Intelligent Models to Identification and Treatment of Outliers in Electrical Load Data

R. M. Salgado, T. C. Machado and T. Ohishi

***Abstract*— Electrical Load data are stored in each time interval generating big databases with high dimensional data. Each data stored contains significant information that can assist the planning and operation of electrical systems. In the data analysis step, many aspects must be considered such as the data consistency and the identification and treatment of outliers. This is a critical step because data quality is directly reflected in the results of the planning and operation of electrical systems. This paper proposes two models for the identification and treatment of outliers in electrical load data. The first model was built using the ensemble technique through a combination of individual models. The second model was created from an expert system that uses a rules database to detect outliers. The processing of the outliers detected is conducted through a combination of non-outliers load in the same time interval. To evaluate the performance, the models were applied in a historical load database measured in the Northeast of Brazil during year of 2006. The results showed that the proposed models showed satisfactory results in terms of detection as well in the treatment of outliers.**

***Keywords*— Outlier Detection, Outliers Treatment, Ensembles, Expert Systems, Electrical Load Data.**

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. 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.

Outliers são geralmente comuns em base de dados, no entanto, sua verdadeira causa de ocorrência é geralmente desconhecida para os analistas de dados [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 aleatórios. Mesmo com ferramentas visuais altamente sofisticadas, na maioria dos casos se torna inviável analisar os dados utilizando processos identificação manual [4], [6]. Fato que se atribui a grande quantidade de informações e alta dimensionalidade dos dados.

R. M. Salgado, Instituto de Ciências Exatas Universidade Federal de Alfenas, Alfenas, Minas Gerais, Brazil, ricardo@bcc.unifal-mg.edu.br

T. C. Machado, Instituto de Ciências Exatas, Universidade Federal de Alfenas, Alfenas, Minas Gerais, Brazil, a08035@bcc.unifal-mg.edu.br

T. Ohishi, Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, Caminas, São Paulo, Brazil, taka@densis.fee.unicamp.br

De acordo com a 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 espaço com alta dimensionalidade, os dados se tornam cada vez mais independentes e a noção de proximidade e densidade perde o seu significado, logo seu poder de separação é minimizado o bastante para se tornar ineficaz [5], [12].

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 um especialista. No entanto, para um grande conjunto de dados este método torna-se pouco eficiente e muito demorado.

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), que usa vários modelos de complexidade variável para detectar os *outliers* com base nas características dos vetores de suporte, calculados pelos 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. O modelo proposto compara a diferença entre a robusta distribuição empírica do quadrado das distâncias e a função de 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) e não-*outliers* (falsos positivos). Entretanto o método Becker torna-se preferível pela sua baixa taxa de classificação de não-*outliers* (falsos positivos). 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 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 cálculo de legitimidade dos elementos do conjunto. Neste processo iterativo baseado em algoritmos genéticos, 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 busca com o objetivo de calcular a função prevista, e o outro, contém 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. 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. O experimento demonstra uma técnica para rejeitar automaticamente os *outliers* a partir dos dados de treinamento, utilizando este método de detecção oferecida pela ferramenta em questão.

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 dois métodos para detecção de *outliers*, um baseado em curvas representativas para cada dia da semana que é baseado em diferenças absolutas dos dados e outro modelo baseado em um combinador Ensemble [13]. O Ensemble proposto será criado a partir das seguintes técnicas: *BoxPlot, Teste de Chauvenet, Teste ZScore, Delete Outlier e Teste de Hampel*. Para validar os modelos propostos foram utilizados dados de carga elétrica reais pertencentes a empresas localizadas na região sudeste do Brasil.

O conteúdo deste trabalho encontra-se organizado da seguinte forma: na seção II são apresentados os métodos de identificação de *outliers* utilizados. A seção III é aborda a metodologia utilizada para realização dos experimentos. Na Seção IV é feita uma discussão sobre os dados utilizados nas simulações realizadas bem como apresenta-se os resultados obtidos a partir da aplicação dos modelos propostos para a identificação e tratamento da série de carga elétrica. A seção V finaliza o trabalhando apresentando as considerações e discussões finais.

II. FUNDAMENTAÇÃO TEÓRICA

Para realizar uma previsão de carga de forma adequada é recomendado que os dados históricos não apresentem inconsistências. Segundo Hawkins [17] um dado inconsistente ou *outlier* é uma observação que se desvia das demais, a ponto de suspeitar-se que tenha sido gerada por um mecanismo diferenciado ou aleatório. Já para Barnett e Lewis [18], um *outlier* é uma observação (ou subconjunto de observações) que parece ser atípica em relação ao restante do conjunto de dados. Dependendo da sua natureza, os *outliers* podem causar um efeito substancial na análise dos dados. A Fig. 1 mostra curvas de carga diária de diversos barramentos contendo *outliers*.

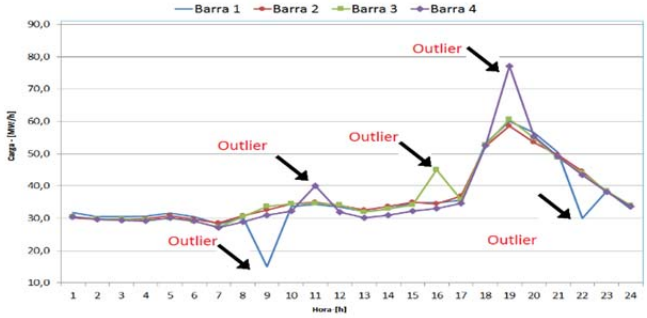

Figura 1. *Outlier*: curva de carga ativa.

Detectar dados inconsistentes não é uma tarefa trivial, como os *outliers* geralmente são gerados por um evento aleatório não é possível mapeá-los por meio de uma equação determinística. Na literatura especializada existem diversos modelos que objetivam a detecção de dados inconsistentes em séries temporais. Os modelos vão desde a observação visual da série até a utilização abordagens baseadas em inteligência computacional. De acordo com a literatura, não há um critério uniforme e determinístico que possa ser usado para definir se um dado de carga suspeito seja realmente um *outlier*. Assim, como não existe um único modelo capaz de detectar todos os *outliers* de uma determinada série de dados o modelo de ensemble hibrido proposto irá combinar várias técnicas por meio de um modelo híbrido baseado em *ensembles* [19] e [20]. Nesta seção, serão apresentados brevemente a descrição teórica dos métodos que serão utilizados para compor o *ensemble* proposto.

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 superior 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.

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 artigo foi escolhido como 0,05. O nível de significância 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 como estimativas para estabelecer este limite. O método de Hampel demonstra-se muito eficaz na detecção de *outliers*.

III. MODELAGEM PROPOSTA

O modelo proposto neste artigo consiste de duas fases: ***1) Identificação de outliers*** e ***2) Tratramento dos outliers***. Abaixo serão descritas detalhadamente cada uma das fases.

1) Idendificação de Outliers

A fase se identificação será realizada através de dois modelos: O primeiro consiste em uma abordagem baseada em *Ensembles* que utiliza como componentes os 5 modelos apresentados na seção II a saber: *BoxPlot, Teste de Chauvenet, Teste Z-Score, Delete Outlier e Teste de Hampel.* A segunda abordagem consiste em um sistema especialista que realiza

comparações entre curvas características visando detectar os *outliers* presentes nas séries de carga elétrica.

Em sequência as duas abordagens para identificação de *outliers* proposta neste trabalho serão detalhadas.

A. 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.

A implementação realizada neste trabalho consiste em combinar os resultados de diferentes modelos de detecção de *outliers*. De acordo com um número mínimo de métodos com validações positivas quanto à constatação deste dado com um *outlier*, ele pode ou não ser rotulado como um *outlier* (Quadro I e Fig. 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 analisado de maneira genérica. Como consequência dessa abordagem de análise do conjunto de dados, os resultados tornam-se confiáveis.

QUADRO I
PSEUDO-CÓDIGO – ENSEMBLE PROPOSTO

Seja D[i] um dado pertencente ao 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 D[i] como *outlier* for maior ou igual à taxa de corte definida pelo ensemble, então o dado é classificado como um *outlier*.

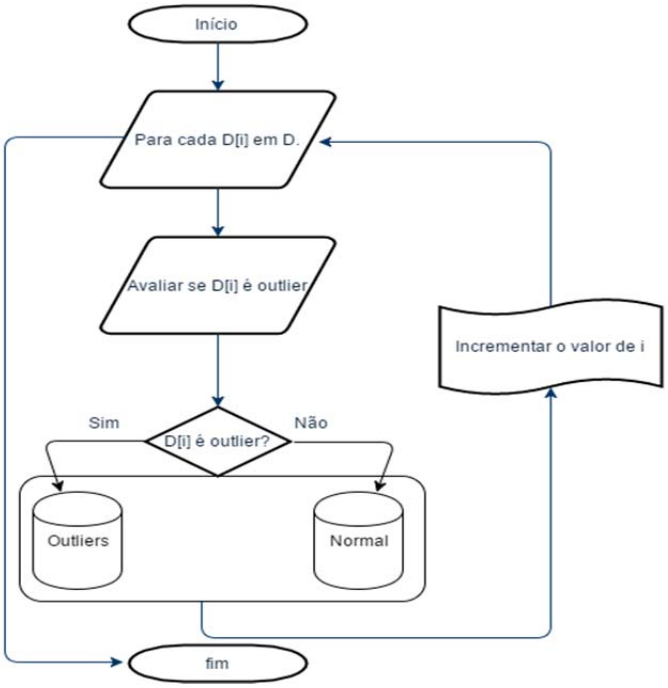

Figura 2. *Fluxograma*: *Ensemble*.

B. Sistema Especialista

Para identificar os *outliers* o sistema especialista encontra curvas típicas para as seguintes classes de dias da semana: **segundas-feiras, dias úteis, sábados, domingos e feriados**. O próximo passo foi calcular uma curva média para cada uma das 4 classes. Para evitar a influência de *outliers* a curva média será encontrada através da medida de posição mediana. Os *outliers* foram encontrados através da comparação entre a curva do dia analisado e a curva média da respectiva classe através da diferença horária entre as cargas. Por exemplo, se o dia analisado for uma quarta-feira (dia útil) então a diferença entre a curva média será calculada para cada hora do dia. Se a diferença for superior a um determinado valor previamente definido então o dado será considerado um *outlier*. A lógica do sistema especialista para detecção de *outliers* está apresentada no Quadro II e na Fig. 3.

QUADRO II
PSEUDO-CÓDIGO – SISTEMA ESPECIALISTA

Seja **St** uma série temporal de carga elétrica com d dias de discretização n.
Passo 0: calcular a curva de carga média para uma das classes pré-definidas; Definir o nível de tolerância ε;
Passo 1: Para cada um dos dias d fazer:
2.1 – Definir o tipo de dia analisado (segundas, dias úteis ou outros);
2.2 – Comparar a carga da curva média em cada um dos n pontos do dia;
2.3 – Se a diferença entra a carga da curva média e do dia analisado d for maior que ε então o dia d em análise será considerado um *outlier*.
Passo 2: Retornar a lista de dias que possuem *outliers* e a posição de cada um deles.

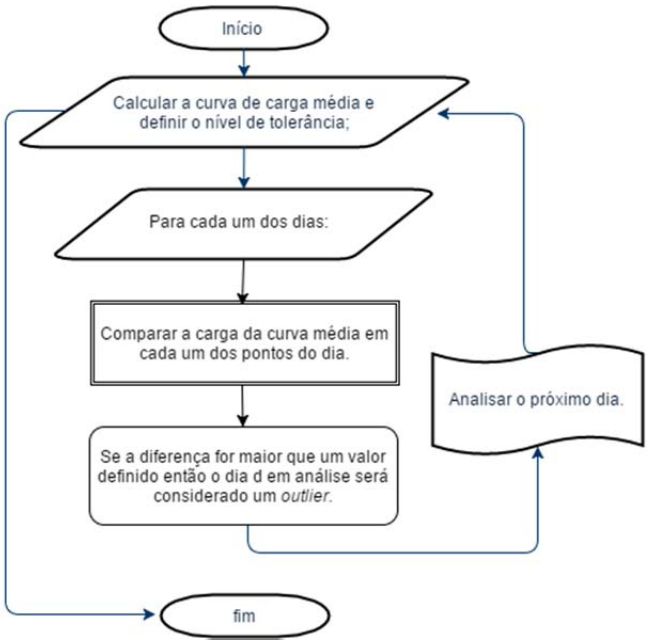

Figura 3. *Fluxograma*: Sistema Especialista.

2) Tratamento dos Outliers

Os dados inconsistentes identificados na fase anterior do modelo, devem ser tratados, isto é, substituídos por um novo dado que represente adequadamente comportamento da série em análise. Em uma série temporal quando um dado for considerado um *outlier*, é possível tratá-lo de forma eficiente

com base no histórico de dados. Por exemplo, em séries de carga diária pode-se corrigir um dado inconsistente levando em consideração medições realizadas em um mesmo intervalo de tempo *t* e em dias do mesmo tipo *d*.

Supondo que um dado carga medido em um determinado dia *d* em uma determinada hora h seja identificado como *outlier*. Este modelo visa a substituição do *outlier* utilizando operadores estatísticos baseados em medidas de posição, entre os quais pode-se citar a média (simples ou ponderada) e/ou a mediana. Assim utilizou-se, como base para a substituição, dados do mesmo tipo ao dado corrompido em outros instantes de tempo, os quais não foram identificados previamente como *outliers*. Por exemplo, se o dado corrompido identificado for às 14h de uma Quarta-feira a ideia é substitui-lo usando a mediana da carga medida às 14h em outras Quartas-feiras consideradas normais. A Fig. 4 mostra o exemplo de um dado D[k] corrompido, onde D[k-1] e D[k+1] são dados normais. A Fig. 5 mostra um exemplo no qual dado D[k] foi tratado pelo método de através da média aritmética simples calculada entre D[k-1] e D[k+1], que são dados normais.

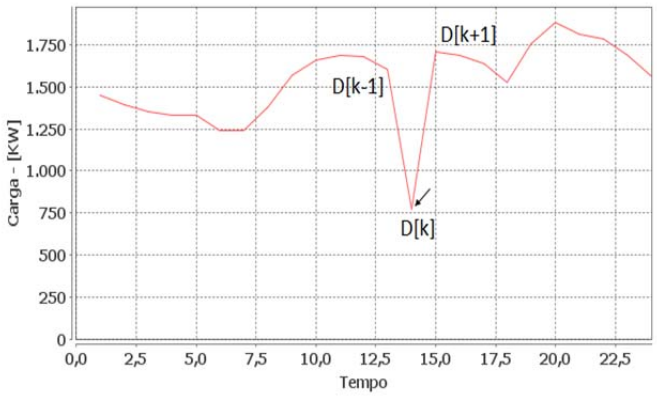

Figura 4. Dado corrompido com vizinhos D[k-1] e D[k+1] normais.

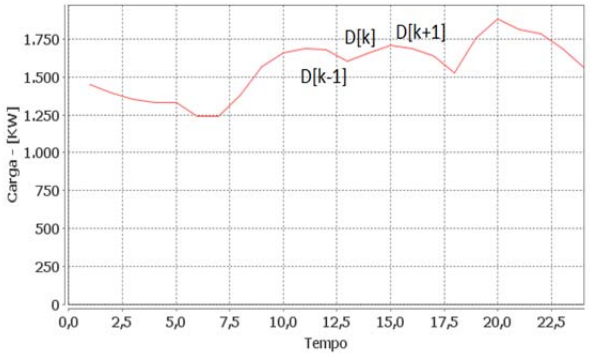

Figura 5. Dado tratado com vizinhos D[k-1] e D[k+1] normais.

IV. ESTUDO DE CASOS

Os estudos de caso de análise e tratamento de dados neste artigo foram feitos com cargas elétricas reais medidas em uma empresa localizada no Nordeste do Brasil, contendo um histórico completo do ano de 2006, com medições horárias para cada dia do ano. A Fig. 6 mostra a série de dados de carga horária, com 24 horas durante os 365 dias do ano de 2006.

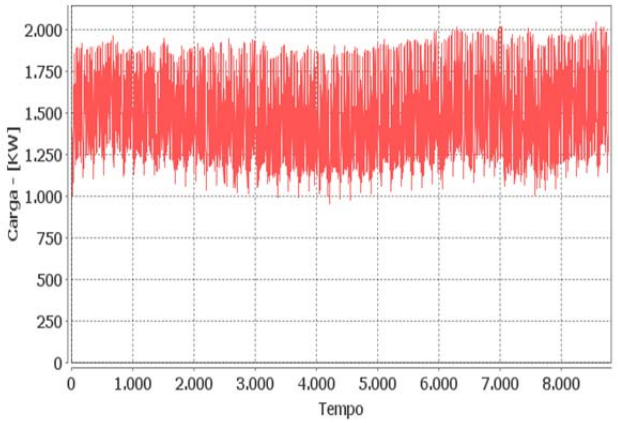

Figura 6. Série histórica de carga elétrica – 01/Jan a 31/Dez/2006.

A) Geração Artificial de Outliers

Para testar os modelos propostos foi utilizado um método de geração artificial de dados *outliers* sobre a série de carga histórica. Tal método se mostra necessário, pois em dados reais não se pode afirmar com certeza que cada um dos dados é um *outlier* ou não. Deste modo, foi desenvolvido um método que escolhe aleatoriamente elementos da série histórica, corrompendo-os através do algoritmo apresentado no Quadro III:

QUADRO III

GERADOR ARTIFICIAL DE OUTLIERS

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 120% a 200% de seu valor original;
Caso contrário, então o seu novo valor será aleatoriamente ajustado entre 10% a 50% de seu valor original.

Após a aplicação do algoritmo de geração artificial de *outliers* 7,5% dos dados da série original foram manipulados de forma a torna-los *outliers*, gerando um total de 657 dados corrompidos ao longo dá série de carga anual. Como pode ser observado na Fig. 7 e na Tabela I, a série corrompida difere significativamente da série original.

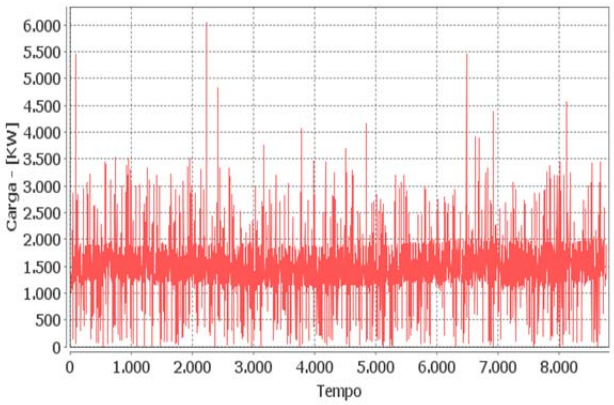

Figura 7. Série de dados corrompida.

TABELA I
ESTATÍSTICAS DE COMPARAÇÃO ENTRE A SÉRIE NORMAL E A CORROMPIDA

Valores Estatísticos	Série Original	Série Corrompida
Variância	227,79	402,65
Desvio Padrão	15,09	20,07
Média	1465,01	1462,68
Máximo	2.042,10	6035,20
Mínimo	951,00	2,90

B) Análise dos resultados do experimento

Com o objetivo de avaliar o desempenho do modelo de tratamento dos dados inconsistentes foi realizada a comparação dos dados normais aos dados tratados, fazendo a sua validação, através do erro relativo médio (ERM), apresentado na Equação 1:

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

No qual 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 bem como a taxa de identificação de falsos positivos. A taxa de corte utilizada no *Ensemble* que mostrou-se mais eficiente foi quando pelo menos dois modelos detectavam o dado em análise como *outlier*. Ou seja, caso um determinado elemento da série seja verificado como corrompido por dois ou mais métodos de detecção ele é rotulado como um *outlier* pelo *Ensemble* proposto.

A Tabela II, a Tabela III apresentam o desempenho dos métodos de detecção de *outliers*.

TABELA II
DESEMPENHO PERCENTUAL DOS MÉTODOS DE DETECÇÃO

Métodos de Detecção	*Outliers Verdadeiros Detectados*	*Outliers Falsos-positivos Detectados*
BoxPlot	96,80%	2,59%
Teste Chauvenet	36,23%	0,00%
Delete Outlier	60,73%	0,15%
Teste de Hampel	75,34%	0,00%
Sistema Especialista	92,09%	0,76%
Teste ZScore	33,33%	0,00%
Ensemble	92,69%	0,15%

TABELA III
DESEMPENHO NÚMERICO DOS MÉTODOS DE DETECÇÃO

Métodos de Detecção	*Outliers Verdadeiros Detectados*	*Outliers Falsos Positivos Detectados*
BoxPlot	636	17
Teste Chauvenet	238	0
Delete Outlier	399	1
Teste de Hampel	495	0
Sistema Especialista	605	5
Teste ZScore	219	0
Ensemble	609	1

Analisando os resultados é possível 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 nos testes realizados. O sistema especialista proposto foi capaz de identificar uma alta taxa de *outliers* verdadeiros, mas não foi tão eficaz quanto o *Ensemble* na detecção de falsos-positivos. 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. A Fig. 8 ilustra a detecção do *Ensemble* proposto.

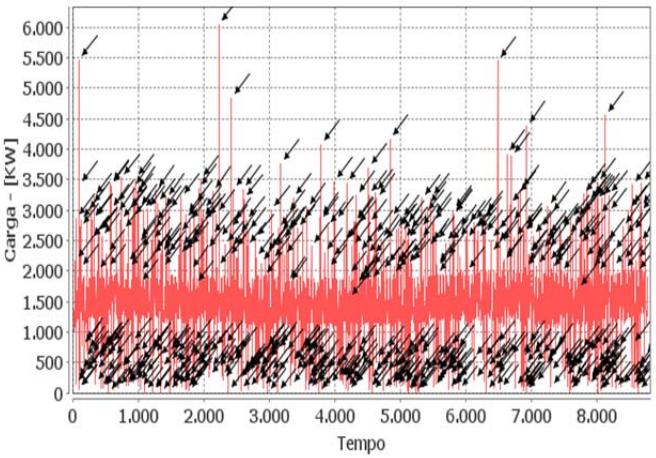

Figura 8. Série de dados com os *outliers* detectados.

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

O conjunto de *outliers* detectado pelo Ensemble foi utilizado para medir a eficiência do modelo de tratamento de dados. O gráfico gerado a partir dos dados tratados pode ser visto na Fig. 9. A Tabela IV apresentam as medidas de posição da série tratada após a aplicação do modelo de tratamento proposto. Como pode ser verificado na Fig. 9, o gráfico da série tratada é visualmente semelhante à série original.

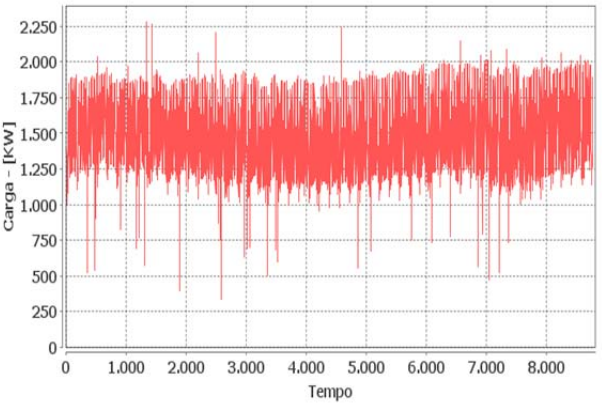

Figura 9. Série de dados tratados.

TABELA IV
ESTATÍSTICAS DA COMPARAÇÃO ENTRE A SÉRIE TRATADA E A CORROMPIDA

Valores Estatísticos	*Série Corrompida*	*Série Tratada*
Variância	*402,65*	*232,32*
Desvio Padrão	*20,06*	*15,24*
Média	*1462,68*	*1464,25*
Máximo	*6035,20*	*2283,10*
Mínimo	*2,90*	*332,30*

Contudo, observando a Fig. 9 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 as abordagens de identificação propostas neste trabalho. Os resultados da execução dos modelos de detecção sobre a série tratada apresentada na Fig. 9 podem ser observados na Fig. 10.

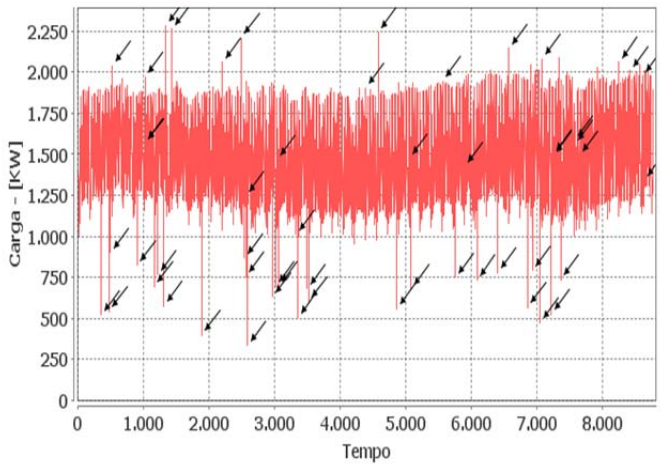

Figura 10. Detecção dos *outliers* da série de dados tratados.

A nova lista com os *outliers* detectados pelo *Ensemble* na série apresentada na Fig. 10 foi aplicado o modelo de tratamento dos dados corrompidos. O resultado foi a obtenção de uma série de dados visualmente e estatisticamente equivalente à série original, como pode ser observado na Fig. 11 e na Tabela V.

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, 84% (após o segundo tratamento). Fato que mostra um ganho muito significativo em relação à originalidade da série de carga em questão.

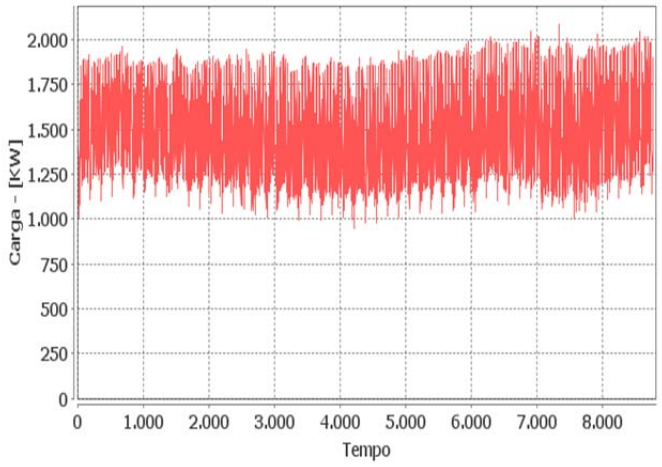

Figura 11. Série de dados tratados (após segundo tratamento).

TABELA V
ESTATÍSTICAS DE COMPARAÇÃO ENTRE A SÉRIE ORIGINAL E A CORROMPIDA APÓS SEGUNDO TRATAMENTO

Valores Estatísticos	*Série Original*	*Série Tratada*
Variância	*227,78*	*226,62*
Desvio Padrão	*15,09*	*15,054*
Média	*1465,07*	*1461,72*
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* de generalização 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 baseado em *Sistema Especialista*, 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 baseado na 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**) ao CNPq pelo suporte financeiro indispensável a esta pesquisa e ao Operador Nacional do Sistema Elétrico (ONS) pelo compartilhamento dos dados.

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;

[17] Hawkins, D. (1980). Identication of outliers. Chapman & HaIl, London.

[18] Barnett, V. e Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons, 3nd edition.

[19] L. Hansen and P. Salamon. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(10), 993-1001, 1990.

[20] A. Sharkey (Ed.). Combining artificial neural nets: Ensemble and modular multi-net systems, Springer-Verlag, London, 1999.

Ricardo Menezes Salgado é Bacharel em Matemática pela Universidade Federal de Viçosa (2002), mestre (2004) e doutor (2009) em Engenharia Elétrica pela Universidade Estadual de Campinas . Atualmente é professor da Universidade Federal de Alfenas, lotado no curso de Ciência da Computação. Atua na área inteligência computacional, sendo consultor e executor de diversos projetos em grandes corporações. Tem experiência como desenvolvedor de sistemas inteligentes de suporte a decisão (plataforma JAVASE) e em análise de dados com ênfase em séries temporais. Ao longo de sua carreira tem atuado principalmente nos seguintes temas: inteligência computacional, previsão de séries temporais, mineração de dados, otimização e reconhecimento de padrões aplicados em diversos setores.

Tadeu Carvalho Machado é graduado em Ciência da Computação pela Universidade Federal de Alfenas (UNIFAL-MG). Atualmente trabalha na Telefônica/Vivo na área de banco de dados de alto desempenho.

Takaaki Ohishi é graduado em Engenharia Elétrica pela Universidade de São Paulo (1978) e doutor em Engenharia Elétrica pela Universidade Estadual de Campinas (1990). Atualmente é professor livre docente da Universidade Estadual de Campinas. Tem experiência na área de Engenharia Elétrica, com ênfase em Sistemas Elétricos de Potência, atuando principalmente nos seguintes temas: pré-despacho, despacho econômico, otimização, previsão de carga e sistemas de potência.