UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

PAULO PIRES FERNANDES NETO

APLICAÇÃO DE MODELO HÍBRIDO ARIMA-RNMC PARA PREDIÇÃO DE SÉRIES TEMPORAIS DE CARGAS DE ENERGIA DAS 5 REGIÕES BRASILEIRAS

CAMPINA GRANDE - PB

2014

PAULO PIRES FERNANDES NETO

APLICAÇÃO DE MODELO HÍBRIDO ARIMA-RNMC PARA PREDIÇÃO DE SÉRIES TEMPORAIS DE CARGAS DE ENERGIA DAS 5 REGIÕES BRASILEIRAS

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Engenheiro Eletricista.

ORIENTADOR: Prof. Dr. Hiran de Melo

CAMPINA GRANDE - PB

2014

PAULO PIRES FERNANDES NETO

APLICAÇÃO DE MODELO HÍBRIDO ARIMA-RNMC PARA PREDIÇÃO DE SÉRIES TEMPORAIS DE CARGAS DE ENERGIA DAS 5 REGIÕES BRASILEIRAS

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Engenheiro Eletricista.

ORIENTADOR: Prof. Dr. Hiran de Melo

Aprovado em ____ de _____ de 2014

BANCA EXAMINADORA

_____

Prof. Dr. Hiran de Melo

Universidade Federal de Campina Grande

Orientador

_____

Professor Convidado

Universidade Federal de Campina Grande

Avaliador

AGRADECIMENTOS

A Deus por ter me dado forças e iluminado meus caminhos para que eu pudesse concluir mais uma etapa da minha vida.

A meus pais e minha irmã, por serem tão dedicados, amorosos e por ter me dado condições para me tornar o profissional e o homem que sou.

A toda a família Fernandes e família Peixoto por todo apoio e suporte.

A Priscila Araújo por todo apoio, amor, carinho e atenção dados nessa caminhada.

A Raul por ser um grande amigo e praticamente um irmão, por ter me acompanhando muitos anos da minha vida, inclusive os de vida acadêmica.

Ao professor e orientador Dr. Hiran de Melo, pela grande contribuição no desenvolvimento deste trabalho, por ser um grande amigo além de um mestre em transmitir conhecimento, onde pude conhece-lo além da vida acadêmica e conviver no dia-dia.

Ao professor e coorientador José Carlos Reston Filho pela oportunidade que me concedeu, pela amizade e pelo treinamento, que me permitiu crescer profissionalmente.

Àqueles, que não por menor importância, não foram citados, mas também tiveram grande contribuição na realização desse grande sonho de ser engenheiro.

RESUMO

Este trabalho propõe um método para a predição futura de médio prazo da série temporal de cargas de energia das 5 regiões brasileiras. O modelo criado tem por objetivo estimar até 5 passos (semanas) a frente, o montante total de energia requisitado pelo conjunto de instalações das regiões do Brasil. Para tanto, utiliza uma abordagem híbrida, conjugando uma técnica clássica de forecasting, o ARIMA, com as redes neurais artificiais de múltiplas camadas. Os benefícios de um modelo híbrido são os de combinar as melhores características de cada técnica, promovendo assim um modelo robusto capaz de capturar as não linearidades de séries temporais complexas, resultando assim em uma previsão mais exata. Os resultados são conclusivos em apontar a metodologia proposta como técnica efetiva na previsão de valores futuros da carga de energia, variável importante no proceso de tomada de decisão na gestão do Sistema Interligado Nacional-SIN.

PALAVRAS-CHAVE— Mineração de Dados, Previsão de séries temporais, Rede neural de mútiplas camadas, modelo ARIMA.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com os avanços da informática, a grande maioria das operações e atividades das instituições privadas e públicas são hoje registradas em banco de dados e se acumulam com o tempo. Utilizar técnicas de mineração de dados é uma forma eficaz para extração de conhecimento, busca de padrões e tendências em grandes volumes de dados e até gerar regras para fazer previsões e correlacionar dados, que podem ajudar na tomada de decisões das empresas.

As regras para a compra de energia pelas distribuidoras exigem altíssimo nível de exatidão nas previsões de carga. Em maior ou menor grau, todas as distribuidoras brasileiras precisaram aperfeiçoar o processo de elaboração das previsões, tendo em vista o novo modelo para o setor elétrico. O setor elétrico sempre trabalhou com duas modalidades de previsão: o de longo prazo, necessário para viabilizar o planejamento da construção de usinas geradoras, e o de curtíssimo prazo, dependente de variáveis sobre as quais não é possível ter controle, como mudanças súbitas de temperatura. O novo modelo para o setor elétrico projetou um ambiente mais competitivo, em que a previsão de carga se tornou fator absolutamente crítico para as distribuidoras. Isto porque a compra de energia passou a ser feita exclusivamente por meio de leilões, com base no critério do menor preço.

A busca da exatidão das previsões é bastante importante porque a lei prevê que a distribuidora utilize 100% da energia que contratou, e as margens de variação permitidas são muito pequenas. Na prestação de contas anual à Agência Nacional de Energia Elétrica (ANEEL), caso a empresa tenha feito uma compra inferior em mais de 4% à energia distribuída, poderá incorrer em penalidades. O objetivo do governo, nesse caso, foi estabelecer regras para assegurar que não falte energia. Em caso inverso, ou seja, se a distribuidora comprar energia a mais, ela somente pode repassar esse custo às tarifas se a margem de erro for de até 3%. Se tiver adquirido energia em um percentual acima deste, a distribuidora deve arcar com os custos sem transferi-los ao consumidor. Com essa regra, o governo quis garantir a menor tarifa possível. Então, em qualquer um dos erros, seja para mais ou para menos, acarretará em prejuízo para as empresas.

A previsão pode ser distribuída em quatro tipos: curtíssimo prazo, curto prazo, médio prazo e longo prazo. Elas se diferenciam de uma pra outra, quanto a classificação do período de previsão. Para a previsão a curtíssimo prazo, o interesse é de alguns

minutos até uma hora à frente. Para a previsão a curto prazo, estima-se uma faixa de 24 horas até uma semana à frente. Já na previsão a médio prazo, a faixa se estende a alguns meses. Finalmente, a previsão a longo prazo, se refere a períodos superiores a um ano.

Neste trabalho, será desenvolvida a previsão de séries temporais para carga de energia para médio prazo. A previsão de médio prazo, que pode variar entre semanas e meses e é, em geral, realizada para que se possa definir e organizar as manutenções que precisam ser efetuadas no sistema de geração e distribuição de energia, que podem ser classificadas como manutenções preventivas ou emergências, além disso, o conhecimento futuro da carga exerce um papel muito importante no planejamento de fluxo de potência, na análise e controle da segurança dos sistemas de energia elétrica, na operação econômica e no planejamento de expansão do sistema elétrico.

Na literatura, podemos encontrar técnicas que se destacam para a previsão de cargas de energia, tais como: Alisamento exponencial, Regressão linear, Filtro de Kalman, ARIMA (Auto Regressive Integrated Moving Average) de Box & Jenkins e as Redes Neurais Artificiais.

Conhecer o comportamento da carga de energia é muito importante na tomada de decisões em sistemas elétricos de potência. O comportamento da carga é influenciado por diversos fatores e de forma complexa, muitas vezes não-linear, fatores esses tais como: temperatura, horários, dias da semana, dias atípicos (feriados, greves), etc.

As redes neurais artificiais vem demonstrando grande eficiência para obtenção da previsão precisa de cargas de energia elétrica, aliado ao hibridismo, se torna uma técnica ainda mais potente que usada isoladamente. Um exemplo de modelo híbrido bastante utilizado na literatura é o modelo no qual separa a série temporal em uma parte linear e outra não linear, em que o ARIMA é responsável pela parte linear da série e a rede neural pela parte não linear dos resíduos do ARIMA.

A previsão de cargas elétricas futuras se demonstra viável, tendo em vista a competitividade de um mercado em franca e contínua expansão e, visando principalmente, minimizar a perda às empresas fornecedoras de energia (HIPPERT et al., 2001).

Desta forma, torna-se necessária a utilização de um sistema com capacidade de alcançar o mínimo possível de erro, através de técnica precisa, possibilitando a maximização dos lucros e diminuindo a possibilidade de perdas, através de estratégias desenvolvidas pelo próprio sistema (LOPES, 2005).

1.1 Objetivo

Toda vez que falamos em características lineares, trata-se de modelos que podem ser mapeados pela equação y = ax + b. O ARIMA em especial faz um mapeamento de entradas do passado influenciando valores do futuro a partir de um mapeamento linear que usa os correlogramas, com valores de autocorrelação e autocorrelação parcial. Assim, as características lineares da série são aquelas que podem ser explicadas pelas relações do correlograma.

Já em relação as não linearidades, a rede neural possui uma função sigmoidal como função de ativação. Ela possui uma leve não-lienaridade nas suas extremidades. No entanto, como uma rede neural possui diversos neurônios ela pode mapear grandes não-linearidades a partir da força do conjunto dos neurônios.

A proposta deste trabalho é a utilização de um sistema híbrido, combinando uma rede neural artificial Perceptron Multicamadas, utilizando o algoritmo backpropagation, com um modelo ARIMA, que irá nos possibilitar a obtenção da previsão de cargas de energia até 5 passos (semanas) a frente para todas as regiões do Brasil. O modelo é desenvolvido para captar as características lineares e não-lineares da série temporal.

1.2 Metodologia

Para este trabalho foi utilizado o software IBM SPSS Modeler. O IBM SPSS Modeler é um pacote de software abrangente no estilo workbench de mineração de dados que fornece uma maneira de construir modelos preditivos através de uma interface gráfica com o usuário, utilizando diversas técnicas avançadas de modelagem. Além disso, também tem a capacidade para executar várias operações de pré-processamento de dados para melhor preparar os dados a serem modelados.

O pacote também oferece suporte na forma de facilidades de visualização, processamento estatístico, navegação e suporte periférico para acesso e manipulação de dados que permitem o desenvolvimento rápido e fácil de experimentos de mineração de dados.

O modelo de mineração de dados adotado foi o CRISP-DM. A sigla significa Cross-Industry Standard Process for Data Mining (CRISP-DM). O método é composto de seis fases, que abordam os principais pontos da mineração de dados. Estas seis fases

formam um processo cíclico e cobrem todas as etapas da mineração de dados, inclusive a fases da incorporação dos resultados.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de Dados

Em décadas passadas, o crescente volume de informações e dados armazenados nas grandes organizações, era um grande problema a se resolver. Nos dias atuais com a evolução da tecnologia e do desenvolvimento de técnicas provenientes da mineração de dados, foi possível não só armazenar, mas também processar esses dados mais rapidamente e com maior precisão.

O problema que existe nos dias atuais se da na competência de análise desses dados, de forma a extrair informações que sejam úteis, que possam instruir a tomada de decisões e que, utilizando-se de técnicas de mineração de dados, possibilite até fazer previsões de situações futuras, como por exemplo, prever futuras vendas de um determinado produto, para um melhor planejamento da produção, o que nos possibilitaria minimizar as chances de ocorrer uma baixa produção e possivelmente o estoque de produtos não atender a grande demanda de vendas de determinado mês, ou então, em situação análoga, podemos prever uma superprodução e evitar o excesso de estoque, evitando assim adicionais custos para a empresa.

O custo de uma má ou boa qualidade dessas informações pode decidir o futuro para o sucesso de uma empresa. Com isto a utilização de técnicas de mineração de dados, torna-se essencial para descobrir padrões e tendências, permitindo a criação de modelos e o conhecimento melhor da realidade, além de guiar decisões de certeza limitada que não seriam possíveis de se obter por meios de técnicas padrões de estatística ou também por meio de ferramentas analíticas.

O processo capaz de descobrir conhecimento (informação) em banco de dados chama-se Knowledge Discovery in Database – KDD. Consoante Fayyad (1996), este processo foi proposto para referir-se as etapas que produzem conhecimento a partir dos

dados, em 1989. Neste processo, a mineração de dados é a fase que transforma os dados em informação.

2.2 O processo de mineração de dados

Conforme Fayyad (1996), o processo tradicional para transformação de dados em informação, consiste em um processamento manual de todas essas informações por especialistas, para posteriormente produzirem relatórios, que deverão ser analisados, porém, esse processo manual, devido ao grande volume de informações, torna-se, impraticável.

A mineração de dados envolve vários algoritmos de extração de conhecimento, diferentes qualidades de dados, várias fontes de dados, também envolve o trabalho simultâneo de problemas complexos e diferentes formas de se medir o sucesso da mineração. Portanto, um roteiro pré-definido para mineração de dados, garante que todas as questões críticas e pontos importantes sejam abordados e que o minerador de dados não se perca em meio às complexidades.

O processo de mineração de dados modelo para o uso com o software IBM SPSS Modeler 14.2, é o Cross-Industry Standard Process for Data Mining (CRISP-DM).

2.3 A metodologia CRISP-DM

O modelo CRISP-DM, segundo CHAPMAN (2000) é composto por seis fases, que abordam os principais temas em mineração de dados. De acordo com OLSON (2008), as seis fases fazem parte de um processo cíclico. Essa metodologia, torna projetos de mineração de dados de larga escala mais rápidos, mais baratos, mais confiáveis e mais gerenciáveis. Estas seis fases cobrem o processo de mineração de dados por completo, inclusive na etapa de incorporação dos resultados.

Na figura 1, é mostrado o ciclo de vida de um projeto de mineração de dados. O ciclo externo na figura simboliza o ciclo natural de mineração de dados, um processo continua após a solução ter sido desenvolvida. A sequência de fases é não rígida, podendo ocorrer a transição para diferentes fases.

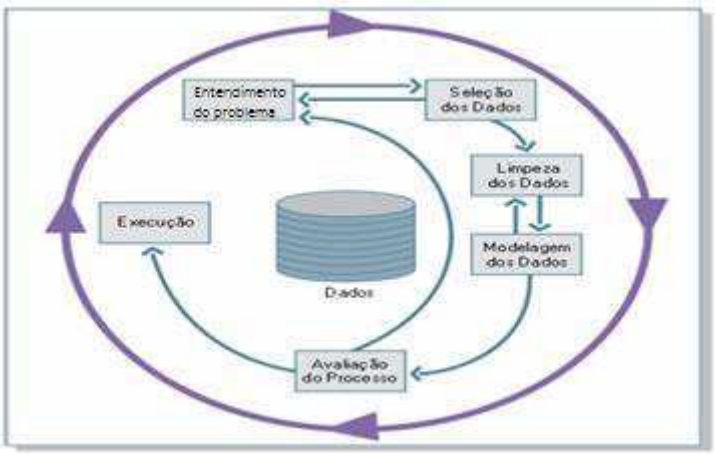

Figura 1 - Processo CRISP-DM.

2.3.1 Entendimento ou compreensão do problema (Problem Understanding)

Esta pode ser considerada a fase mais importante da mineração de dados. Nesta etapa, o foco não só é entender qual o objetivo que se deseja atingir com a mineração de dados, mas também traçar objetivos, avaliar a situação, determinar metas e elaborar um plano de projeto. No caso deste projeto, o foco é prever cargas de energia elétrica, através de técnicas de mineração, conjugando um modelo ARIMA e as redes neurais de múltiplas camadas, para que haja uma melhor exatidão das previsões.

2.3.2 Compreensão dos dados (Data understanding)

Tendo em mente que os dados fornecem a "matéria prima" para a mineração de dados, é preciso além de tudo entender os dados, visando a familiarização por parte do minerador, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formação de hipóteses. Esta etapa inclui a coleta inicial de dados, descrição dos dados, exploração dos dados e verificação da qualidade dos mesmos.

2.3.3 Preparação dos dados (Data preparation)

Devido às diversas origens possíveis, é muito comum que os dados não estejam preparados para que os métodos de mineração de dados sejam aplicados diretamente. Esta é a fase que visa a limpeza, transformação, integração e formatação dos dados da

etapa anterior. Nesta etapa é que são tratados os ruídos, dados inconsistentes e estranhos. Esta etapa abrange todas as atividades para a construção do conjunto de dados final a partir do conjunto de dados inicial. Esta costuma ser a fase que se exige mais esforço, correspondendo geralmente a 50% de todo o trabalho.

2.3.4 Modelagem dos dados (Modeling)

Basicamente, esta é a fase que as técnicas (algoritmos) de mineração de dados serão aplicadas, onde a escolha dessas técnicas irá depender dos objetivos desejados. Essas técnicas serão usadas para se extrair informações dos dados. Nesta etapa várias técnicas são utilizadas e seus parâmetros são calibrados e ajustados para se obter valores otimizados. Geralmente existem várias técnicas para o mesmo problema de mineração, algumas delas possuem requerimentos específicos na forma dos dados, consequentemente voltar para a etapa de preparação de dados pode ser necessário.

A maioria das diversas técnicas de mineração de dados baseiam-se em conceitos de estatística, reconhecimento de padrões, aprendizagem de máquina, classificação e clusterização.

2.3.5 Avaliação dos dados (Evaluation)

Considerada uma fase crítica do processo de mineração, essa etapa é necessária a participação de especialistas nos dados, tomadores de decisão e conhecedores do negócio. Esta etapa visa garantir que o modelo gerado atenda às expectativas da organização. Nesta etapa, faz-se avaliação de resultados, a revisão do processo de mineração de dados e a determinação das próximas etapas.

Visando obter a confiabilidade dos modelos, devem ser executados testes e validações, tais como validação cruzada, e indicadores para auxiliar a análise dos resultados precisam ser obtidos.

2.3.6 Implantação (Deployment)

É a etapa de colher os benefícios. Esta fase tem como objetivo a integração de seus novos conhecimentos aos processos de modo a resolver um problema de negócio. Esta fase inclui o plano de implantação, manutenção, monitoramento, elaboração de um relatório final e revisão do projeto.

Geralmente há uma tendência para que o processo caminhe de maneira linear através dos passos na ordem indicadas acima. Porém, pode ocorrer a influência mútua entre fases de maneira não linear.

Um ponto muito fundamental na mineração de dados é a sua natureza iterativa. Raramente é suficiente planejar um projeto, executa-lo e, em seguida dar o trabalho como finalizado. Trata-se de um esforço continuo. O conhecimento obtido com um ciclo de mineração de dados, geralmente levam a novas questões, novos problemas e novas oportunidades.

2.4 Técnicas de mineração

Em mineração de dados usam-se bases de dados para gerar modelos, que podem ser aplicados posteriormente para predição, classificação, avaliação e apoio a decisão. Aprendizagem de máquina ou modelagem é como são conhecidas, muitas das técnicas utilizadas em mineração de dados. Antes de se considerar qual técnica será usada, deve-se avaliar o negócio e os dados que temos em mãos, em determinados aspectos.

O primeiro aspecto é a disponibilidade dos dados, pois os dados precisam estar em um formato acessível. O segundo aspecto é a abrangência dos dados, pois para fazer um projeto de mineração de dados de valor, é importante que os dados contenham todos os elementos pertinentes. Outro aspecto é o ruído, que podem se apresentar na forma de discrepâncias (outliers) ou também a falta de dados. Quanto mais ruídos os dados tiverem, maior a dificuldade para se fazer previsões precisas. Em seguida, temos a questão da suficiência desses dados, se os dados têm uma boa cobertura dos possíveis resultados, serão alcançados resultados razoáveis mesmo para uma pequena quantidade de registros.

Por fim, temos o aspecto do conhecimento, pois nem sempre a pessoa encarregada de minerar os dados, conhece de forma satisfatória toda a problemática envolvida e a natureza dos dados, então é desejável que quanto mais pessoas envolvidas

que conheçam a base de dados e sua problemática envolvida, melhor será o resultado final.

2.4.1 Classificação

A classificação é uma das mais utilizadas técnicas de mineração de dados, que visa identificar a qual classe um determinado registro pertence. Essa técnica pode ser utilizada tanto para entender dados existentes, tanto para prever como novos dados irão se comportar. São comuns as tarefas de classificação de clientes em baixo, médio e alto risco de empréstimo bancário, de clientes potencialmente consumidores de um determinado produto a julgar pelo seu perfil, entre outras.

Os algoritmos de árvores de decisão são os mais utilizados para fins de classificação.

2.4.2 Estimação

A estimação é similar a classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Estimar algum índice é determinar seu valor mais provável diante dos outros índices semelhantes sobre os quais se tem conhecimento. Então, podemos estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após a análise dos dados, o modelo é capaz de dizer qual será o valor gasto por um consumidor novo.

Os algoritmos de regressão e as redes neurais são bastante utilizados nesses casos.

2.4.3 Previsão (Prediction)

A previsão está associada à avaliação de um valor futuro de uma variável a partir de dados históricos do seu comportamento no passado. Pode-se prever, por exemplo,o valor de uma ação de uma determinada empresa três meses adiante, podemos também

prever futuras vendas de um determinado produto para o planejamento e controle da produção, etc.

Os algoritmos utilizados para previsão são as redes neurais, regressão, árvores de decisão, entre outros.

2.4.4 Agrupamento (Clustering)

A análise de agrupamentos tem o objetivo de formar grupos ou elementos similares entre si. Um agrupamento, ou cluster, é uma coleção de registros similares entre si, porém diferente dos outros registros dos demais agrupamentos. A diferença entre agrupamento e classificação, é que na classificação as classes são pré-definidas pelo pesquisador, enquanto no agrupamento, não são.

Na análise de agrupamentos, os grupos ou classes são constituídos com base na semelhança entre os dados, cabendo ao minerador das classes resultantes, avaliar se estas terão alguma utilidade. A análise de agrupamentos é uma técnica normalmente preliminar, utilizada quando não se sabe nada, ou quase nada, sobre os dados. Agrupar ou segmentar um mercado, é um forma usual de análise de agrupamentos, onde consumidores são reunidos em classes representantes dos segmentos desse mercado.

Os algoritmos utilizados para agrupamentos, geralmente são algoritmos estatísticos específicos para esta finalidade, porém as redes neurais e os algoritmos genéticos são utilizados neste sentido. A figura 2 abaixo, mostra um agrupamento de três clusters.

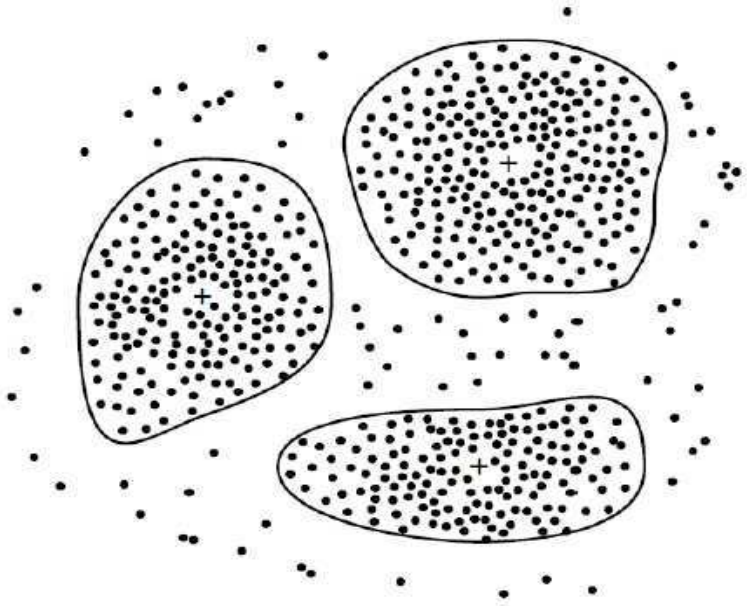

Figura 2 - Registros agrupados em três clusters.

2.4.5 Associação

É a descoberta de relações de associação ou correlações entre um conjunto de itens. Apresentam a forma: SE *atributo* X ENTÃO *atributo* Y. É uma das tarefas mais conhecidas devido a ótimos resultados obtidos, principalmente na análise da "Cesta de compras", do qual deseja-se conhecer quais os produtos que são comumente comprados em conjunto pelos consumidores. Isto possibilita a otimização do layout interno dos supermercados e a realização de vendas dirigidas nas quais os itens são oferecidos já em conjuntos com preços menores.

Os algoritmos utilizados para regras de associação constituem-se no procedimento mais utilizado nesses casos.

2.5 Tipos de algoritmos

Vários tipos de algoritmos são utilizados nas diferentes técnicas de mineração de dados, por serem os mais amplamente usados, três (3) dos principais tipos de algoritmos são brevemente analisados a seguir.

2.5.1 Árvores de Decisão

Uma árvore de decisão é um fluxograma, semelhante a estrutura de uma árvore, onde cada nó significa um atributo ou teste, cada ramo representa o resultado e cada folha representa a distribuição dos registros. O método de árvores de decisão representa um tipo de algoritmo de aprendizado de máquina que utiliza a aproximação dividir-para-conquistar. Com este método, permite ao usuário definir o objeto de saída, então a partir de um grupo de dados é possível identificar os fatores mais importantes correlacionados com este objeto.

Inicialmente todos os registros são associados ao nó da raiz da árvore de decisão, após isso o algoritmo seleciona uma partição dos dados e divide o conjunto de registros no nó da raiz de acordo com o valor do atributo selecionado. Isto tudo com o objetivo de separar as classes os registro de classes diferentes tendam a serem associadas a distintas partições. Esse processo é repetido inúmeras vezes, produzindo subconjuntos até que um critério de parada seja satisfeito. Para diminuir a quantidade de ramos da árvore usamos métodos de poda, ou seja, determinar quantas sub-árvores ou particionamentos serão necessários gerar.

O sucesso do uso de algoritmos de árvores de decisão, se da ao fato de ser uma técnica extremamente simples, que não necessita parâmetros de configuração, geralmente tem um bom grau de assertividade, tem uma eficiência computacional alta e facilidade de interpretação. São indicados aos casos em que se têm muitos atributos categóricos e que não se conhece a distribuição dos dados.

2.5.2 Redes Neurais Artificiais

Uma rede neural artificial é um processador maciçamente paralelo e distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torna-lo disponível para o uso (HAYKIN, 1999). Essas unidades de processamento simples (neurônios) têm o objetivo de calcular determinadas funções matemáticas, eles são dispostos em uma ou mais camadas e interligados por um grande número de conexões, essas conexões estão associadas a pesos que armazenam o conhecimento representado no modelo e ponderam as entradas recebidas por cada neurônio da rede.

Os fundamentos das Redes Neurais Artificiais (RNA's) são inspirados em sistemas neurais biológicos, com intenção de simular a forma como o cérebro humano aprende, recorda e processa as informações.

A forma como as RNA's adquirem conhecimento a partir de um ambiente é feita através de um processo de aprendizagem (treinamento). Na fase de treinamento, os pesos das conexões da rede vão sendo ajustados de forma que as informações extraídas dos dados possam ser representadas internamente, através de repetidas iterações, ajustando os parâmetros do modelo. Após muitas repetições, o modelo que se aproxima muito dos pontos dentro do grupo de dados pode ser internamente definida.

As vantagens de algoritmos baseados em redes neurais estão em sua alto tolerância a valores discrepantes e incorretos, além de sua robustez computacional ao lidar com erros no conjunto de treinamento, lida com a não-linearidade do modelo, também possui alta adaptabilidade e mapeamento de entradas e saídas. Em função dessas características, torna-se possível que uma rede neural se adapte a uma resposta previamente determinada, realize a classificação a padrões e tendências, se adapte a modificações com o decorrer do tempo, obtenha robustez computacional, podendo manter a questão *elasticidade vs plasticidade*, tudo em decorrência da metodologia a ser utilizada, por meio da melhor estrutura neural adequada a cada situação em específico (HAYKIN, 1999).

As suas dificuldades estão na necessidade de definição de muitos parâmetros como a sua estrutura e valores iniciais dos pesos, além de longos períodos de treinamento e da dificuldade em se identificar de forma clara as relações entre entradas e saídas.

Na figura 3 podemos observar uma imagem ilustrativa de uma rede neural perceptron de multicamadas.

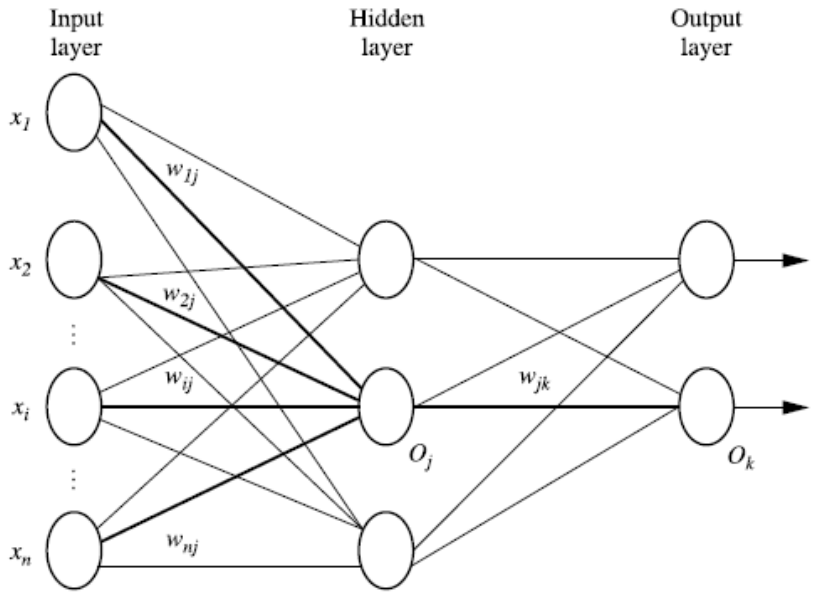

Figura 3 -Rede neural Perceptron Multicamadas.

2.5.3 Rede Neural Perceptron Multicamadas

A rede neural Perceptron Multicamadas (MLP) é uma importante classe de redes neurais e também uma das mais versáteis quanto a sua aplicabilidade, pois podem ser utilizados em diversos tipos de problemas, principalmente para reconhecimentos de padrões, controle e processamento de sinais. Uma das suas principais características é a capacidade universal de aproximar funções. Com uma camada intermediária de neurônios, seria suficiente para aproximar qualquer função contínua e com duas camadas, seriam suficientes para aproximar qualquer função matemática podendo ser contínua ou não, de acordo com Cybenko (1989).

A figura 4 mostra o modelo de um neurônio artificial e nela os sinais de entrada são ponderados por pesos sinápticos e somados no corpo celular do neurônio. A ativação do neurônio se dá por meio de uma função de ativação não-linear f(.), transformando o sinal de entrada em estado de ativação.

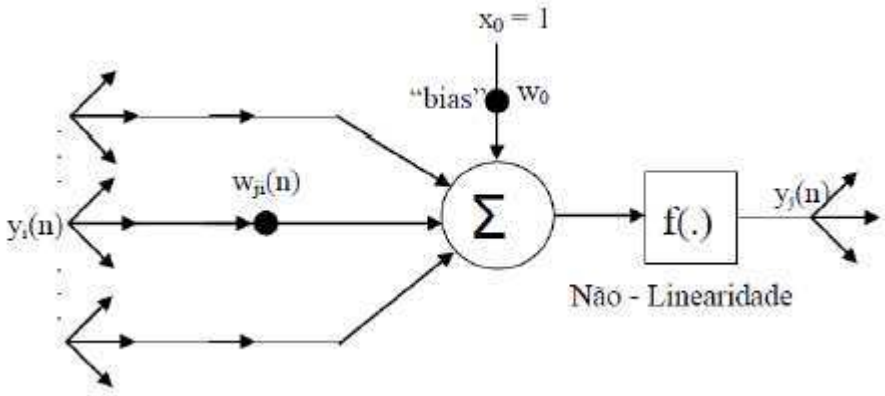

Figura 4 - Modelo de um neurônio artificial.

Cybenko (1989) mostra que em uma rede MLP, o número de camadas intermediárias é determinado pela natureza da função a ser aproximada. Esse número é definido empiricamente, pois depende da distribuição dos dados a serem utilizados para treinamento e validação, tornando-se necessário a análise prévia dos dados e do problema a ser enfrentado.

Para os autores Braga, Ludemir e Carvalho (2000), uma grande dificuldade surgiu com o treinamento de redes MLP, pelo fato de que poderia ocorrer convergência para um mínimo local, por causa da distribuição dos dados. Para tentar resolver esse problema, a rede MLP deveria ser implementada com uma camada intermediária formada por um conjunto de redes Perceptron para cada grupo de entradas linearmente

separáveis. Outra solução seria treinar toda a rede de uma única vez, porém, esse caso surgia um outro problema: Como treinar os neurônios da camada intermediária?

WERBOS (1974) apresenta um método para resolução do problema chamado de algoritmo backpropagation ou retropropagação do erro. O treinamento se da de forma supervisionada. Quanto a sua estrutura, as redes neurais MLP são compostas por camadas sucessivas, ou pelo menos uma camada intermediária de neurônios situada entre a camada de entrada e a camada de saída, assim essas redes tem no mínimo duas camadas (HAYKIN, 1999), conforme a figura 5.

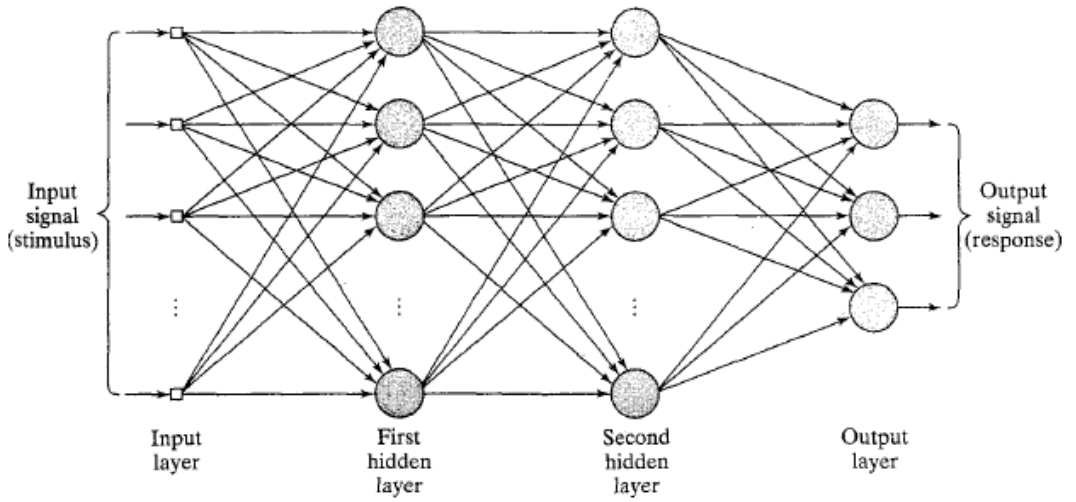

Figura 5 - Arquitetura de uma rede MLP (HAYKIN, 1999).

Os sinais de entrada da rede MLP, se propagam sequencialmente rumo à camada de saída, passando por todos os neurônios das camadas estruturais da rede, da esquerda para a direita, de acordo com a figura 5. As camadas intermediárias tem a função de extrair características, com pesos que são uma codificação das características dos sinais de entrada, permitindo então a rede criar uma representação particular, em um formato mais complexo e com mais informações. Enfim, os neurônios da camada de saída recebem os sinais vindos da última camada intermediária e produz uma resposta padrão que será a saída da rede neural.

O algoritmo backpropagation ou retropropagação do erro pode ser melhor entendido pelos seguintes passos:

- Apresenta-se um padrão X à rede neural, produzindo uma saída Y;
- É obtida a diferença entre o valor desejado e a saída, ou seja, é efetuado o cálculo de erro de cada saída;

- O erro é retropropagado pela rede, estando associado à derivada parcial do erro quadrático de cada elemento associado aos pesos;
- Os pesos de cada elemento são ajustados;
- Apresenta-se um padrão desconhecido à rede, repetindo o processo até sua convergência, ou seja, |erro| ≤tolerância arbitrada.

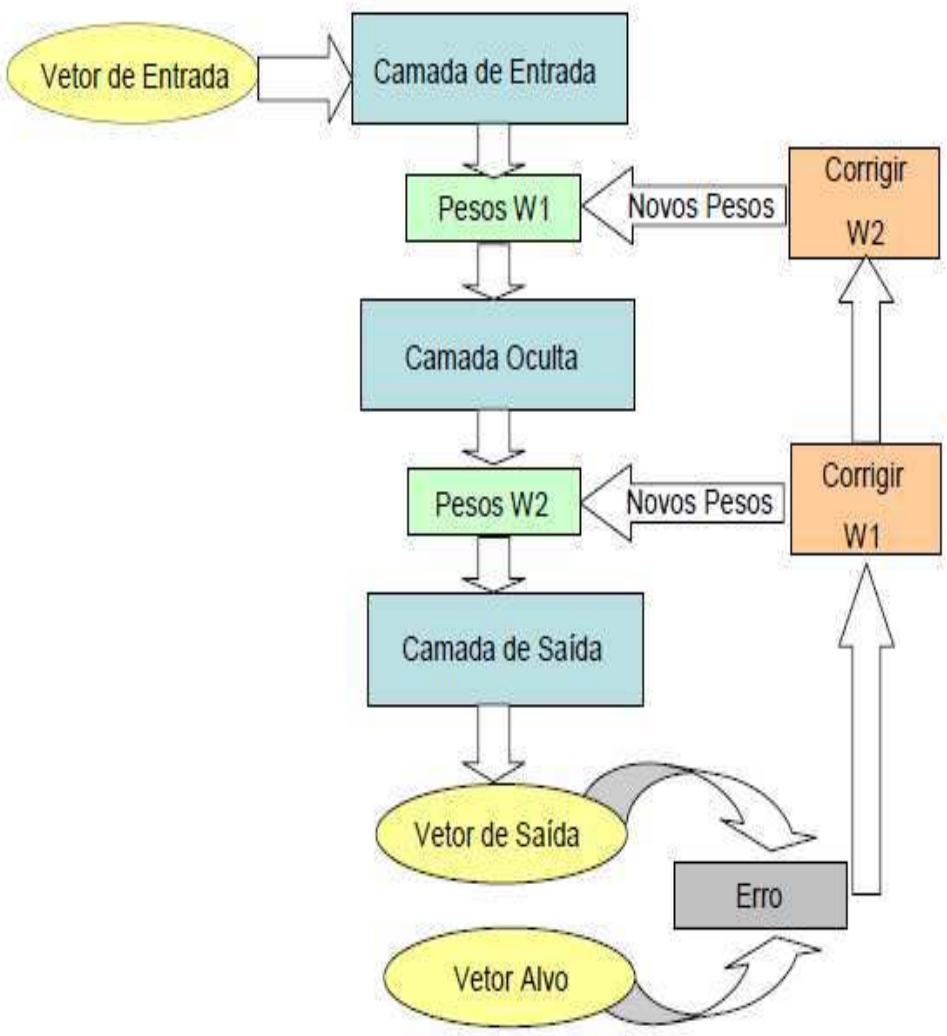

Figura 6 – Algoritmo de retropropagação do erro.

Basicamente, o processo de retropropagação do erro é constituído de duas fases: uma fase de propagação do sinal funcional (direta) e uma de retropropagação do erro (inversa). Na fase positiva, os vetores de dados são aplicados às unidades de entrada, e seu efeito se propaga pela rede, passando camada a camada. Então, um conjunto de saídas é produzido como resposta da rede. Durante a fase positiva, os pesos das conexões são mantidos fixos. Na retropropagação do erro, fase inversa, por outro lado,

os pesos são ajustados de acordo com uma regra de correção do erro. Especificamente, a resposta da rede em um instante de tempo é subtraída da saída desejada para produzir um sinal de erro. Este sinal de erro é propagado da saída para a entrada, passando camada a camada, originando o nome "retropropagação do erro". Os pesos são ajustados de forma que a "distância" entre a resposta da rede e a resposta desejada seja reduzida.

A figura 6 mostra o diagrama do algoritmo de retropropagação do erro, podemos visualizar de forma clara todas as suas fases. O algoritmo procura o mínimo da função erro no espaço dos pesos sinápticos usando o método de gradiente descendente. A combinação de pesos que minimiza a função de erro é considerada uma solução de o problema de aprendizagem. Como este método requer o cálculo do gradiente da função de erro em cada passo de iteração, temos de garantir a continuidade e diferenciabilidade da função erro.

2.6 Validação Cruzada (Cross-Validation)

A essência do aprendizado backpropagation é codificar uma relação funcional entre as entradas e saídas, representada por um conjunto de exemplos rotulados {x, d}, nos pesos sinápticos e limiares de um Perceptron de múltiplas camadas (MLP). A esperança é que a rede torne-se bem treinada de modo que ela aprenda bastante sobre o passado para pode generalizar sobre o futuro. Com essa visão, o processo de aprendizado se equivale a uma escolha de parametrização da rede para este conjunto de exemplos. Mais especificamente podemos ver o problema de seleção da rede como sendo de escolha, dentro de um conjunto de estruturas de modelos candidatas (parametrizações), a "melhor" de acordo com um tipo de critério.

Neste contexto, uma ferramenta padrão em estatística conhecida como validação cruzada (cross-validation) fornece um procedimento de grande valia. Validação cruzada é um processo de aprendizagem supervisionada em mineração de dados, após o pré-processamento e a formatação, os dados são fragmentados em dois subconjuntos, denominados base de treinamento e base de testes.

Na primeira fase, um algoritmo de indução de conhecimento é aplicado à base de treinamento. Com isso se obtém um modelo "treinado", que representa de certa forma as informações extraídas. O conjunto de treinamento é dividido em dois

subconjuntos, o conjunto de estimação, usado para selecionar o modelo, e o conjunto de validação, usado para testar ou validar o modelo. Na segunda fase, o modelo obtido é aplicado ao fragmento da base de dados chamado de base de testes. Como essa base de testes é previamente rotulada, se pode medir a taxa de acerto do modelo, comparando-se o resultado obtido com a rotulação disponível na base de testes.

A técnica de validação cruzada consiste em dividir a base de dados em "x" partes (folds). Destas, "x-1" partes são utilizadas para o treinamento e uma serve como base de testes. O processo é iterativo, de forma que cada parte seja usada uma vez como conjunto de testes. Portanto, ao final a correção total é calculada pela média dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas.

2.7 Modelo ARIMA

Uma série temporal é o conjunto de observações feito sequencialmente ao longo do tempo, onde a ordem dos dados é de crucial importância. Ela pode ser estacionária, quando ela se desenvolve aleatoriamente no tempo em torno de uma média constante, ou pode ser não-estacionária, quando ela varia em torno dessa média. A classe de modelos ARIMA (Auto Regressive Integrated Moving Average), proposto por Box & Jenkins, ou ARIMA (p, d, q), em que as letras "p" e "q" referem-se, respectivamente, ao número de parâmetros AR e MA existentes no modelo e "d" representa quantas diferenciações foram necessárias para estacionarizar a série. Caso ela seja estacionária, d=0. Os modelos ARIMA são capazes de descrever de maneira satisfatória séries estacionárias e não estacionárias, com exceção das séries que apresentam um comportamento explosivo, por exemplo, o crescimento de uma colônia de bactérias.

O principal foco de Box & Jenkins se baseia na previsão. Essa metodologia permite que valores futuros de uma série, no caso deste trabalho a previsão de cargas de energia, sejam previstos utilizando apenas seus valores presentes e passados, através da correlação temporal que existe entre os valores exibidos pela série.

Pelo fato de que a maioria dos processos encontrados são raramente estacionários, ou seja, não se desenvolve no tempo por meio de uma média constante, então para torna-las estacionárias, é necessária a aplicação de diferenças. O número de iteração para tornar uma série estacionária é denominado ordem de integração.

Caso $M_t = \Delta^d Z_t$ é estacionária, podemos representar M_t por um modelo ARIMA (p,q) ou seja,

$$\phi(B)M_t = \theta(B)\, a_t$$

Se M_t for uma diferença de Z_t, então Z_t é uma integral de M_t, daí dizermos que Z_t segue um modelo auto-regressivo, integrado, de médias móveis, ou ARIMA, denotado por ARIMA (p,q,d) e representado pela equação abaixo:

$$\phi(B)\Delta^d Z_t = \theta(B)a_t$$

onde d é o número de diferenças para tornar a série estacionária; Z_t são os valores atuais no período t; a_t são os erros aleatórios no período t;$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 -...- \phi_p B^p$ é o operador auto-regressivo de ordem p; $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 -...- \theta_p B^p$ é o operador das médias móveis de ordem *q*; Δ^d é a diferenciação da série.

Abaixo é citado alguns casos particulares ARIMA (MORETIN; TOLOI, 2006).

- ARIMA (0, 1, 1): $\Delta Z_t = (1-\theta B)a_t$;
- ARIMA(1, 1, 1): $(1 - \varphi B)\, \Delta Z_t = (1 - \theta B)\, a_t$;
- ARIMA (*p*, 0, 0) = AR(*p*);
- ARIMA (0, 0, *q*) = MA(*q*);
- ARIMA (*p*, 0, *q*) = ARMA (*p*, *q*);

Existe também uma variação do processo estocástico ARIMA capaz de captar a sazonalidade de uma série temporal, denominada SARIMA. Muitas séries temporais contêm uma componente periódica sazonal que se repete a cada *s* observações, por exemplo, para dados mensais *s* = 12. Para essas séries, o uso do modelo SARIMA (p,d,q)(P,D,Q) é o mais adequado. Esses modelos contém uma componente não sazonal com parâmetros (p,d,q) e outra sazonal (P,D,Q). Uma série que apresenta sozonalidade, na parte de modelagem, o objetivo é calcular esta componente sazonal e na sequência, subtrair a componente do modelo.

2.8 Metodologia de Box & Jenkins

Suponha que exista um sistema que atue como um filtro da Figura 7, que é estimulado por uma série de ruídos brancos, resultantes de um processo de geração de

números aleatórios, e que com esse estímulo seja gerada pelo sistema uma seqüência de valores observados seguindo um padrão, que corresponde à série temporal Yt.

Figura 7 - Geração de série temporal Yt.

Em situações da realidade, tem-se o caminho inverso ou seja, conhece-se o conjunto de observações seqüenciais Y_t geradas pelo sistema em questão, ao qual busca-se associar um modelo que corresponda aos processos internos ao sistema que as gerou (Figura 8). Uma vez que se estabeleça um modelo operacional para essa representação, a série aleatória e_t de valores em torno de zero corresponde à seqüência de valores (resíduos) que resulta ao extrair de Y_t os valores obtidos com o modelo ajustado à essa mesma série de valores observados Y_t.

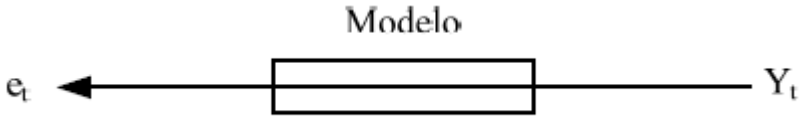

Figura 8 - Associação de modelo à série de observações Y_t.

Segundo a sistemática da metodologia de Box-Jenkins os modelos ARIMA descrevem tanto o comportamento estacionário como o não-estacionário. Dessa forma, pode-se afirmar que essa é uma metodologia de modelagem flexível em que as previsões com base nesses modelos são feitas a partir dos valores correntes e passados dessas séries.

A metodologia de Box & Jenkins para a previsão se baseia no ajuste de modelos tentativos denominados ARIMA à séries temporais de valores observados de forma que a diferença entre os valores gerados pelos modelos e os valores observados resulte em séries de resíduos de comportamento aleatório em torno de zero.

Pode-se associar o conceito inicial de um filtro estimulado por uma sériealeatória do tipo ruído branco à metodologia de Box & Jenkins, podemos ver isso através da figura abaixo.

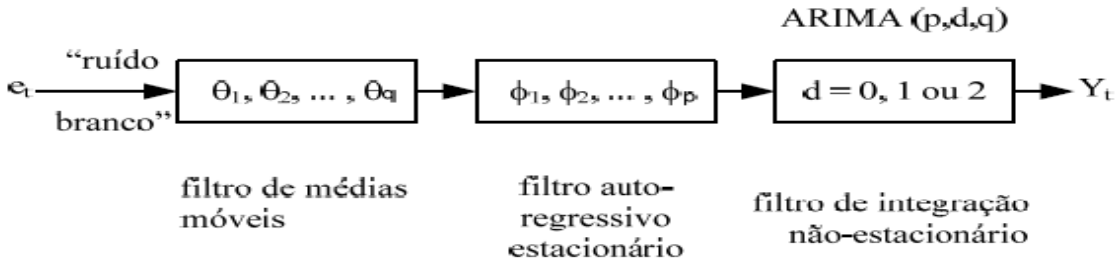

Figura 9 - Os filtros de médias móveis, autoregressivos e de integração não-estacionária.

Na Figura 9 é representado um conjunto de sucessivos filtros aos quais associam-se os parâmetros dos modelos ARIMA (p,d,q) que representam os sistemas estimulados pela série e_t que geraram a série temporal Y_t: o filtro de médias móveis (parâmetro q), o filtro autoregressivo estacionário (parâmetro p) e o filtro de integração não-estacionário (parâmetro d).

A metodologia de Box-Jenkins corresponde a três estágios principais: (1) identificação, (2) estimação, e (3) Verificação, aos quais se segue a aplicação do modelo para a previsão ou controle do sistema de geração dos valores observados Y_t.

1. **Identificação**. Consiste em descobrir qual dentre as várias versões do modelo ARIMA descreve o comportamento da série. Nesta etapa calcula-se autocorrelações, autocorrelações inversas, autocorrelações parciais e correlações cruzadas.

O procedimendo desta fase de identificação consiste em tomar diferenças da série quantas vezes necessárias para obter-se uma série estacionária, de modo que o processo $\Delta^d Z_t$ seja reduzido a um ARMA (p, q). O número de diferenças, d, necessárias para que o processo seja estacionário, é alcançado quando a função de autocorrelação amostral de $W_t = \Delta^d Z_t$ decresce rapidamente para zero. Outro procedimento é identificar o processo resultante ARMA (p, q), analisando as autocorrelações e autocorrelações parciais estimadas.

Portanto, a realização do processo de identificação necessita de outros procedimentos, tais como a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP).

Uma função de autocorrelação (FAC) é a correlação entre o comportamento anterior da carga(para este caso) e o fator que se deseja influenciar no valor futuro da

carga. O coeficiente de autocorrelação de ordem j, ou seja, a autocorrelação entre Z_t e Z_{t-j} é obtido pela equação abaixo:

$$\rho = \frac{Cov\,(Zt, Zt-j)}{V(Zt)} = \frac{Yj}{Yo}$$

Nesta, a sequência de pares (j, ρ_j), j = 1 , 2, ..., é denominada função de autocorrelação.

O coeficiente de autocorrelação ρ_j envolve parâmetros desconhecidos, assim na prática é necessário trabalhar com coeficiente de autocorrelação "amostral" r_j, definido pela equação abaixo:

$$r_j = \frac{\sum_{t=j+1}^{N}(Zt-\overline{Z})(Zt-j-\overline{Z})}{\sum_{t=1}^{N}(Zt-\overline{Z})^2}, \; j = 1,2, \ldots$$

Onde ρ_j é a autocorrelação teórica, Y_j é a covariância de Z_t, Y_o é a variância de Z_t, r_j é a autocorrelação amostral e N é o número de observações da série Z_t.

As autocorrelações amostrais são apenas estimativas de autocorrelações teóricas ρ_j, portanto tendem ao mesmo padrão, assim pode-se concluir muitas propriedades do processo estocástico subjacente a partir de um estudo da função de autocorrelação amostral (O'DONOVAN, 1983). Mesmo sendo de grande importância, nem sempre a função de autocorrelação amostral permite especificar o modelo apropriado, dessa forma é preciso outra característica da série temporal sendo esta a função de autocorrelação parcial amostral expressa.

A função de autocorrelação parcial amostral (FACP) é também por sua vez, uma estimativa da autocorrelação parcial teórica ρ_{jj}, calculada dos valores da série temporal observada. A autocorrelação parcial teórica é definida como sendo a autocorrelação entre quaisquer duas variáveis Z_t e $Z_{t\,+\,j}$, separadas *j* atrasos de tempo sendo que as variáveis Z_{t+1}, Z_{t+2}, ..., Z_{t+j-1} são eliminadas. As autocorrelações parciais amostrais possuem a seguinte forma:

$$r_{11} = r_1$$

$$r_{22} = \frac{r2 - r1^2}{1 - r1^2}$$

A partir de r_{33} tornam-se cada vez mais complicadas.

Exatamente como a forma ocorrida com a FAC amostral, a FACP amostral tende ao mesmo padrão que a FACP teórica, de maneira que se pode utilizar a FACP amostral para identificar o modelo apropriado no processo estocástico.

A identificação é a fase mais crítica da metodologia de Box & Jenkins, vários pesquisadores usando a mesma série podem identificar modelos diferentes. Isto porque, trabalhando com a FAC e a FACP amostrais, muitas vezes fica difícil decidir se elas estão decrescendo ou se são truncadas. Muitos pesquisadores preferem utilizar outros procedimentos de identificação, que não depende de quem está analisando a série de tempo. Esse procedimento faz uso de critérios de seleção de modelo construídos com base na variância estimada de a_t , no tamanho da amostra e nos valores *p* e *q*.

Como nessa fase é possível que vários modelos diferentes se adaptem bem a uma determinada série temporal, então, deve-se utilizar o modelo mais simples, com menos parâmetros.

2. **Estimação**. Consiste em estimar os parâmetros do modelo identificado.Também produz diagnósticos de estatística para ajudar a avaliar a adequação do modelo.

Nesta fase são encontrados os valores dos coeficientes p e q. Os métodos que podem ser utilizados é o de mínimos quadrados e de máxima verossimilhança. Qualquer que seja o método adotado irá necessitar do uso de um computador, pois o processo de estimação é extremamente trabalhoso.

3. **Verificação**. Tem por finalidade avaliar se o processo de estimação foi bem sucedido.

Através de uma série de testes, sendo o principal a análise dos resíduos (erros de predição), ajusta-se o modelo. Se o modelo não for satisfatório, o ciclo é repetido, voltando-se a fase de identificação.

Nesta etapa, verifica-se se o modelo representa, ou não, de forma adequada os dados. Caso represente, é possível fazer a previsão, se não, outra especificação deverá ser escolhida para modelar a série, portanto, retorna a etapa inicial de identificação e estimação. Existem vários testes de verificação para um dado modelo ajustado a uma série. Em geral, tais testes são baseados nas autocorrelações estimadas dos resíduos.

No teste de autocorrelação residual, se o modelo for adequado os erros devem ser não-correlacionados, os resíduos do modelo estimado $\hat{a}_t$ deverão estar próximos do ruído branco a_t, sendo $\hat{a}_t$, uma aproximação de a_t. Se $\hat{r}_j$ indicarem as autocorrelações dos resíduos

$\hat{a}_t$, assim seus coeficientes de autocorrelação devem ser estatisticamente iguais a zero. As autocorrelações $\hat{r}_j$ são calculadas pela equação abaixo:

$$\hat{r}_j = \frac{\sum_{t=j+1}^{N} \hat{a}t\hat{a}t-j}{\sum_{t=1}^{N} \hat{a}t^2}$$

Contudo, conforme ressaltou Durbin (1970) para valores pequenos de *j*, a variância de $\hat{r}_j$ pode ser bem menor do que $\frac{1}{\sqrt{N}}$. Para valores moderados ou grandes de *j*, a distribuição é válida e podem-se realizar testes de hipóteses e construir intervalos de confiança para avaliar a significância de cada $\hat{r}_j$. Para o teste conjunto utiliza-se a estatística de Box e Pierce (1970), que apesar de não detectar quebras especificas no comportamento do ruído branco, pode indicar se esses valores são muito altos.

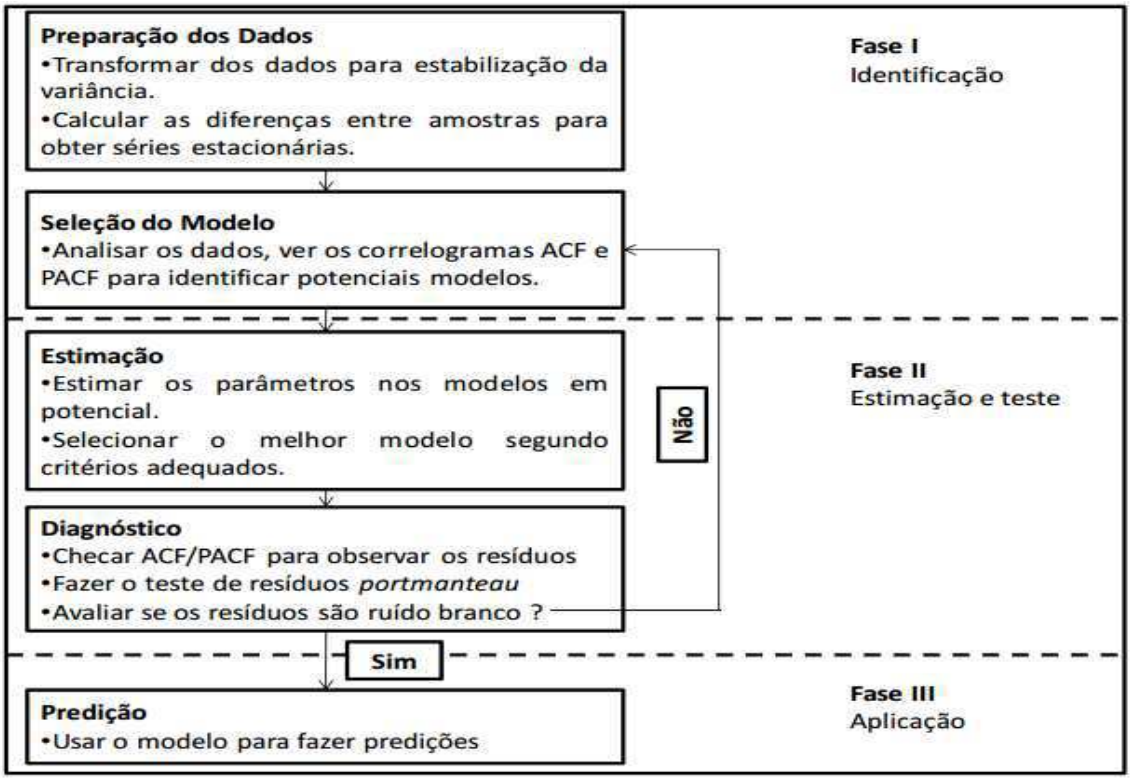

Figura 10 - Fluxograma do ciclo iterativo de Box & Jenkins.

Na fase da verificação, se a saída não for algo como um ruído branco, volta-se para a primeira fase, se for, pode-se usar o modelo para fazer a previsão.

2.9 Métodos Híbridos

Os métodos híbridos representam uma técnica já bastante difundida na literatura especializada e que demonstra a viabilidade de sua utilização, visando especialmente extrair as melhores características de modelos distintos, em favor da obtenção dos melhores resultados.

No presente trabalho, optou-se pelo modelo ARIMA de Box & Jenkins, combinando com a Rede Neural Perceptron Multicamadas, via algoritmo backpropagation, responsável pela previsão de cargas, valendo citar como suas principais características a facilidade de solução de problemas complexos e trabalhar bem com a não-linearidade.

Para fins de previsão de cargas elétricas, os modelos híbridos têm sido bastante difundidos, com resultados satisfatórios em relação a outros já descritos na literatura especializada. O presente trabalho demonstra de forma clara a viabilidade de se combinar métodos distintos, no intuito de extrair as melhores características de cada modelo, visando a precisa previsão de cargas de curto prazo.

2.10 Medidas de precisão da previsão

As medidas de precisão são uma aplicação de extrema importância no estudo das técnicas de previsão. Dado a complexidade das séries temporais reais, os valores futuros das variáveis tornam-se bastante difíceis de prever, deste modo, é fundamental incluir informação acerca da medida em que a previsão pode desviar-se do valor real da variável. Este conhecimento adicional fornece uma percepção melhorada sobre o quão precisa pode ser a previsão (STEVENSON, 1996).

O minerador de dados necessita de uma medida de precisão para usar como base de comparação ao escolher uma entre as várias técnicas disponíveis, com o objetivo de fazer a escolha mais acertada entre as técnicas, devido ao fato de que algumas técnicas oferecem uma maior precisão que outras.

Enquanto algumas aplicações de previsões envolvem uma série de previsões, por exemplo, as receitas mensais de uma empresa, outras envolvem uma única previsão que conduz a uma única decisão, como por exemplo, o tamanho de um shopping center. É importante monitorizar os erros de previsão para determinar se estão dentro de limites

razoáveis, quando são efetuadas previsões periódicas. Devem ser implementadas medidas corretivas no caso de os erros de previsão não se encontrarem dentro desses limites.

No presente trabalho o conjunto de observações são muito altos, logo, se o erro encontrado estiver baixo em relação ao número alto de observações, podemos dizer que o resultado é extremamente satisfatório.

A diferença entre o valor real e a previsão do valor da origem ao erro de previsão:

$$\mathbf{E_t = A_t - P_t}$$

Onde,

Et = Erro no período t

At = Valor real no período t

Pt = Previsão para o período t

Quando a previsão é muito baixa, ou seja, menor que o valor atual, o resultado são erros positivos. Os erros negativos ocorrem quando a previsão tem um valor mais elevado do que o valor atual.

As decisões podem ser influenciadas de duas formas distintas pelos erros de previsão. A primeira consiste na escolha de outros meios/alternativas de previsão. A outra consiste na avaliação do sucesso ou fracasso da técnica utilizada.

A seguir são apresentados os vários erros de previsão, as suas definições e fórmulas.

2.10.1 Erro Médio

Consoante o autor Dilworth (1992), o valor da previsão raramente é igual ao valor real devido às variações aleatórias que caracterizam a variável que, contudo, não deve diferir muito da média dos valores reais ao longo desses mesmos períodos. Desse modo, a previsão do modelo não deve ser tendenciosa, ou seja, a variável não deve ser subestimada ou sobrestimada. O erro médio deve ser muito próximo de zero caso o modelo de previsão seja isento, sendo calculado pela soma dos erros de previsão de uma

série de períodos e dividindo essa soma pelo número de erros usados para calcular a soma. O erro médio pode ser calculado através da seguinte equação:

$$\text{EM} = \frac{\sum_{t=1}^{n} et}{n}$$

Onde *n* é o número de períodos usados e o numerador da equação é chamado de soma corrente dos erros de previsão. Caso a soma de erros positivos seja igual a soma dos erros negativos, o modelo de previsão é imparcial, ou seja, a soma é próxima de zero.

2.10.2 Desvio médio absoluto

Uma das medidas mais comuns de erro de previsão é o desvio médio absoluto (DMA), que não leva em conta se um erro foi subestimado ou sobrestimado, caracterizando-se por ser a média dos erros cometidos pelo modelo de previsão durante uma série de períodos de tempo. Outra forma de se chamar o DMA é o de erro médio absoluto (EMA). Para calcular o DMA, subtrai-se o valor da previsão ao valor real em cada período de tempo, tendo em conta que o resultado deverá ser positivo, ou seja, sempre em módulo, soma-se e divide-se pelo número de valores que foram usados para obter a soma.

$$\text{DMA} = \frac{\sum_{t=1}^{n} |et|}{n}$$

Onde se percebe que devido ao módulo, ignora-se a direção dos desvios.

2.10.3 Erro quadrático médio

O erro quadrático médio (EQM), também pode ser usado como uma medida do erro de uma previsão. Ele pode ser determinado somando os erros de previsão ao

quadrado e dividindo pelo número de erros usados no cálculo. O erro quadrático médio pode ser calculado pela seguinte equação:

$$\text{EQM} = \frac{\sum_{t=1}^{n} e^2 t}{n}$$

2.10.4 Erro Percentual

Segundo Machado (2009), o erro percentual mede a percentagem do erro em relação ao valor real. Podemos calcular subtraindo ao valor real no período t a previsão no respectivo período e divide-se o resultado pelo valor real utilizado anteriormente.

$$\text{EP}_t = \frac{(At - Pt)}{At} \text{x } 100$$

2.10.5 Erro médio percentual

Podemos calcular dividindo-se o erro percentual pelo número de períodos. Se os erros positivos forem compensados pelos erros negativos, o resultado deve ser aproximadamente nulo.

$$\text{EMP} = \frac{\sum_{t=1}^{n} EPt}{n}$$

2.10.6 Erro médio percentual absoluto

De acordo com Heizer (2004), tanto para valores do desvio médio absoluto, como do erro quadrático médio dependem da importância do modelo que está sendo previsto, o que pode causar problemas ao nível da dimensão dos resultados. Se a previsão do modelo é medida em milhares, os valores do desvio médio absoluto e do erro quadrático médio podem ser muito grandes. A utilização do erro médio percentual absoluto (EMPA) é uma forma eficaz para resolver este problema. O EMPA é a média

da diferença absoluta entre os valores previstos e atuais, expressa em porcentagem dos valores atuais. Assim, se existem previsões e valores reais para ***n*** períodos, o erro médio percentual absoluto é:

$$\text{EMPA} = \frac{\sum_{t=1}^{n} |EPt|}{n}$$

2.10.7 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson é uma medida do grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor 1 indica uma relação linear perfeita e o valor -1 também indica uma relação linear perfeita mas inversa, ou seja quando uma das variáveis aumenta a outra diminui.

O coeficiente de correlação de Pearson é normalmente representado pela letra ***r*** e a sua fórmula de cálculo é:

$$r = \frac{n . \sum Xi.Yi - (\sum Xi).(\sum Yi)}{\sqrt{[\, n.\sum Xi^2 - (\sum Xi)^2] \,.[n.\sum Yi^2 - (\sum Yi)^2]}}$$

Onde n é o número de observações.

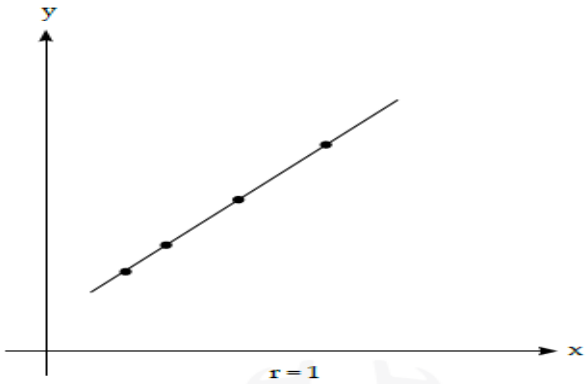

Figura 11 – Correlação linear perfeita (positiva).

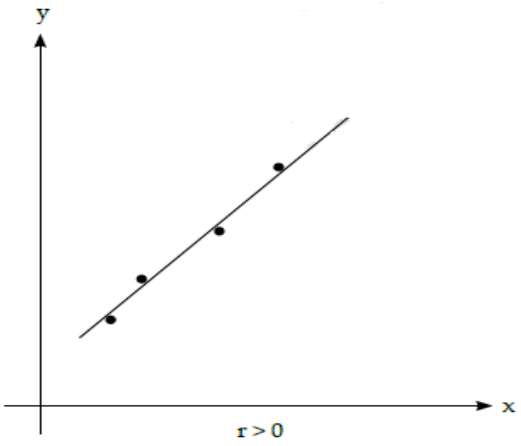

Figura 12 – Forte correlação positiva.

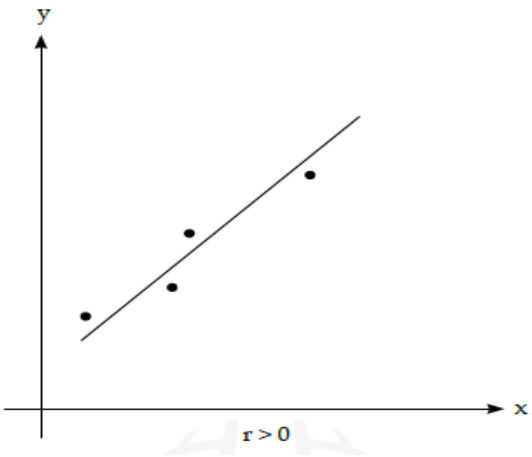

Figura 13 – Fraca correlação positiva.

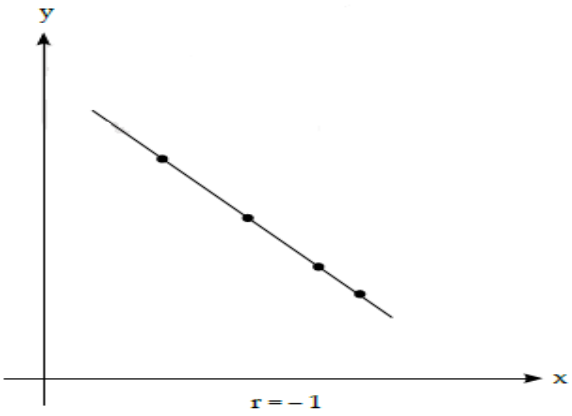

Figura 14 – Correlação linear perfeita (negativa).

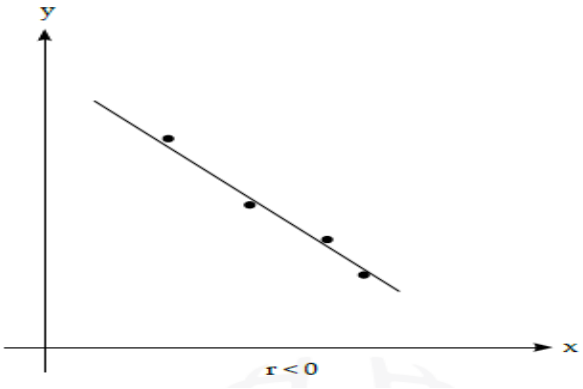

Figura 15 – Forte correlação negativa.

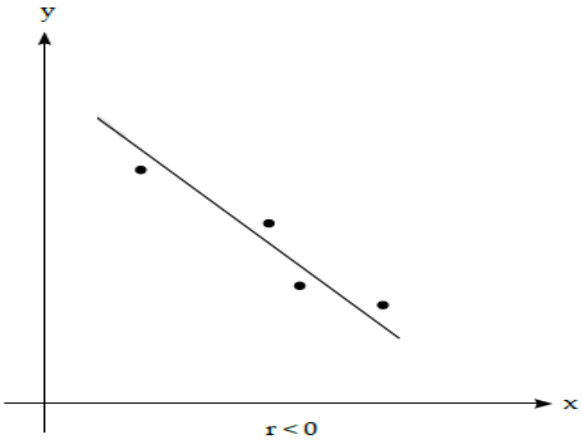

Figura 16 – Fraca correlação negativa.

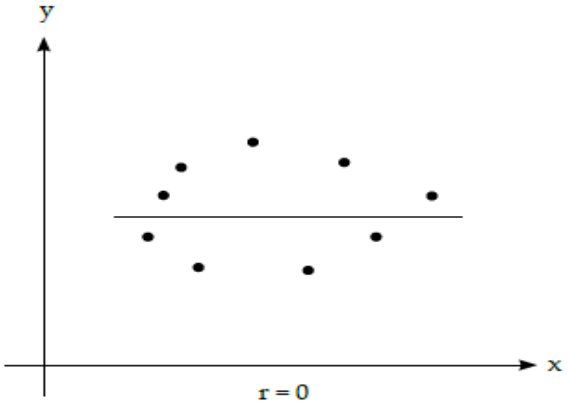

Figura 17 – Ausência de correlação linear.

Entretanto, quando r = 0, isso não significa que entre x e y não existe qualquer relação, mas que não existe entre essas variáveis uma relação linear. O coeficiente de

correlação r, portanto, mede a intensidade da relação linear entre as variáveis x e y, o que não implica que uma delas tenha efeito direto ou indireto sobre a outra variável. Pode acontecer de x e y estarem sendo influenciadas por outra(s) variável(eis) e, em consequência, ser estabelecido entre elas uma relação matemática.

3 APLICAÇÕES E RESULTADOS

A série histórica de valores de carga de energia foi obtida junto ao Operador Nacional do Sistema – ONS a partir de dados coletados do Informativo Preliminar Diário da Operação - IPDO, no período de 4/01/2003 à 29/12/2009 para a região nordeste, 4/08/2001 à 31/12/2009 para a região sul, sudeste, centro-oeste e norte. A série está registrada em base semanal e na unidade de medida Mega Watt Médio (MWMed).

A figura 18 apresenta um diagrama de blocos do modelo proposto. No primeiro passo para a modelagem híbrida, usa-se um modelo ARIMA sazonal, com os valores atrasados no tempo como entrada para modelar a parte linear da série temporal e para criação da previsão até 5 passos a frente. Calcula-se os valores reais menos os valores previstos que será o *resíduo* do ARIMA, esse resíduo será a entrada da RNMC. Não se espera que nesta etapa sejam captadas as não-linearidades da série temporal.

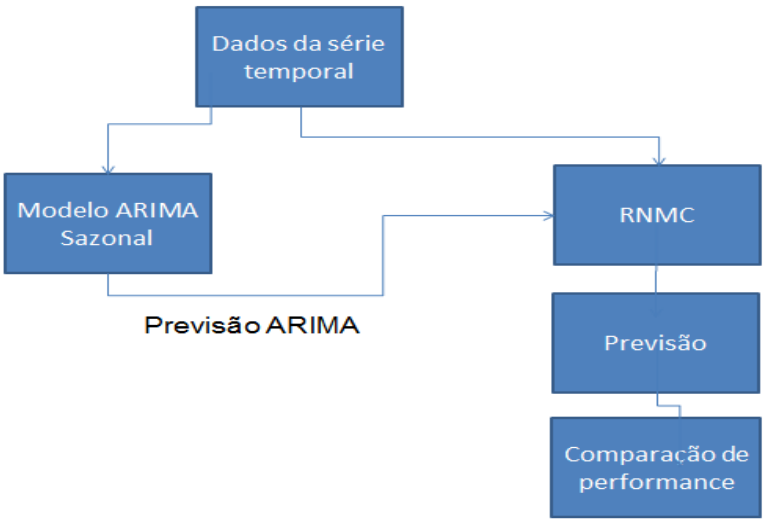

Figura 18 - Diagrama de blocos do modelo híbrido.

Para este fim, no segundo passo será usada uma Rede Neural de Múltiplas Camadas - RNMC. Antes do resíduo da previsão do ARIMA entrar na RNMC, por estratégia e para melhor obtenção de resutados, foi utilizada uma técnica de pré-processamento de dados, chamada de Principal Components Analysis (PCA), que através desta técnica, ajudadará o modelo a reduzir a complexidade dos seus dados, onde o PCA é um procedimento matemático que utiliza uma transformação ortogonal para converter um conjunto de observações de variáveis possivelmente correlacionadas a um conjunto de valores de variáveis linearmente descorrelacionadas chamadas componentes principais. O objetivo é encontrar um pequeno número de campos derivados que resumem as informações do conjunto original de campos. Um ponto forte desse método é que o PCA pode efetivamente reduzir a complexidade de seus dados sem sacrificar grande parte do conteúdo da informação.

Então utilizando a ferramenta de modelagem Factor / PCA do IBM SPSS Modeler, foram criados a partir deste pré-processamento, 5 fatores onde cada fator será uma entrada na RNMC e resume as informações principais do conjunto original de campos.

A RNMC criada possui 5 neurônios na camada de entrada, que são os 5 fatores gerados pelo PCA, também possui 10 neurônios na primeira camada intermediária, 5 neurônios na segunda camada intermediária, e o alvo que será a carga, na camada de saída. O software oferece uma opção antes da geração da RNMC, de criar um modelo default, onde o próprio software faz a escolha da melhor quantidade de neurônios, porém, utilizando-se desta opção, não foram encontrados os melhores resultados. Portanto, os neurônios das camadas intermediárias foram escolhidos empíricamente, onde a melhor combinação encontrada foi de 10 para a primeira e 5 para a segunda, pelo fato de que ao aumentar ou diminuir além desta combinação, os resultados não seriam os melhores possíveis. A RNMC gerada pode ser observada pela figura 19 abaixo.

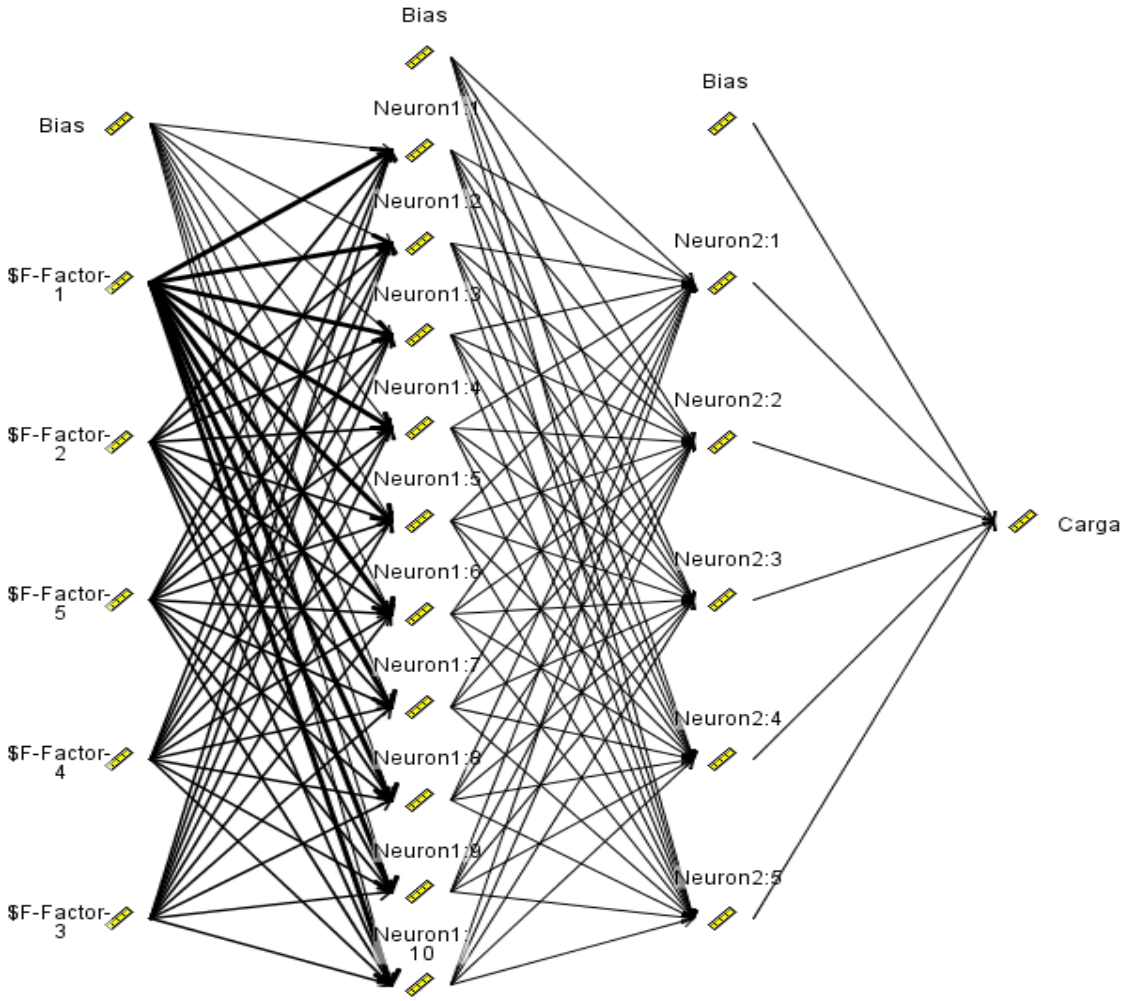

Figura 19 – RNMC gerada.

A RNMC capta as relações que não foram absorvidas pela etapa anterior usando os valores residuais das estimativas como entrada. Na terceira e última etapa, outra RNMC é usada para estimar valores residuais a serem acrescidos ao forecast realizado no primeiro passo. Produziu-se neste trabalho estimativas futuras até 5 passos a frente.

Primeiramente iremos observar o que foi gerado pela modelagem ARIMA, observando o ajustes de curvas, para 5 passos a frente das 5 regiões brasileiras, onde a linha tracejada em azul (Carga) é a série original e a linha contínua em vermelho ($ N-Carga) é o ajuste de curvas ARIMA.

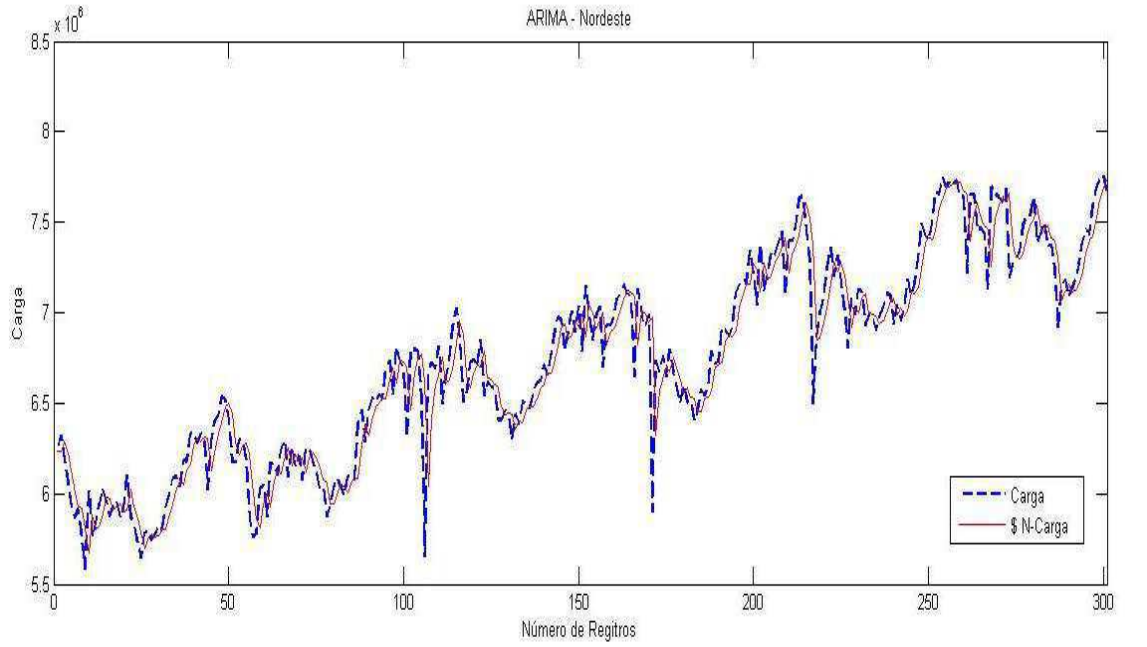

Figura 20 – Previsão Modelo ARIMA para 5 passos a frente – Região Nordeste.

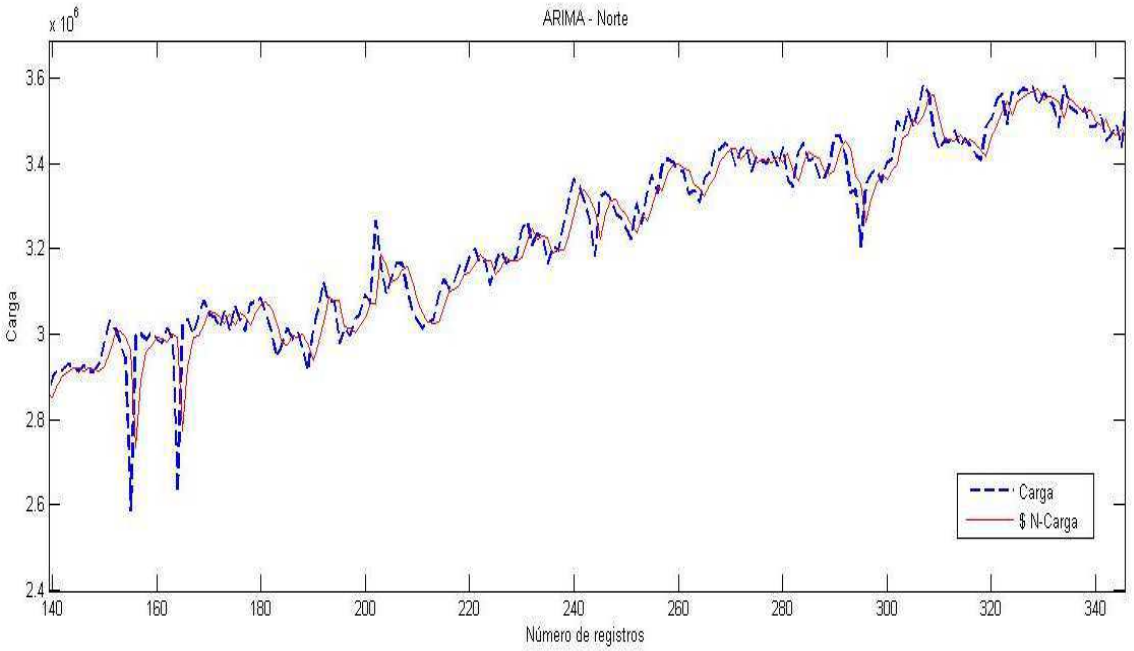

Figura 21 – Previsão Modelo ARIMA para 5 passos a frente – Região Norte.

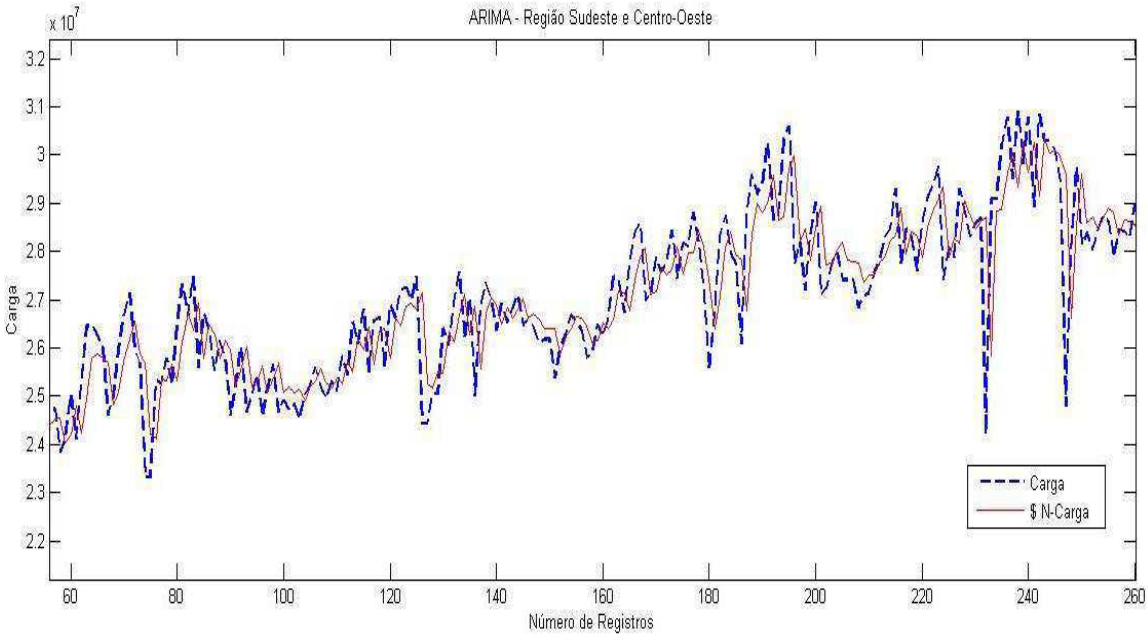

Figura 22 – Previsão Modelo ARIMA para 5 passos a frente – Região Sudeste e Centro-Oeste.

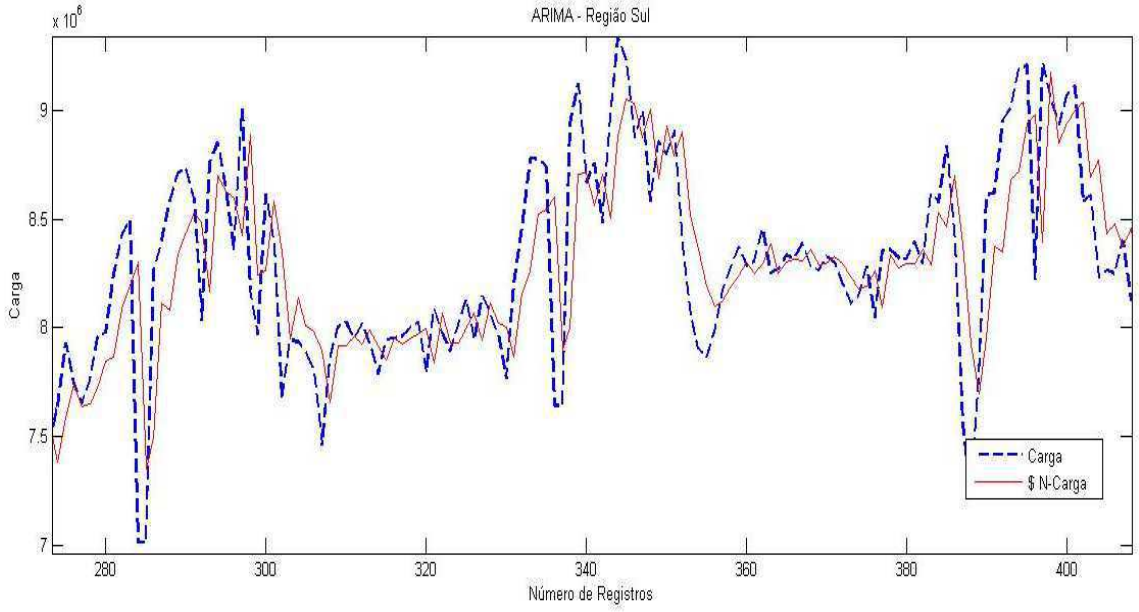

Figura 23 – Previsão Modelo ARIMA para 5 passos a frente – Região Sul.

Agora, as figuras a seguir ilustram as previsões para 5 passos a frente para o modelo Híbrido ARIMA – RNMC.

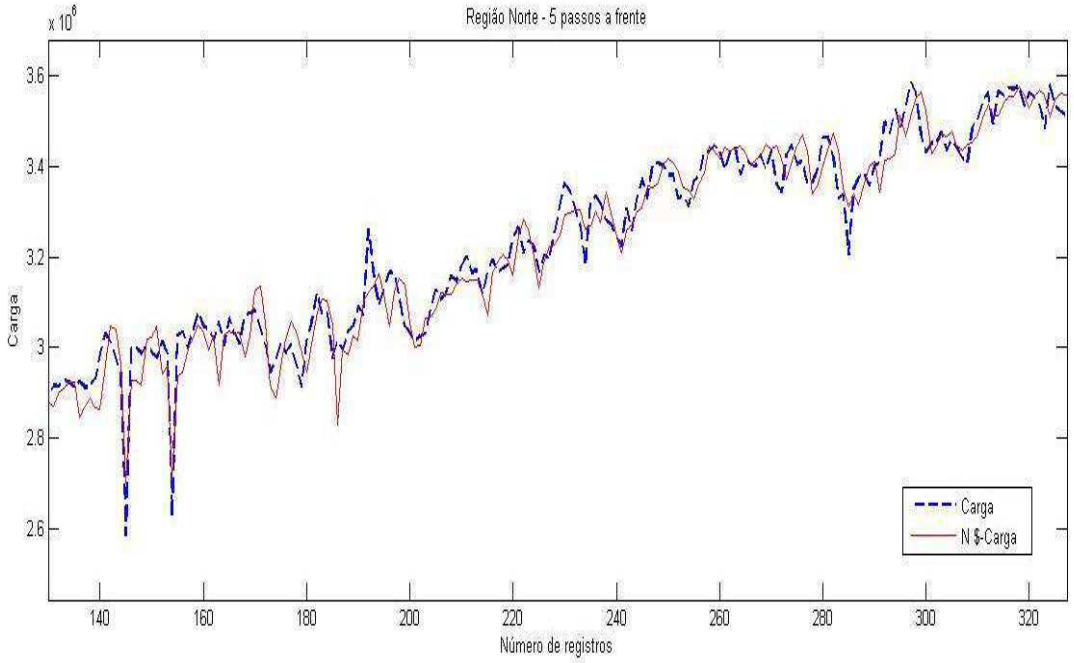

Figura 24 – Previsão para 5 passos a frente – Região Norte.

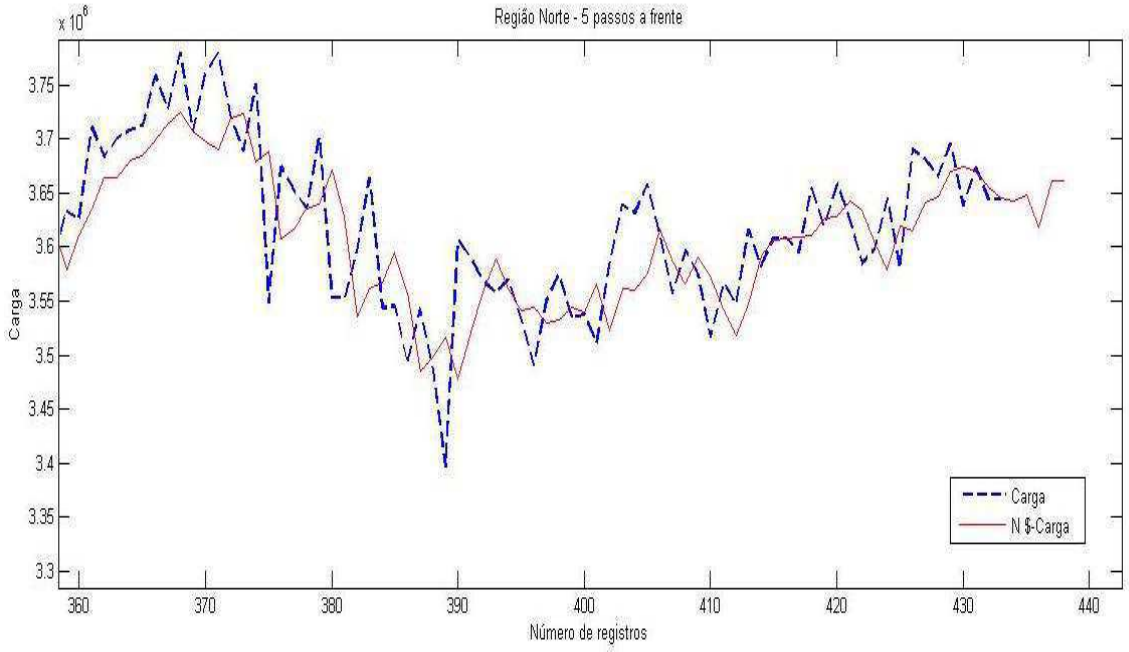

Figura 25 – Visão aproximada da Previsão para 5 passos – Região Norte.

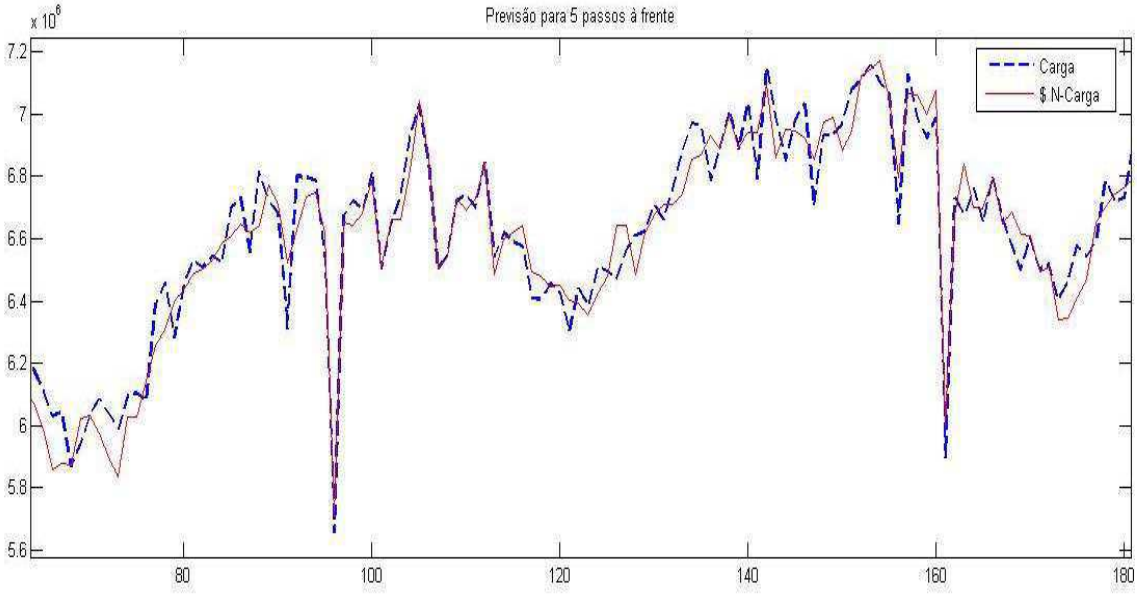

Figura 26 – Previsão para 5 passos a frente – Região Nordeste.

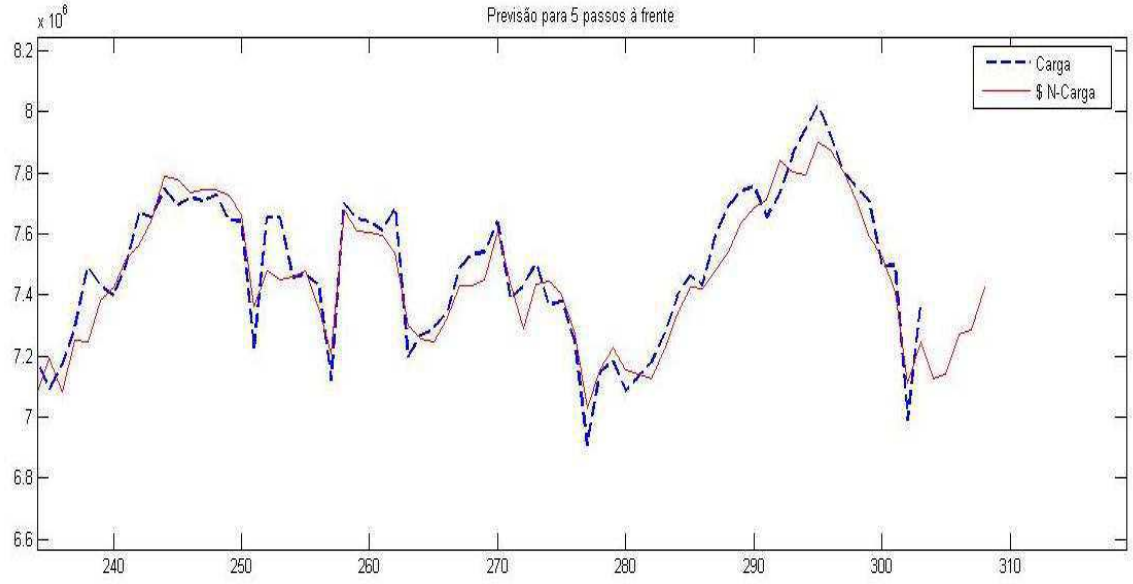

Figura 27 – Visão aproximada da Previsão para 5 passos – Região Nordeste.

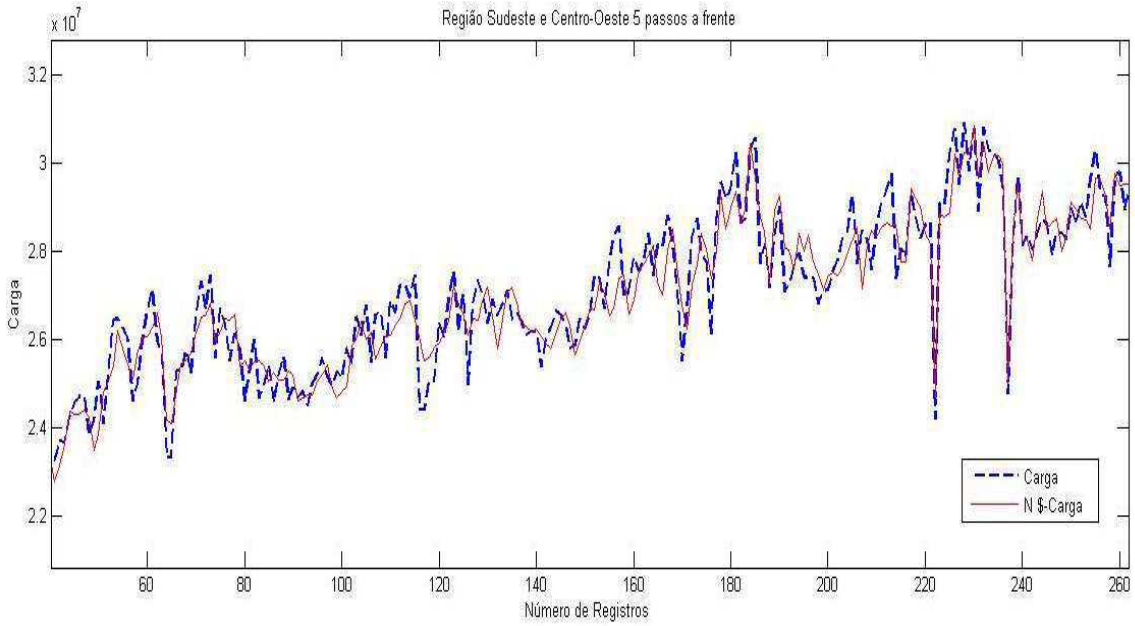

Figura 28 – Previsão para 5 passos a frente – Região Sudeste e Centro-Oeste.

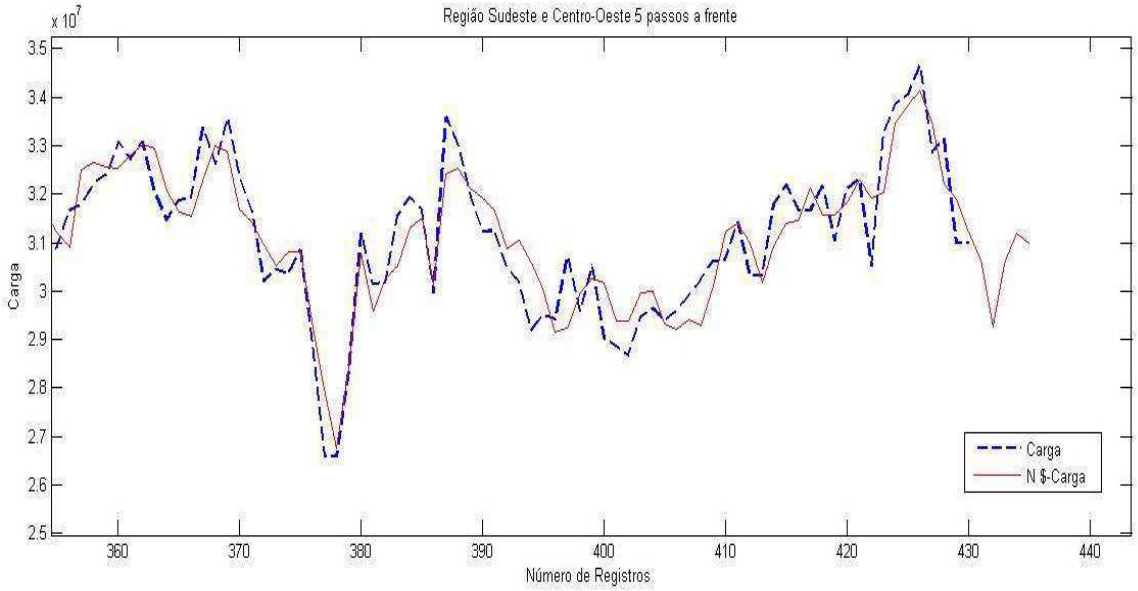

Figura 29 – Visão aproximada da Previsão para 5 passos – Região Sudeste e Centro-Oeste.

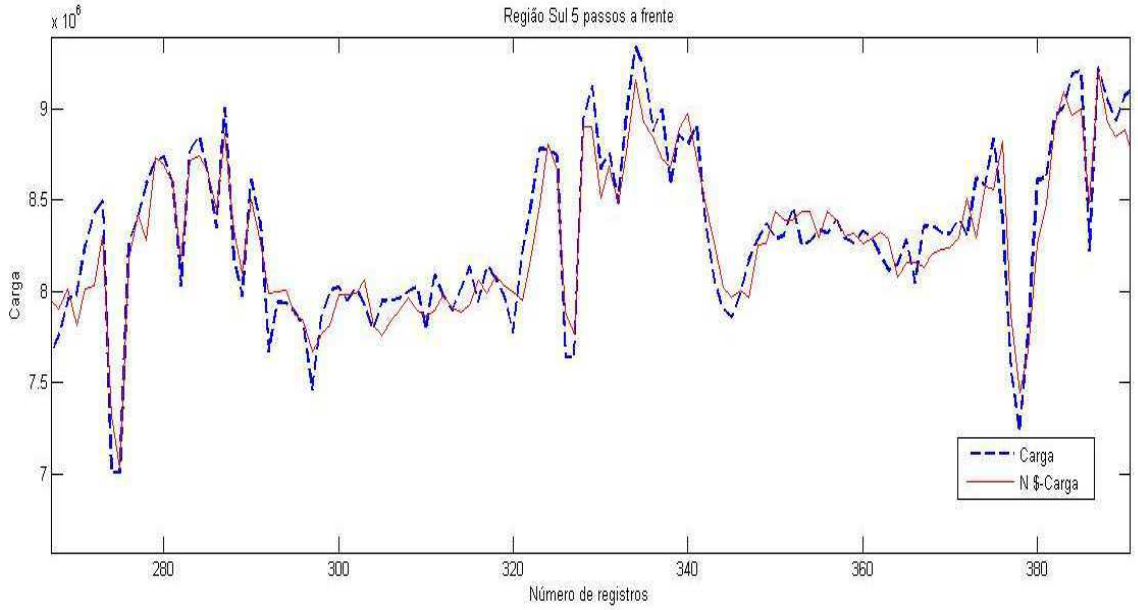

Figura 30 – Previsão para 5 passos a frente – Região Sul.

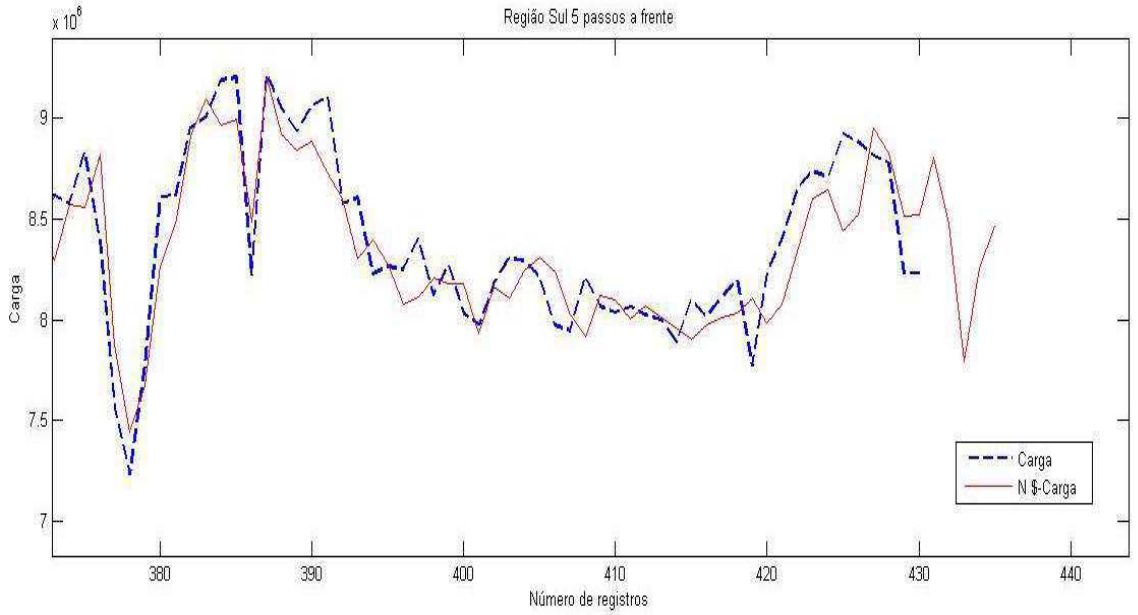

Figura 31 – Visão aproximada daPrevisão para 5 passos – Região Sul.

As figuras 20 a 31 apresentam a comparação entre a série temporal alvo (linha tracejada) e a série obtida pelo modelo híbrido (linha contínua) para até 5 passos a frente, em modo zoom (lupa de aproximação), para cada região. As figuras 25, 27, 29 e 31 especificamente, apresentam no detalhe os resultados dos valores do sistema preditor no momento onde a série alvo tem a interrupção de seus dados.

Tabela 1 - Índice de correlação linear e Erro Médio. – NORDESTE

	Passos a Frente	Real	Previsto	Erro (%)	Correlação Linear
NORDESTE	1	7366,57	7191,20	2,38	0,987
	2	7430,00	7326,67	1,39	
	3	7402,71	7439,98	0,50	
	4	7380,57	7559,19	2,42	
	5	7437,71	7484,73	0,63	

Tabela 2 - Índice de correlação linear e Erro absoluto Médio.- NORTE

	Passos a Frente	Real	Previsto	Erro (%)	Correlação Linear
NORTE	1	3695,43	3624,25	1,92	0,994
	2	3638,14	3652,57	0,39	
	3	3673,43	3675,94	0,06	
	4	3644,43	3686,29	1,14	
	5	3644,43	3668,51	0,66	

Tabela 3 - Índice de correlação linear e Erro Médio.- SUDESTE e CENTRO-OESTE

	Passos a Frente	Real	Previsto	Erro (%)	Correlação Linear
SUDESTE	1	31668,00	31254,20	1,30	0,980
	2	31857,86	31077,39	2,44	
	3	31139,86	31483,90	1,10	
	4	31982,43	32127,05	0,45	
	5	31962,43	32461,75	1,56	

Tabela 4 - Índice de correlação linear e Erro Médio.- SUL

	Passos a Frente	Real	Previsto	Erro (%)	Correlação Linear
SUL	1	8885,29	8778,96	1,19	0,971
	2	8815,71	8641,86	1,97	
	3	8780,00	8587,96	2,18	
	4	8632,29	8574,12	0,67	
	5	8732,29	8704,81	0,31	

As tabelas 1 a 4 apresentam os passos (semanas) a frente, os valores originais da série, os valores previstos, o erro percentual e o índice de correlação linear de Pearson, respectivamente, para todas as regiões do brasil.

Através da visualização das tabelas podemos perceber o quão próximos são os valores da previsão dos valores originais. Também percebemos, o baixo erro percentual para todas as regiões previstas, dentro do limite esperado pelas regras estabelecidas pela ANEEL e por fim percebemos o método de avaliação dos coeficientes de correlação linear de Pearson, onde todas as previsões são bastante satisfatórias com índices bem próximos de 1.

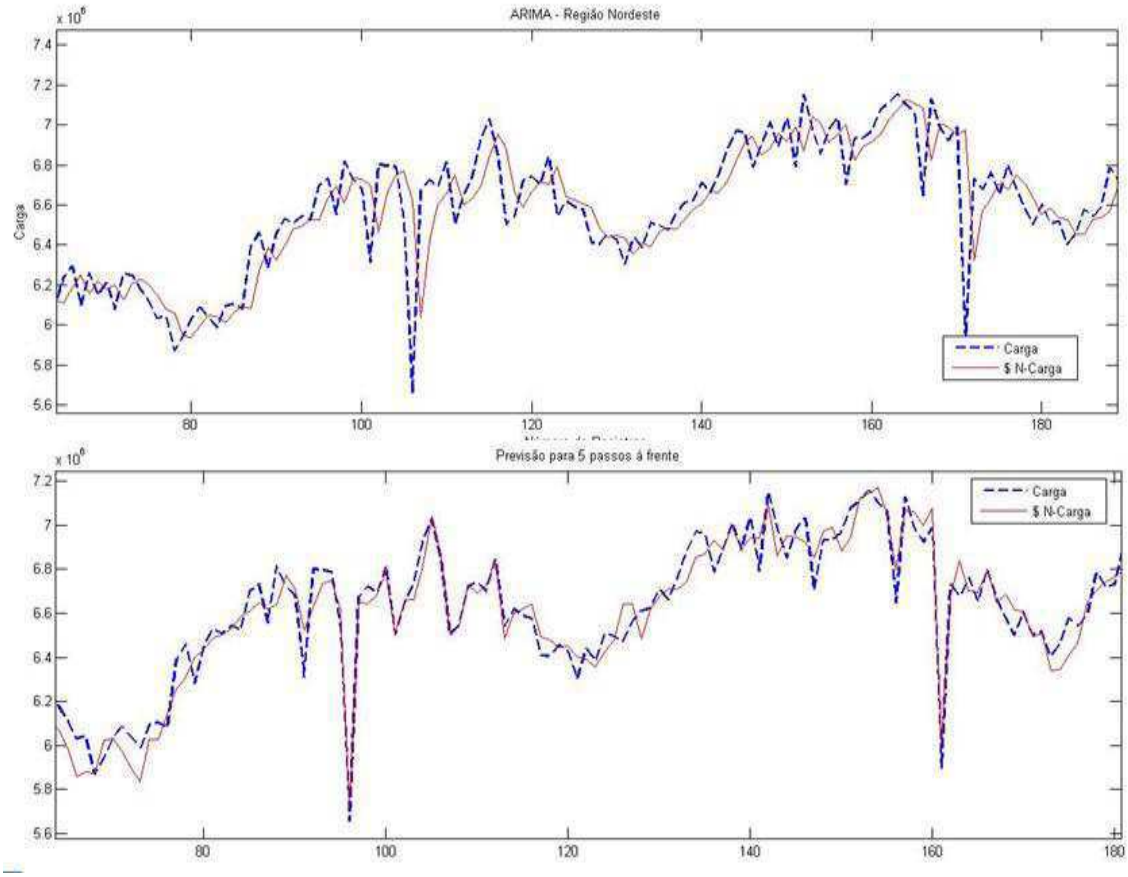

Figura 32 – Comparação entre a previsão do modelo ARIMA (figura de cima) e o modelo Híbrido ARIMA –ANN (figura de baixo).

A figura 32 mostra de forma clara a comparação gráfica entre as previsões utilizando apenas o modelo ARIMA e o modelo Híbrido ARIMA-RNMC, escolhida uma das regiões (Nordeste), nos posibilitando visualizar a grande vantagem de se utilizar um modelo híbrido para a previsão futura de cargas de energía.

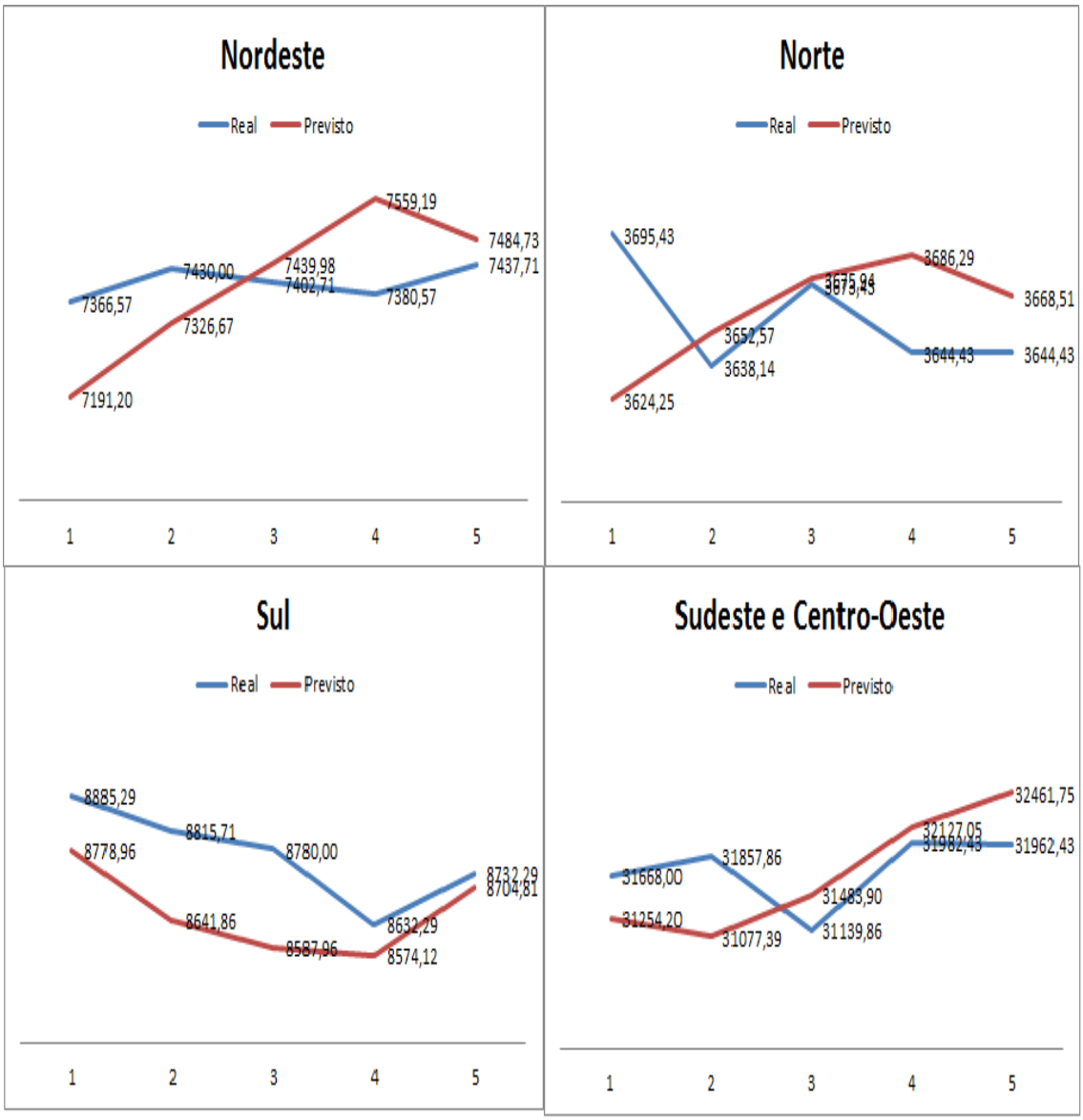

Figura 33 – Gráfico das comparações entre valores reais e previstos para 5 passos a frente.

A figura 33 mostra a comparação gráfica entre os valores reais e os valores dos passos previstos no nosso modelo híbrido para as regiões do Brasil. Portanto, é uma forma de visualizar a partir do momento em que a rede neural deixa de ser treinada e passa a ser validada.

4 CONCLUSÕES

A utilização do modelo híbrido gerou resultados mais consistentes do que o uso isolado de redes neurais de múltiplas camadas e da metodologia ARIMA. O modelo híbrido apresentou excelentes indicadores de correlação linear de Pearson entre séries alvo e predita e com boa capacidade de generalização para até 5 passos a frente.

O aprendizado das redes neurais de múltiplas camadas foi de implementação simples através da ferramenta IBM SPSS Modeler e possibilitou a criação de um modelo computacional leve que consumiu apenas 5s para gerar os resultados.

Os resultados são conclusivos em apontar o modelo híbrido como um método efetivo para a predição futura da série de carga de energia para as regiões brasileiras para até 5 passos a frente, além de que as previsões alcançaram índices de erro muito baixos, como o previsto pelas regras da ANEEL.

A técnica pode ser testada com outras séries temporais, principalmente aquelas que afetam as decisões de operação do Sistema Interligado Nacional – SIN e do Operador Nacional do Sistema – ONS.

5 REFERÊNCIAS BIBLIOGRÁFICAS

ABURTO, L.; WEBER, R. **Improved supply chain management based on hybrid demand fore-casts. Applied Soft Computing**, Santiago, v. 7, n. 1, 2007 - p. 136-144.

BOX, G. E. P.; PIERCE, D. A. **Distribution of residual autocorrelations is autoregressive-integrated moving average time series models. Journal of the American Statistical Association**, New York, v. 65, n. 332, p. 1509-1526: 1970.

BOX, G.E.; JENKINS, G.M**. Times series analysis: forecasting and control.** San Francisco: Holden - Day, 1976. 575 p.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDEMIR, T. B**. Redes neurais artificiais: teoria e aplicações**. 2.ª ed. Rio de Janeiro: LTC: 2000. 250 p.

BROCKWELL, P., J., and DAVIS, R.A**. Introduction to time series and forecasting.** Springer, New York: 2002.

BROWN, R.G. **Smoothing, Forecasting and Prediction of Discrete Time Series**. 12 ed. New Jersey: Prentice-Hall: 1963, 468 p.

CHAPMAN, P et. al. **CRISP-DM 1.0.** CRISP-DM consortium: 2000.

CIOS, K. J; PEDRYCZ, W; SWINIARSKI, R. W; KURGAN, L. A. **Data Mining – AKnowledge Discovery Approach.** Springer: 2007.

Cross Industry Standard Process for Data Mining (CRISP-DM) Disponível em: <http://www.crisp-dm.org/>. Acesso em: 12 de jan. de 2014.

CYBENKO, G. **Approximation by superpositions of sigmoidal function. Mathematics of Control Signals, and Systems**, New York, v. 2, n. 4:1989 - p. 303-314.

DASH, P. K., SATPATHY, H., LIEW, A. C., and RAHMAN, S. **A real-time short-term load forecasting system using functional link network**, IEEE Trans., 1997, PWRS-12, (2), p. 675-680.

DESOUKY, A., El, and ELKATEB, M, M. **Hybrid adaptive techniques for eletric-load forecast using ANN and ARIMA**, IEEE Proc. Gener., Transm. Distrib: 2000, 147, (4), pp. 213-217.

DILWORTH, James B. - **Operations management: design, planning, and control for manufacturing and services.** Singapura: McGraw-Hill: 1992.

DURBIN, J. **Testing for serial correlation in least-squares regression when some of the re-gressors are lagged dependent variables.** Econometrica, Chicago, v. 38, n. 30, p. 410-412: 1970.

FARUK, D.O. **A hybrid neural network and ARIMA model for water quality time series pre-diction.** Engineering Applications of Artificial Intelligence, Aydin, v. 23, n. 4, 2009 - p. 586-594.

FAYYAD, U.; PIATESKI, S. and SMYTH, P. **The KDD Process for Extracting Useful Knowledge** from Volumes of Data. In: Communications of the ACM, November 1996/vol. 39, no. 11, p. 27-34.

FERREIRA, V.H.; ALVES da Silva, A.P. **Toward Estimating Autonomous Neural Network-Based Electric Load Forecasters,** IEEE Transactions on Power Systems, v.22, n.4, pp. 1554-1562.

GEORGE, E. P. BOX, GWILYM, M. JENKINS, GREGORY, C. REINSEL. **Time series analysis: Forecasting and Control**: 2008.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** Elsevier: 2006.

HARVEY, A.C. **Forecasting, Structural Time Series Models and the Kalman Filter**, Cambrigde University Press, U.K: 1999.

HAYKIN, S. (1999) **Neural networks: a comprehensive foundation.** 2 ed. New Jersey: Prentice-Hall, 1999, p.842.

HEIZER, Jay.; RENDER, Barry. **Operations management.** 7ª ed. Upper Saddle River, NJ: Pearson Education: 2004.

HIPPERT, H. S.; PEDREIRA, C. E.; SOUZA, R. C. **Neural networks for short-term load forecasting: are view and evaluation.,** IEEE Transactions on Power Systems, Piscataway, v. 16, n. 1: 2001 - p. 44-55.

IBM SPSS Modeler Disponível em: <http://www-03.ibm.com/software/products/pt/spss-modeler/> Acesso em: 10 de jan. de 2014.

KALPAKIS, D., GADA, and PUTTAGUNTA, V. **Distance measures for effective clustering of ARIMA time-series.** Proc. Of IEEE Int. Conf. on Data Mining: 2001, pp. 273-280.

MACHADO, Virgílio Cruz; CABRITA, Maria do Rosário. **Técnicas de Previsão. Caparica**: FCT/UNL, 2009. cap. II.

MOGHRAM, I. e RAHMAN, S. **Analysis and evaluation of five short-term load forecasting techniques.** IEEE Transactions on Power Systems, Vol 4: 1989 – N. 4, pp. 1484-1491.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais.** 2 ed. São Paulo: Edgard Blucher: 2006 - 544 p.

O'DONOVAN, T. M. **Short term forecasting: an introduction to the Box-Jenkins approach,** New York: John Wiley & Sons: 1983 – p. 256.

OLSON, D. L; DELEN, D. **Advanced Data Mining Techniques.** Springer, 2008.

PARK, B and KARGUPTA, H. **Distributed Data Mining: Algorithms, Systems, and Applications, Data Mining Handbook:** 2002.

PENGE, T., HUBELE, N., and KARADY, G. (1993) **An adaptive neural network approach to one week ahead load forecasting.** IEEE Trans., 1993, PWRS-8, (3), p. 1195-1202.

STEVENSON, William J. **Production / operations management.** 5ª ed. Chicago: Irwin: 1996, p 496.

U. M. FAYYAD, et al. (Eds.) **Advances in Knowledge Discovery and Data Mining**, AAAI Press/MIT Press: 1996.

WEI, W. W. S. **Time series analysis: univariate and multivariate methods**. 2 ed. New York: Pearson: 2006. p. 634.

WERBOS, P. J. **Beyond regression: new tools for prediction and analysis in the behavioral sciences.** 102 f. Dissertação (Mestrado) - Harvard University, Harvard, 1974.

ZHANG, G. P. **Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing,** Atlanta, v. 50: 2003, p. 159-175.

ANEXO – SOFTWARE IBM SPSS MODELER

A1. Sobre o IBM SPSS Modeler

O software IBM® SPSS Modeler é um conjunto de ferramentas de mineração de dados que lhe permitem desenvolver rapidamente modelos preditivos utilizando o conhecimento do problema a que se quer resolver e implantá-los em operações de negócios para melhorar a tomada de decisão. Concebido em torno do modelo CRISP-DM padrão da indústria, SPSS Modeler suporta todo o processo de mineração de dados, a partir de dados para melhores resultados de negócios ou problemas.

O SPSS Modeler oferece uma variedade de métodos de modelagem tiradas de aprendizado de máquina, inteligência artificial, e as estatísticas. Os métodos disponíveis na paleta Modeling permitem obter novas informações a partir de seus dados e desenvolver modelos preditivos. Cada método tem alguns pontos fortes e é mais adequada para determinados tipos de problemas.

A2. Visão geral do software

Como uma aplicação de mineração de dados, o IBM® SPSS Modeler oferece uma abordagem estratégica para encontrar relações úteis em grandes conjuntos de dados. Em contraste com os métodos estatísticos mais tradicionais, você não precisa necessariamente saber o que você está procurando quando você começa. Você pode explorar os dados, ajuste de diferentes modelos e investigar diferentes relações, até encontrar informações úteis.

Uma visão geral da interface do software pode ser observada na figura abaixo, onde percebe-se que o software fornece ao usuário uma maneira de contruir modelos para determinados objetivos através de uma interface gráfica com o usuário.

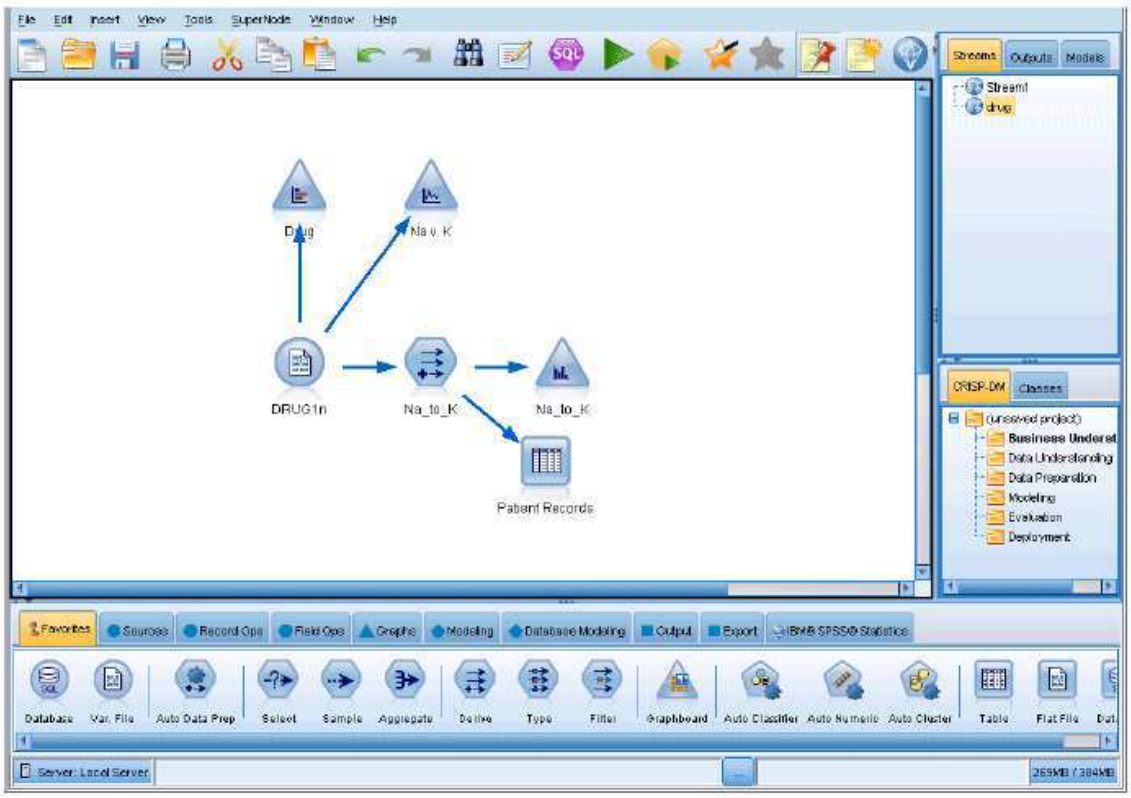

Figura 34 – Interface do sofware.

Em cada ponto do processo de mineração de dados, a interface é fácil de usar, o IBM® SPSS Modeler convida os seus conhecimentos de negócio específico. Algoritmos de modelagem, tais como previsão, classificação, segmentação e detecção de associação, assegurar modelos potentes e precisos. Os resultados do modelo podem ser facilmente implantados e lidos em bancos de dados, IBM® SPSS Statistics, e uma grande variedade de outras aplicações.

Trabalhar com SPSS Modeler é um processo de três etapas de trabalho com dados.

- Em primeiro lugar, você lê os dados no SPSS Modeler.
- Em seguida, você executa os dados por meio de uma série de manipulações.
- Finalmente, você envia os dados para um destino.

Esta seqüência de operações é conhecida como *data stream*, porque os fluxos de dados vão passando registro por registro da fonte através de cada manipulação e, finalmente, para o destino, ou um modelo ou tipo de saída de dados.

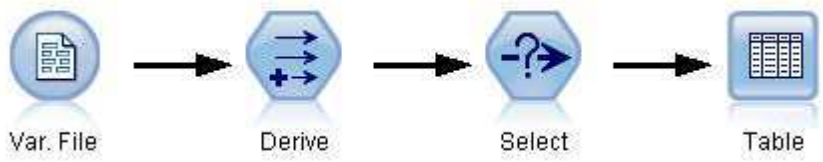

Figura 35 – Sequência de uma data stream básica.

A3. IBM SPSS Modeler stream canvas

A *stream canvas* é a maior área da janela do IBM ® SPSS Modeler e é o lugar onde você vai construir e manipular *data streams*.

Streams são criados pelo desenho de diagramas de operações de dados relevantes para o seu objetivo na tela principal da interface. Cada operação é representado por um ícone ou nó, e os nós estão ligados entre si em uma stream que representa o fluxo de dados através de cada operação. Você pode trabalhar com múltiplos fluxos de uma só vez em SPSS Modeler, ou na mesma tela corrente ou abrindo uma nova tela corrente. Durante a sessão, as streams são armazenadas no gerenciador de streams, no canto superior direito da janela do SPSS Modeler.

A4. Paleta de nós ou funções do software

A maioria das ferramentas de modelagem e de manipulação dos dados no IBM® SPSS Modeler se encontra na paleta de nós, na parte inferior da janela abaixo da stream canvas.

Por exemplo, na guia da 'Record Ops' contém os nós que você pode usar para executar operações sobre os registros de dados, como a seleção, fusão e acrescentar.

Para adicionar nós para a tela, clique duas vezes em ícones da Paleta de nós ou arraste e solte-os na tela. Você, então, pode conectá-los e criar uma stream, o que representa o fluxo de dados.

Figura 36 – Record Ops na Paleta de nós.

Cada guia da paleta contém uma coleção de nós relacionados utilizados para diferentes fases das operações das streams , tais como:

- *Sources* – São nós que trazem dados para o SPSS Modeler.
- *Record Ops* – Nós que realizam operações em registros de dados , como a seleção , fusão, e acrescentar .
- *Field Ops* – Nós que executam operações em campos de dados , como filtrar , derivar novos campos e determinar o nível de medida para determinados campos.
- *Graphs* – Nós que apresentam graficamente os dados antes e depois da modelagem. Os gráficos incluem gráficos como histogramas, nós web e gráficos de avaliação.
- *Modeling* – Nós que usam os algoritmos de modelagem disponíveis no SPSS Modeler , como redes neurais, árvores de decisão, algoritmos de agrupamento e sequenciamento de dados.
- *Database modeling* – Nós que usam os algoritmos de modelagem disponíveis no Microsoft SQL Server, IBM DB2 e Oracle.
- *Output* – Nós que produzem uma variedade de saída para dados , gráficos e os resultados do modelo que podem ser vistos no SPSS Modeler.
- *Export* – Nós que produzem uma variedade de saídas que pode ser vistas em aplicações externas, tais como a coleta de dados do IBM® SPSS ou Excel.
- *SPSS Statistics* – Nós que importam dados de, ou exportam dados para , IBM® SPSS Statistics, ou outros softwares que o minerador desejar, bem como executam procedimentos SPSS Statistics.