialista no assunto, que os verifica de maneira manual. No entanto, para um grande conjunto de dados este método torna-se pouco eficiente e muito demorado, uma vez que seja possível propor um algoritmo eficiente no campo.

Nos estudos de Elsa M. Jordaan, Dow Benelux BV, Guido e Smits e Dow Benelux BV [5], foi utilizada uma abordagem de detecção de *outlier* baseada em modelos Support Vector Machine (SVM) sem ambiguidade, que usa vários modelos de complexidade variável para detectar, com base nas características dos vetores de suporte, os *outliers* obtidos nos modelos SVM. Foi verificado que, a decisão não depende da qualidade de um modelo único, mas da robustez da abordagem como um todo. Além disso, sendo uma abordagem iterativa, os *outliers* mais graves são removidos primeiro, permitindo que os modelos na próxima iteração possam aprender com os dados "mais limpos" e, portanto, revelar *outliers* que foram "mascarados" no modelo inicial.

A abordagem usada por Zhu Cui, KitagawaHiroyuki, Papadimitriou Spiros e Faloutsos Christos, [7] consiste em um novo e refinado método de detecção de *outliers*, interativo e adaptável às intenções do usuário. O fato de que a avaliação de um dado normal muitas vezes depende do usuário e ou do conjunto de dados faz com que o problema de detecção de *outlier* ser difícil de resolver. Assim, esta metodologia permite que o usuário dê alguns exemplos de *outliers*, agindo de forma interativa com o sistema. Experimentos com dados reais e sintéticos demonstram que o método interativo pode ter sucesso ao incorporar esse método, incluindo resultados positivos no feedback e na detecção de *outliers* falsos, como determinado em seus estudos.

Filzmoser [8] mostra um método para a detecção de *outliers* multivariados que propõe medidas para o tamanho de estrutura e amostra de dados. O método identifica *outliers* no espaço multivariado. Nos estudos ainda foi ressaltado sua simplicidade de implementação e sua facilidade para calcular. Este compara a diferença entre a robusta distribuição empírica do quadrado das distâncias e a função de distribuição da distribuição *Chi-quadrado*. O método conta não só para uma dimensão diferente dos dados, mas também para tamanhos de diferentes amostras.

Filzmoser [4] comparou o desempenho de três métodos para a identificação de *outliers* multivariados, Rousseeuw, Becker e Filzmoser, que se baseiam na distância de Mahalanobis robusta, que contam com uma estimativa da localização e covariância. Nas simulações descritas, foi observado que o desempenho dos métodos de Filzmoser e de Rousseeuw são comparáveis, apresentando aproximadamente as mesmas porcentagens de *outliers* simulados (artificiais), mas os não-*outliers* (falsos-positivos) também foram detectados em ambos os métodos. Já o método Becker é preferível pela sua baixa taxa de classificação de falsos *outliers*. No entanto, o seu comportamento como um identificador de *outlier* foi bastante pobre para dados de baixa dimensão e muito melhor para dados de dimensão superior.

Baragona, Calzini e Battaglia [9] propuseram um algoritmo genético para a identificação de *outliers*, em uma determinada série temporal. Tal método mostrou-se bastante eficaz, baseando-se em um grande conjunto de dados na procura dos candidatos a *outliers*. O método utiliza uma função objetiva com um conjunto distinto de valores para o calculo de legitimidade dos elementos do conjunto. Neste processo iterativo de algoritmo genético, ao contrário de outros métodos iterativos, os *outliers* não são identificados e removidos um de cada vez, mas sim analisando o conjunto de dados como um todo. Sendo assim, o valor de cada dado é calculado sobre o padrão dos *outliers* detectados. Esta característica parece ser

capaz de lidar eficazmente com o efeito *swamping*, que surge quando as observações que são compatíveis com a maioria dos dados ainda são detectadas incorretamente como *outliers*, os falsos-positivos, estes possivelmente decorrentes de algum tipo de diversidade acidental, ou mesmo do problema de mascaramento, o qual é peculiar neste contexto, em que os *outliers* consecutivos realmente têm grande probabilidade de ocorrer.

Chiang [10] se baseou no método Gentleman and Wilk's para detecção de *outliers*, que consiste em encontrar um subgrupo de dados que tem a soma mínima de resíduos ao quadrado. O método proposto modifica cada subgrupo analisado para o que tem a soma mínima de erro de previsão ao quadrado. Em seguida, o algoritmo encontra os melhores dados de construção para os parâmetros definidos, baseando-se nos *resíduos de Jackknife* absolutos. Todo um conjunto de dados é dividido em dois grupos. O primeiro deles é um grupo de vasculhamento com o objetivo de calcular a função prevista, e o outro, contem os valores corrompidos, o qual é examinado. O algoritmo verifica-se útil e rápido para encontrar vários *outliers*, baseando-se apenas na divisão de dados e resíduos brancos, sendo muito mais simples do que modelos não lineares. Neste sentido, o diagnóstico de um único *outlier* em modelos lineares pode ser estendido para a detecção de *outliers* múltiplos. Logo, os efeitos de mascaramento e proliferação no método não são um problema.

A abordagem de Lukashevich H., Nowak S., Dunker P. [11] consiste na detecção automática de *outliers* através da formação de conjuntos de imagens, utilizando o apoio da ferramenta SVM. Nos experimentos foram usados quatro tipos de imagens: *Snow & Skiing, Family & Friends, Architecture & Buildings e Beach*. A ferramenta SVM oferece este método de detecção de *outliers* que conta com um conjunto de abordagens que podem lidar com uma pequena quantidade de incertezas, o que é aceitável. Em suma, o experimento demonstra uma prova por conceito para rejeitar automaticamente os *outliers* a partir dos dados de treinamento, utilizando este método de detecção oferecida pela ferramenta em questão.

No trabalho de Prabhjot Kaur, Anjana Gosain [12], são utilizados algoritmos de agrupamento, chamados clusters. Os algoritmos consideram todos os dados como normais, ao mesmo tempo dividindo-os em clusters. No entanto, os resultados dos mesmos não são capazes de produzir clusters eficientes. Sendo assim, para a identificação de *outliers* é necessário dividir o conjunto de dados em *clusters fuzzy*, que consistem em um grupo de dados similar. Sendo assim, os *outliers* não pertencem a qualquer outro grupo semelhante, permanecendo isolados em vários clusters. Os resultados mostram claramente que identificar *outliers*, antes de aplicar qualquer um dos algoritmos de clusters, pode melhorar o desempenho drasticamente. A técnica proposta de identificação *outliers* se revelou ótima, mostrando ser mais eficiente quando se agrupam os dados.

Onoghojobi [1] tentou superar as dificuldades associadas à captura de *outliers* utilizando um modelo linear dinâmico, reformulando os critérios de desempenho para o processo de identificação de *outliers* multivariados, baseando-se em um método de detecção de *outlier* chamado *Becker*. Em seus estudos, foi observado que a técnica de detecção de *outliers* de *Becker* fornece melhor eficiência para identificadores de anormalidades se comparada aos critérios de desempenho convencionais para procedimentos multivariados de identificação de *outliers.*

Este trabalho propõe um método para detecção de outliers, chamado *Chave Inglesa*, que é baseado em diferenças absolutas de dados e um combinador *Ensemble* [13] destes modelos com as seguintes técnicas: *BoxPlot, Teste de Chauvenet, Teste ZScore, Delete Outlier e Teste de Hampel*. Para testar os modelos propostos foram utilizados dados reais de empresas do setor elétrico.

O conteúdo deste trabalho encontra-se organizado da seguinte forma, na Seção III são apresentados os métodos de identificação de *outliers* utilizados, na Seção IV é abordada a metodologia utilizada para realização dos experimentos, já na Seção V é feita uma discussão sobre os dados utilizados para o tratamento das cargas e na Seção VI são feitas as considerações finais do trabalho.

III. FUNDAMENTAÇÃO TEÓRICA

É característico de séries temporais de cargas elétricas a sua alta variabilidade, assim como a sua predisposição a algumas características que somente podem ser observadas como um todo. Sendo assim, apenas um modelo de detecção de dados corrompidos torna-se pouco eficiente quando abrange toda a dimensão deste problema, pois cada tipo de método de detecção tem bons resultados para alguns tipos de curvas características e resultados menos satisfatórios para outras.

Nesta seção, serão apresentados os métodos utilizados para identificação de *outliers.*

A. BoxPlot

O Boxplot, ou o método também conhecido como diagrama em caixa, é um gráfico proposto por Tukey [14], sendo utilizado para revelar o centro, a dispersão e a distribuição dos dados, além da presença de *outliers*.

Basicamente este método estatístico é construído com base na mediana, no quartil inferior (Q1), no quartil superior (Q3) e no intervalo interquartil (IQR), que é calculado pela subtração entre Q3 e Q1. Os limites dessa caixa são delimitados por duas linhas que são calculadas de acordo com 150% o valor do IQR. Este valor é somado com Q3 para obter-se o valor do limite superior e subtraído de Q1 para resultar no limite inferior, formando-se assim uma distância segura à normalidade, que é utilizado para isolar os dados aberrantes. Assim, os valores inferiores ao limite inferior e superiores ao limite superior são caracterizados como *outliers*.

O método é simples de ser aplicar, além de revelar outras medidas importantes, como a mediana, a dispersão e a assimetria dos dados.

B. Teste de Chauvenet

O teste foi proposto por Chauvenet em 1960, e especifica a eliminação de um único valor duvidoso, caso seja necessário. Para eliminar um segundo valor seria necessário recalcular a média e o desvio padrão para o novo conjunto de dados e só então aplicar novamente o critério. Porém, o método não especifica nenhum limite para a aplicação do método. Entretanto, como a cada novo cálculo o desvio padrão

diminui, é muito provável que essa aplicação sucessiva resulte na eliminação de um grande número de dados. Sendo assim, é preferível aplicar o critério uma única vez para cada conjunto de dados, eliminando todos os valores que se encontram fora do intervalo estabelecido, tal como é usado neste trabalho.

C. Teste Z-Score

O teste Z-score é uma medida estatística de relação, em termos de desvios padrões e em relação à média. Assim, um valor z-score calculado, definido neste método, determina o número de vezes, de acordo com o desvio padrão, que cada valor está acima ou abaixo da média. Com este cálculo pode-se utilizar destes valores para a identificação de *outliers*, pois um valor z-score muito alto ou muito baixo indica que determinado valor está fora do padrão de comportamento do restante do conjunto de dados.

O valor z-score é calculado pela subtração entre cada dado do conjunto e a média amostral, e posteriormente faz-se a divisão deste valor pelo desvio padrão amostral.

Em seguida, é realizada uma comparação do valor z-score calculado com um valor padrão fixado, de acordo com o tamanho do conjunto de dados igual a *n*. Conforme o resultado dessa comparação, o valor é classificado como um *outlier* ou não da seguinte forma: Se $n \leq 50$ *e z-score* $\leq -2{,}5$ *ou z-score* $\geq 2{,}5$*; ou se* $50 < n < 1000$ *e z-score* $< -3{,}3$ *ou z-score* > 3*; ou ainda, se* $n \geq 1000$ *e z-score* $\leq -3{,}3$ *ou z-score* $\geq 3{,}3$, este dado é tido como um *outlier*, caso contrário o mesmo é classificado como um dado normal.

D. Delete Outlier

O método Delete Outliers é baseado no Teste de Grubbs[15]. Este método de detecção de *outliers* consiste em identifica-los nos extremos de um conjunto de dados, ou seja, para verificar se o menor e o maior valor do conjunto são *outliers*, comparando o valor suspeito com os demais valores do conjunto de dados. O método calcula um valor para o menor e maior valor da amostra utilizando o desvio padrão como denominador e comparando posteriormente com o valor crítico tabelado para o nível de significância desejado, que neste trabalho é escolhido como 0,05. Valor que pode ser alterado caso necessário. Sendo assim, caso os valores calculados excedam os limites inferior ou superior calculados de acordo com o valor crítico tabelado, então o dado em questão é considerado um *outlier*.

E. Teste de Hampel

Hampel [16] introduziu o conceito do ponto de ruptura, que é descrito como uma medida para estimar a existência dos *outliers*. O algoritmo define o ponto de ruptura como a menor porcentagem de dados que pode estimar a tomada de grandes valores arbitrários, os *outliers*. Assim, o maior ponto de ruptura que é estimado tem a maior robustez do conjunto. Por exemplo, mesmo uma única observação grande pode fazer a média da amostra e a variância cruzar qualquer limite, o que não seria aceitável. Assim, o autor sugeriu a mediana e o desvio absoluto médio (MAD) como estimativas para estabelecer este limite. O método de Hampel demonstra-se muito eficaz na detecção de *outliers*.

III. Metodologia Proposta

A metodologia escolhida para este trabalho se baseia no uso de um método de detecção chamado *Chave Inglesa*, juntamente com os outros cinco métodos já citados. Com o objetivo de fazer a junção dos resultados e obtenção de um modelo de identificação com alta capacidade de generalização, é proposto também o método *Ensemble* [13], que combina os resultados dos seis métodos utilizados.

Posteriormente, é utilizado o tratamento de dados, para que os dados corrompidos sejam normalizados de maneira a obter o máximo de proximidade aos dados normais.

A. Teste Chave Inglesa

O teste *Chave Inglesa* foi desenvolvido e proposto neste trabalho. Esta técnica consiste em calcular a legitimidade de cada unidade da série de dados através de comparação entre a diferença absoluta de cada valor ao seu vizinho de acordo com uma tolerância definida para cada série de dados. O Quadro 1 apresenta o pseudocódigo do método aqui descrito.

```
Dado o conjunto de dados D com dimensão n, e os
parâmetros c0 > 0, onde c0 é a constante de ajuste de
acordo com a variabilidade da série, e c1 > 0 onde c1 é
uma constante suficientemente grande com o intuito de
verificar a existência de quedas de demanda, e a variável
I que determina o intervalo de aceitação.

1. Iniciar I com o valor da variância; o número de
aceitos, aceitos = 1; a média de intervalos aceitos,
media = 0; o número de outliers detectados em sequência,
sequencia = 0; o último elemento normal, ultimo = D[1] e
o contador k = 2.

2. Verificar a aceitação inicial para cada elemento D[k]
onde k <= n, com o seguinte procedimento:

    se (|D[k] - ultimo| > (variancia + I) × c0 ÷ 2)
        D[k].estado <- corrompido;
        sequencia <- sequencia + 1;
    senão
        sequencia <- 0;
        media <- media + |D[k] - ultimo|;
        aceitos <- aceitos + 1;
        I <- (I + (media ÷ aceitos)) ÷ 2;
        ultimo <- D[k];
        D[k].estado <- normal;

    queda[k] <- sequencia;
    k <- k + 1;

3. Reiniciar o intervalo com o valor da variância, o
número de aceitos = 1, a média = 0 e o contador k = n-1.

4. Verificar a aceitação final para cada elemento Dk onde
k >= 1, com o seguinte procedimento:

    se (|D[k] - ultimo| < (variancia + I) × c0 ÷ 2)
         se (D[k].estado == corrompido)
                 D[k.estado <- normal;

        media <- media + |D[k] - ultimo|;
        aceitos <- aceitos + 1;
        I <- (I + (media ÷ aceitos)) ÷ 2;
        ultimo <- D[k];
    senão
        se (queda[k] > c1 e |D[k] - D[k + 1]| <
(variancia + I) × c0 ÷ 2))
                 D[k].estado <- normal;
                 ultimo <- D[k];

    k <- k - 1;
```

Quadro 1: Pseudo-código do algoritmo Chave Inglesa.

O método inicialmente define o conjunto de dados *D*, com dimensão *n*. Os parâmetros *c0*, *c1* > *0*, onde *c0* é uma constante de ajuste que altera a tolerância do valor da subtração absoluta entre os dados vizinhos para a aceitação, e *c1* é uma constante suficientemente grande com o intuito de verificar uma sequência de *outliers* detectados em sequência, e posteriormente analisar a existência de quedas de demanda.

Posteriormente o método define também um valor variável a cada iteração chamada I, também chamado de intervalo, que consiste numa média ponderada de todas as subtrações absolutas de dados vizinhos aceitos, esta última é previamente instanciada como a própria variância do conjunto de dados amostral.

Assim, para cada dado do conjunto, é comparado o valor absoluto da subtração entre mesmo e o seu vizinho mais próximo aceito como um dado normal e o valor obtido pela média aritmética entre a variância e o valor atual de I, multiplicado pela constante *c0* dividido por dois. Caso o primeiro valor seja menor em relação ao segundo assume-se que tal valor seja um dado normal inicialmente, e assim o valor de I assume o valor da média aritmética entre o próprio valor I e a média aritmética de todas as diferenças absolutas de dados vizinhos que foram aceitas até a iteração em questão, reinstanciando também as variáveis *sequencia média*, *aceitos* e *ultimo*. Caso contrário o dado é instanciado inicialmente como um *outlier*. Este procedimento é feito para todos os dados da série, a começar pelo segundo, até o ultimo elemento da série.

O mesmo método descrito acima é realizado novamente, mas nesta etapa o sentido de varredura é inverso, ou seja, começa-se do penúltimo terminando no primeiro elemento da série. Sendo assim, caso o dado em questão seja avaliado novamente como um *outlier*, ele é definido definitivamente como tal. Caso o elemento inicial seja verificado como um outlier, todo o procedimento tem que ser refeito, pois a toda as série de dados é validada de acordo com o elemento inicial da mesma, e neste caso, a varável *ultimo* é iniciada com o primeiro elemento verificado normal desta série de dados.

O método descrito anteriormente tem excelentes resultados se aplicados a séries bem comportadas, entretanto é característico de séries temporais de cargas elétricas a ocorrências de quedas de demanda, que são associadas a períodos de tempos específicos que podem ou não ser considerados *outliers*. Na Figura 1, no intervalo ente 300 e 400 no eixo horizontal do gráfico, podemos observar um exemplo de queda de demanda.

Para tanto é proposta uma nova regra quando há necessidade em classificar esses tipos de dados como dados normais. Basta adicionar um contador na primeira etapa do processo, a variável *sequencia*, que será incrementado a cada *outlier* detectado em sequência, caso haja um dado normal entre os mesmos, o contador voltará a ser instanciado com zero. Então, caso a variável *sequencia* seja maior que a constante *c1*, então é possível que sejam detectadas partes independentes da série, identificadas como quedas de demanda. Da mesma maneira, utiliza-se o mesmo método descrito anteriormente para validar cada um de seus elementos.

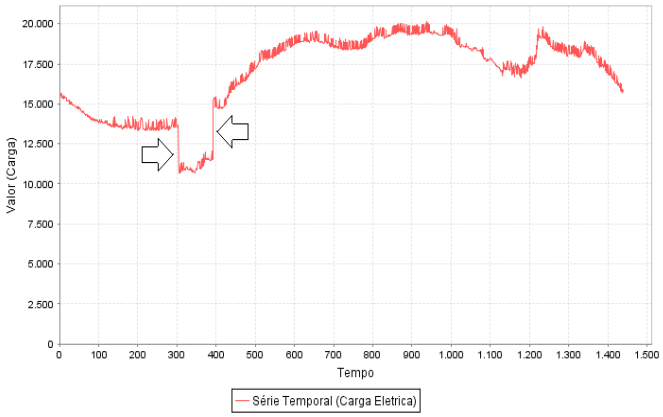

Figura 1: Queda de demanda.

B. *Ensemble*

O *Ensemble* consiste basicamente em um sistema de combinação de resultados dos métodos que o compõem, ou seja, o *ensemble* aplica o conhecimento gerado em cada um dos métodos, levando em conta pontos comuns verificados por todos ou grande parte destes.

Neste método, os resultados de diferentes modelos de detecção de *outliers* são combinados e reanalisados pelo *ensemble* dado a dado. De acordo com um número mínimo de métodos com validações positivas quanto à constatação deste dado com um *outlier*, o mesmo é verificado como um *outlier* ou não (Quadro 2).

Uma das principais vantagens deste modelo é a sua grande capacidade de manter-se imparcial quanto aos resultados isolados de cada método, considerando assim a veracidade de cada dado analisando de maneira genérica. Como consequência dessa abordagem de análise do conjunto de dados, os resultados tornam-se mais estáveis, gerando assim resultado mais confiáveis.

```
Dado o conjunto de dados D com dimensão n, para cada um
dos elementos D[i], onde i <= n:
Se o número de métodos de detecção que avaliaram o
elemento em questão como outlier for maior ou igual à
taxa de corte definida pelo Ensemble, então o dado é
classificado como um outlier.
```

Quadro 2: Pseudo-código do *Ensemble*.

C. *Tratamento de dados*

O modelo de tratamento dos dados proposto é denominado Tratamento por Média.

O Tratamento por Média consiste em uma média aritmética simples entre o último dado e o dado posterior ao em questão detectados como normais, ou seja, para cada dado corrompido detectado, são escolhidos dois dados normais, o mais próximos possível do dado em questão, e em seguida, é feito um cálculo, através de média aritmética entre eles, obtendo assim, o seu valor tratado (Equação 1).

$$D_k = \left(D_{i-1} + D_{j+1}\right) \div 2 \qquad (1)$$

Onde D_k é o elemento em questão, e D_{i+1} e D_{j+1}, são o último e o primeiro elementos da série detectados como normais, respectivamente, *sendo i, j, k* $\geq$*;* $i \leq k$*;* $j \geq k$ e *i, j, k* $\leq n$.

A Figura 2 mostra o exemplo de um dado *D[k]* corrompido, onde *D[k-1]* e *D[k+1]* são dados normais.

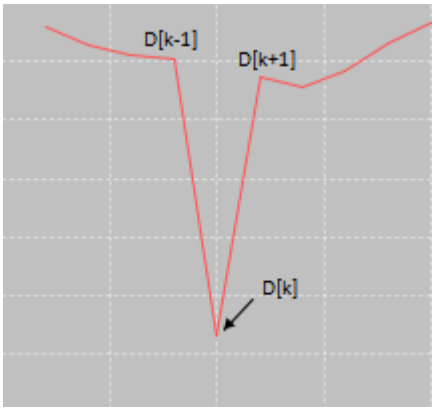

Figura 2: Dado corrompido com vizinhos D[k-1] e D[k+1] normais.

A Figura 3 mostra o exemplo do dado *D[k]* corrompido que foi tratado pelo método de Tratamento por Média, onde *D[k-1]* e *D[k+1]* são dados normais.

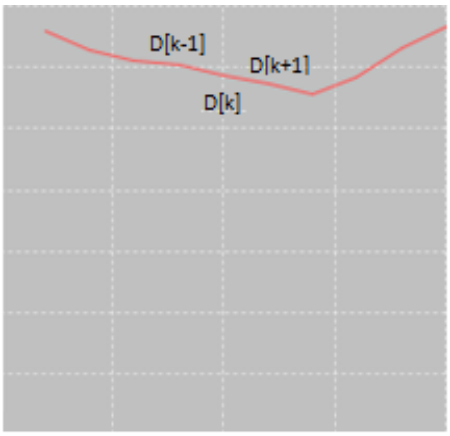

Figura 3: Dado tratado com vizinhos D[k-1] e D[k+1] normais.

Caso não exista um dado normal anterior ou posterior ao dado em questão, então se utiliza dois dados normais os mais próximos possíveis que possam fornecer um intervalo de confiança aceitável para a realização dos cálculos.

Para um dado inicial corrompido, se usa a seguinte Equação 2:

$$D_1 = D_{i+1} - \left(D_{i+1} - D_{j+1}\right) \qquad (2)$$

Onde D_1 é o elemento inicial da série, e D_{i+1} e D_{j+1}, são o primeiro e o segundo elementos da série detectados como normais, respectivamente, sendo $i \geq 1$ e $j \geq i$.

Para um dado final da série que está corrompido, se usa a seguinte Equação 3:

$$D_n = D_{n-i+1} - \left(D_{n-i+1} - D_{n-j-1}\right) \qquad (3)$$

Onde D_n é o último elemento da série, e D_{n-i+1} e D_{n-j-1}, são o penúltimo e o antepenúltimo elementos da série respectivamente, detectados como normais.

IV. Estudo de Casos

Os estudos de caso de análise e tratamento de dados neste artigo foram feitos com cargas elétricas reais, com um histórico no período de um ano (2006), com métricas unitárias de hora em hora para cada dia do ano. Com a definição dos dois tipos de métricas, por cada hora ou total, é possível detectar tanto outliers numa série definida pela hora quanto pela carga de todas as horas em todos os dias do ano em questão, assim como o ano todo em uma só análise.

A Figura 4 mostra a série de dados de carga horária, com 24 horas durante os 365 dias do ano de 2006.

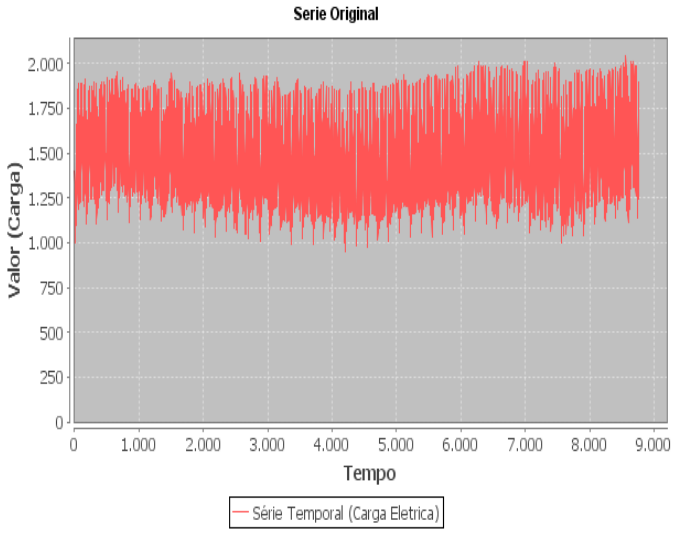

Figura 4: Série de dados de carga anual normal do experimento.

A Figura 5 mostra a série de dados de carga horária, as 06:00 durante os 365 dias do ano de 2006.

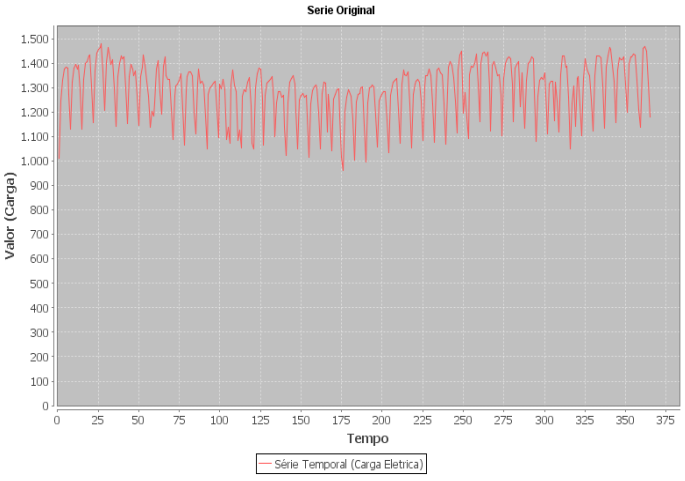

Figura 5: Série de dados horária de carga normal do experimento.

A. Geração artificial de Outliers

Para testar os modelos propostos neste trabalho foi utilizado um método de geração artificial de dados corrompidos. Tal método se mostra necessário, pois em dados reais não se pode afirmar com certeza que cada um dos mesmos é um outlier ou não. Deste modo, foi escolhido um método que escolhe aleatoriamente de 4% a 8% de elementos no conjunto, corrompendo-os da seguinte forma:

- Aleatoriamente é escolhido se o dado será corrompido com valor superior ou inferior ao dado original;
- Caso seja escolhido corromper com valor superior ao dado original, então o seu novo valor será aleatoriamente escolhido entre 150% a 250% de seu valor;
- Caso contrário, então o seu novo valor será aleatoriamente escolhido entre 0% a 50% de seu valor original.

Como podem ser percebidos na Figura 6 e na Tabela 1, os dados corrompidos diferem bastante da série original.

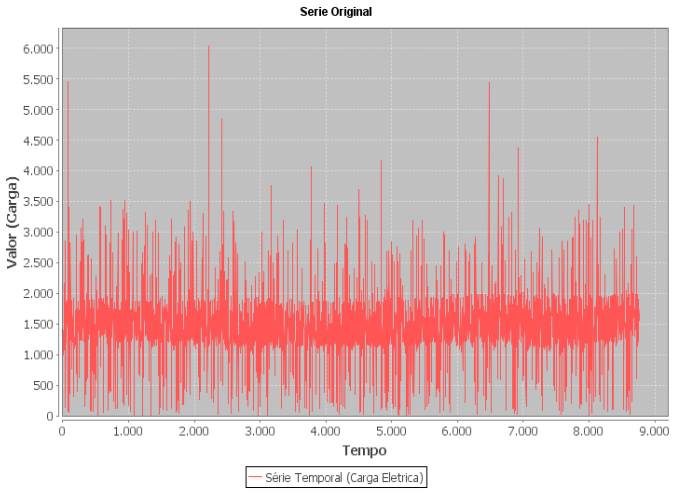

Figura 6: Série de dados corrompida.

Tabela 1 – Estatísticas de comparação entre a série normal e a corrompida.

Valores Estatísticos	***Série Original***	***Série Corrompida***
Variância	227,7788	402,6523
Desvio Padrão	15,0923	20,0661
Média	1465,0067	1462,6793
Máximo	2.042,10	6035,20
Mínimo	951,00	2,90

B. Análise dos resultados do experimento

Com o objetivo de avaliar os resultados comparando os dados normais aos dados tratados, fazendo a sua validação, foi utilizado o erro relativo médio (ERM), apresentado na Equação 4:

$$\mathrm{ERM} = 100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \qquad (4)$$

Onde x_i é o valor normal, $\hat{x}_i$ é o valor da série normalizada e n é a dimensão da série. Nesse mesmo sentido, como nos testes há possibilidade de fazer a comparação com os dados reais, pode ser determinada também a taxa de acerto dos métodos de detecção assim como o *Ensemble*, assim como a taxa de identificação de falsos positivos.

A partir dos testes realizados, foi concluído que os parâmetros para que o método *Chave Inglesa* proposto tenha um melhor rendimento seja de dois para a constante *c0* e de 5% da dimensão do conjunto de dados total para a constante *c1*. Assim, o intervalo de aceitação do método tem variação máxima definida pela própria variância, e a métrica de verificação para transferências de carga é suficiente tendo em conta o conjunto de dados utilizado.

A taxa de corte utilizada no *Ensemble*, que é o número mínimo de métodos detectores que verificam um dado como *outlier* para ser verificado como tal, verificou-se ser mais eficiente quando maior ou igual a dois. Ou seja, caso um determinado elemento da série seja verificado como corrompido por dois ou mais métodos de detecção ele é tido como um *outlier*.

A Tabela 2, a Tabela 3 e a Figura 7 avaliam o aproveitamento dos métodos descritos e do *Ensemble* proposto.

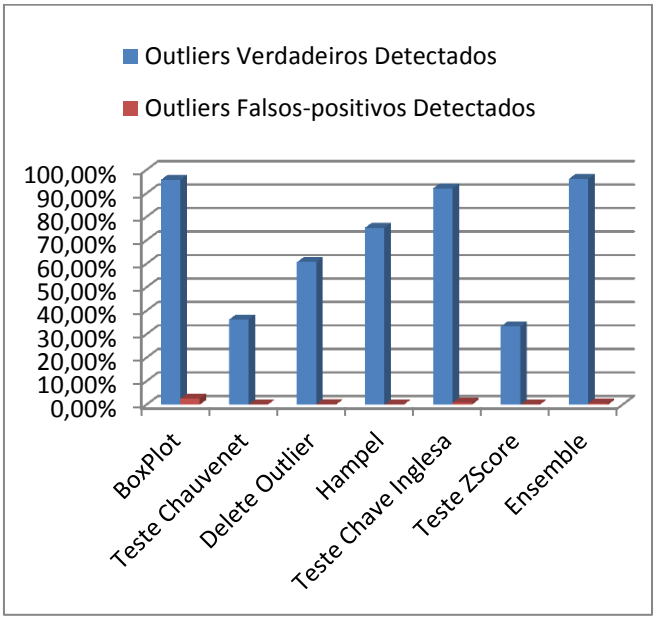

Figura 7: Gráfico do aproveitamento dos métodos de detecção.

Tabela 2 – Aproveitamento dos métodos de detecção.

Métodos de Detecção	***Outliers Verdadeiros Detectados***	***Outliers Falsos-positivos Detectados***
BoxPlot	96,803%	2,603%
Teste Chauvenet	36,225%	0,00%
Delete Outlier	60,73%	0,25%
Teste de Hampel	75,342%	0,00%
Teste Chave Inglesa	92,085%	0,819%
Teste ZScore	33,333%	0,00%
Ensemble	92,694%	0,16393%

Tabela 3 – Estatísticas de comparação entre a série normal e a corrompida.

Métodos de Detecção	***Outliers Verdadeiros Detectados (Total de 657)***	***Outliers Falsos-positivos Detectados***
BoxPlot	636	17
Teste Chauvenet	238	0
Delete Outlier	399	1
Teste de Hampel	495	0
Teste Chave Inglesa	605	5
Teste ZScore	219	0
Ensemble	609	1

Analisando os resultados podemos observar que o método *BoxPlot* conta com uma maior taxa de detecção de outliers verdadeiros comparando-se com os outros métodos utilizados, inclusive o *Ensemble*. Contudo, pode ser observado também que a taxa de detecção de falsos-positivos foi a mais alta do comparativo, por este motivo, o método *Ensemble* mostra-se mais interessante, pois a normalização de *outliers* falsos-positivos é mais dispendiosa mesmo se obtendo um ganho maior de *outliers* verdadeiros detectados.

Os testes foram realizados analisando cada hora para todos os dias do ano de 2006 analisando a série horária. A Figura 8 ilustra a detecção do *Ensemble* proposto.

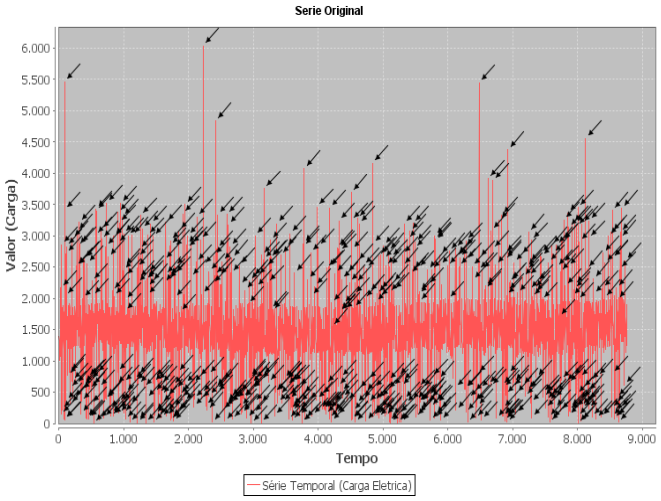

Figura 8 – Série de dados com os outliers detectados.

Como pode ser percebido, o método descrito reconhece grande parte dos dados corrompidos, mesmo com alto grau de perturbação da série.

No experimento descrito, foi aplicado o Tratamento de Dados por Média. Assim, obtemos o gráfico tratado a partir da série de dados que foi detectada como pode ser visualizado na Figura 9 e na Tabela 4. Como pode ser verificado na Figura 9, o gráfico da série em questão se assemelha à série original.

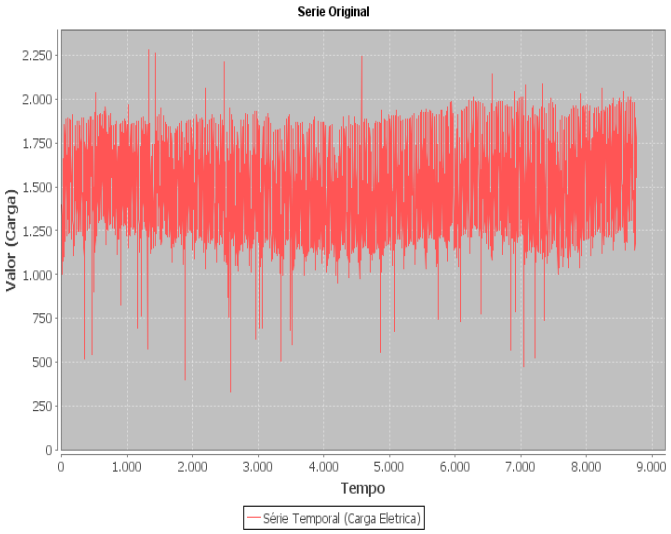

Figura 9 – Série de dados tratados.

Tabela 4 – Estatísticas da comparação entre a série normal e a corrompida

Valores Estatísticos	***Série Corrompida***	***Série Tratada***
Variância	402,6523	232,3181
Desvio Padrão	20,0661	15,2419
Média	1462,6793	1464,2555
Máximo	6035,20	2283,10
Mínimo	2,90	332,30

Contudo, visivelmente ainda pode ser constatada a existência de alguns *outliers*. O que se deve ao fato de que os métodos de detecção utilizados não obtiveram generalidade suficiente na combinação de seus componentes mediante a variabilidade da amostra especificada. Para tanto, a partir desses resultados, é possível obter uma série de dados menos conturbada realizando uma segunda filtragem destes dados, utilizando o mesmo método descrito neste trabalho. Os resultados da detecção podem ser observados na Figura 10.

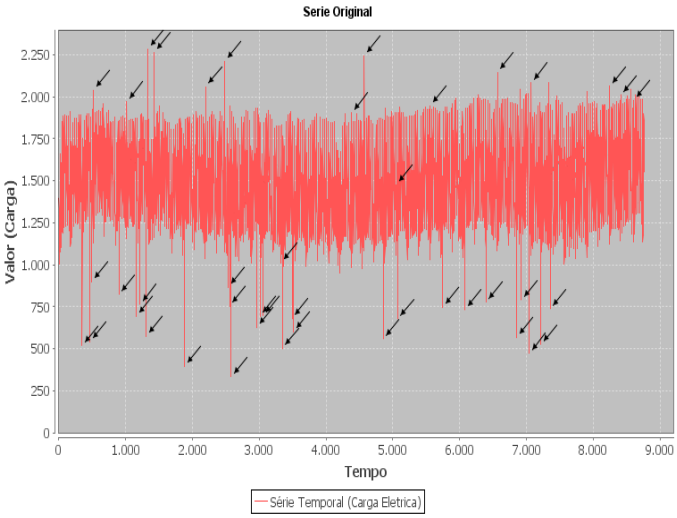

Figura 10 – Série de dados tratados.

Ao final do procedimento descrito acima, é realizado o tratamento dos dados corrompidos detectados da filtragem inicial realizada. Assim obtemos uma série de dados graficamente muito parecida com a série original, como pode ser observado na Figura 11 e na Tabela 5.

O ERM da série corrompida em relação à série original foi de 5,85%, já o ERM da série original em relação à série tratada calculado foi de 0, 8377%. Fato que mostra um ganho muito significativo em relação à originalidade da série de carga em questão.

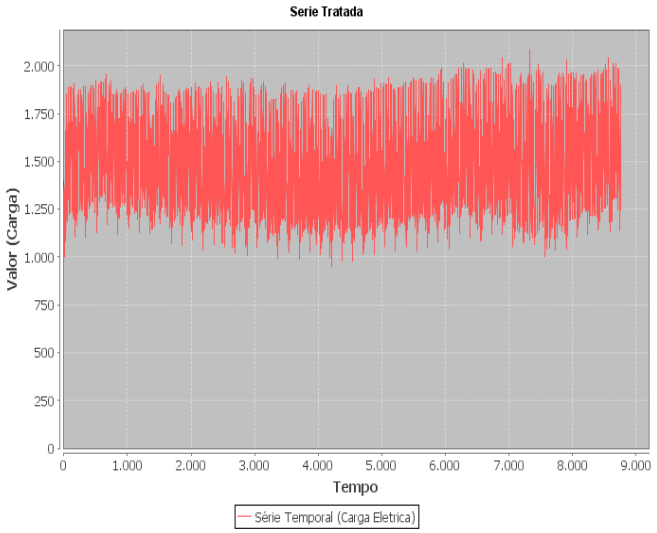

Figura 11 – Série de dados tratados.

Tabela 5 – Estatísticas de comparação entre a série normal e a corrompida.

Valores Estatísticos	***Série Original***	***Série Tratada***
Variância	227,7788	226,6226
Desvio Padrão	15,0923	15,0539
Média	1465,0067	1461,7234
Máximo	2042,10	2042,10
Mínimo	951,00	951,00

V. CONCLUSÃO

Este trabalho apresentou um estudo sobre identificação e tratamento de *outliers* em dados de demanda de carga elétrica. Com os resultados obtidos foi possível perceber a capacidade do modelo *Ensemble* proposto de generalização imparcial das análises do conjunto de dados, mesmo que tenha sido necessário realizar duas etapas de detecção e tratamento dos dados no experimento. O método de detecção *Chave Inglesa* propost, obteve um rendimento acima da média mediante aos outros métodos utilizados, onde a capacidade de identificação de outliers verdadeiros em relação aos falsos-positivos detectados foi a melhor dentre os métodos utilizados.

A principal contribuição deste trabalho foi determinação do modelo combinação dos métodos de detecção de *outliers*, que obtiveram uma maior capacidade de filtragem de dados aberrantes verdadeiros, onde juntamente com a abordagem de tratamento de dados possibilitou produzir melhores resultados.

Os resultados obtidos nas simulações através do *Ensemble* foram muito satisfatórios, o nível de acerto de *outliers* verdadeiros foi alto, contando também com níveis muito baixos de detecção de falsos-positivos, frente a outros trabalhos existentes na literatura. Para propostas futuras, sugere-se a realização de novas simulações acrescentando outros tipos de métodos de detecção e outras técnicas de tratamento de dados a fim de obter resultados com maior nível de precisão. Pode-se também realizar uma análise refinada dos dados, ou seja, realizar a verificação em intervalos menores da série. Espera-se que com isso possam ser descartados alguns erros de variação que são comuns concatenando uma série a outra dentro do espaço amostral.

AGRADECIMENTOS

Agradecemos à Universidade Federal de Alfenas (UNIFAL-MG), ao Laboratório de Inteligência Computacional (LInC) e ao CNPq pelo suporte financeiro indispensável a esta pesquisa.

REFERÊNCIAS

[1] Onoghojobi, B. (2010); An Instant of Performance Criteria for Outlier Identification;

[2] Anscombe, F.J.(1960); "Rejection of outliers".Technometrics;I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350;

[3] Last & Kandel, (2001); Automated detection of outliers in real-world data;

[4] Filzmoser P. (2005); Identification of Multivariate Outliers: A Performance Study;

[5] Elsa M. Jordaan, Dow Benelux BV, Guido e Smits, Dow Benelux BV; (2004); Robust Outlier Detection using SVM Regression;

[6] Laurikkala J., Juhola M. & Kentala E.; (2000); Informal Identification Of Outliers In Medical Data;

[7] Zhu Cui, KitagawaHiroyuki, Papadimitriou Spiros, Faloutsos Christos, (2004), Example-based Outlier Detection with Relevance Feedback;

[8] Filzmoser P. (2004); A Multivariate Outlier Detection Method;

[9] Baragona R., Calzini C., Battaglia F.; (2007); Genetic Algorithms For Outlier Identification Of Additive And Innovational Type In Time Series;

[10] Chiang Jung-Tsung (2008); The Algorithm for Multiple Outliers Detection Against Masking and Swamping Effects

[11] Lukashevich H., Nowak S., Dunker P.; (2009); Using One-Class Svm Outliers Detection For Verification Of Collaboratively Tagged Image Training Sets;

[12] Prabhjot Kaur, Anjana Gosain (2009), Improving the performance of Fuzzy Clustering algorithms through Outlier Identification;

[13] Haykin, S. (2001). Redes Neurais - Princípios e Praticas . Bookman, Porto Alegre, Brasil;

[14] Tukey John (1977); Understanding Robust and Exploratory Data Analysis;

[15] Alfassi, Z. B., Borger, Z. & Ronen (2005); Y. *Statistical Treatment of Analytical Data*. USA and Canada: CRC Press LLC,. 273 p.

[16] Hampel F. R., (1971) ”A general qualitative definition of robustness,” Annals of Mathematics Statistics, 42, 1887–1896;