UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA DE PRODUÇÃO

PARAMETRIZAÇÃO DE REDES NEURAIS ARTIFICIAIS EM PROBLEMAS DE SÉRIES TEMPORAIS NÃO LINEARES EMPREGANDO PBCA (*PLACKETT-BURMAN CORRELATION ANALYSIS)*

Mariângela de Oliveira Abans

Itajubá, abril de 2018

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA DE PRODUÇÃO

Mariângela de Oliveira Abans

PARAMETRIZAÇÃO DE REDES NEURAIS ARTIFICIAIS EM PROBLEMAS DE SÉRIES TEMPORAIS NÃO LINEARES EMPREGANDO PBCA (*PLACKETT-BURMAN CORRELATION ANALYSIS)*

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de *Doutor em Ciências em Engenharia de Produção*.

Área de Concentração: Engenharia de Produção

Orientador: Dr. Pedro Paulo Balestrassi

Itajubá, 20 de abril de 2018

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA DE PRODUÇÃO

Mariângela de Oliveira Abans

PARAMETRIZAÇÃO DE REDES NEURAIS ARTIFICIAIS EM PROBLEMAS DE SÉRIES TEMPORAIS NÃO LINEARES EMPREGANDO PBCA (*PLACKETT-BURMAN CORRELATION ANALYSIS)*

Tese aprovada por banca examinadora em 20 de abril de 2018, conferindo ao autor o título de ***Doutor em Ciências em Engenharia de Produção***.

Banca Examinadora:
Dr. Pedro Paulo Balestrassi (Orientador)
Dr. Antônio Carlos Zambroni de Souza
Dr. Eder Martioli
Dr. Ronã Rinston Amaury Mendes
Dr. José Henrique de Freitas Gomes

Itajubá
2018

DEDICATÓRIA

Ao meu esposo Max e aos meus filhos Ariel e Rodrigo,

porque como família crescemos,

como família lutamos,

e como família vencemos

AGRADECIMENTOS

Ao Prof. Pedro Paulo Balestrassi, meu orientador nesta jornada e nesta vida.

Ao Prof. Antônio Carlos Zambroni de Souza, que me acolheu na UNIFEI e que sempre me apoiou e incentivou.

Ao Prof. Anderson Paulo de Paiva, pelas frutíferas trocas de ideias e por seu tempo.

Aos Doutores Eder Martioli, Ronã Rinston Amaury Mendes, José Henrique de Freitas Gomes e novamente ao Antônio Carlos Zambroni de Souza por terem aceito o desafio de lapidar o projeto inicial em uma pesquisa de mérito científico e de aplicação em mais de uma área do conhecimento.

Aos professores e funcionários do Instituto de Engenharia de Produção e Gestão da UNIFEI, pelo apoio e ajuda constantes.

Aos colegas do IEPG, que compartilharam comigo estes anos de muito trabalho, alegrias, preocupações e de regozijo, afinal. Em especial, à Gabi Amorim, João Éder, Juliana, Paulinho, Fabrício, Vinícius Paes, Vinícius Rennó, Taynara, Lucas, Tábata, Júlio, Bárbara, Gabi Belinato, Carol, Dani, Rachel e Rafael Miranda.

Aos amigos e colegas do Laboratório Nacional de Astrofísica, que sempre me apoiaram e incentivaram, e que me ajudaram a seguir em frente apesar das adversidades esperadas e inesperadas.

Ao Júlio Didier Maciel, pela inestimável ajuda com a programação em MSExcel® e com os cálculos.

À CAPES, CNPq, FAPEMIG e MCTIC, pelo apoio financeiro.

Aos meus amigos e familiares, vivos e *in memoriam*, com os quais encontrei a força, a perseverança, a tenacidade e a esperança.

Aos meus queridos Ariel e Rodrigo, pelos quais tudo vale a pena, pelos quais a própria vida já terá valido a pena.

Ao Max, como se palavras pudessem mensurar minha gratidão. Você foi o porto seguro nas tempestades que eu enfrentei.

"*Most data will never be seen by humans*"

Petr Škoda

RESUMO

Séries temporais são encontradas em fenômenos naturais, mercadológicos e financeiros, e em processos de manufatura, entre outros, o que as torna importantes objetos de pesquisa. Quando não são lineares, sua modelagem é complexa devido ao grande número de parâmetros, à existência de fatores exógenos, à alta volatilidade e à presença de sazonalidade. Tem sido demonstrado que, nesses casos, RNAs apresentam bom desempenho tanto na apreensão do comportamento dos dados como na sua previsão dentro dos limites de exatidão requeridos. Apesar do grande número de parâmetros de uma RNA, delineamentos têm sido usados com sucesso, porém a alta demanda de recursos humanos, computacionais, financeiros e de tempo são obstáculos para sua total exploração. Neste trabalho, implantou-se uma nova metodologia denominada Análise de Correlação em *Plackett_Burman* (*Plackett_Burman Correlation Analysis*−PBCA) na parametrização de Redes Neurais Artificiais − RNAs, com o propósito de modelar e prever Séries Temporais Univariadas Não Lineares. Esta é uma metodologia de seleção de variáveis significativas baseada no Delineamento de Experimentos (DOE) de *Plackett_Burman* com rebatimento e que propõe a análise de correlação entre as séries de resíduos, consideradas como sinais. Sua vantagem é requerer um número de experimentos menor que o Delineamento Fatorial Completo e ser capaz de identificar as significativas interações de segunda ordem entre todas as variáveis. Nesta tese, implementou-se esta nova metodologia PBCA na construção de RNAs previsoras para auxiliar os profissionais responsáveis por decisões tático-estratégicas baseadas em séries temporais. A metodologia foi aplicada a dois casos: (i) número de horas disponíveis para pesquisa em um observatório astronômico profissional de classe mundial e (ii) distribuição de carga elétrica fornecida a uma empresa brasileira, com o objetivo de fornecer previsões de curtíssimo e curto prazos para a tomada de decisões sobre o melhor uso das infraestruturas envolvidas. Ambas as séries foram primeiramente estudadas através da aplicação de técnicas e modelos ditos "clássicos" a fim de estabelecer *benchmarking* para comparação. Os resultados deste trabalho sugerem a adequação apenas parcial da metodologia para estes fins. Não é possível aplicar a PBCA totalmente devido (i) à maneira de definir as RNAs, (ii) ao fato do processo não ser modelável, afinal e (iii) à impossibilidade de uso das interações de ordem dois entre as variáveis significativas no software Statistica®. São também apresentados possíveis desdobramentos da pesquisa e aplicações em outras áreas do conhecimento.

Palavras-chave: Séries Temporais – Redes Neurais Artificiais – Delineamento de Experimentos – *Plackett_Burman* – Análise de Correlação

ABSTRACT

Time series are found in natural, market and financial phenomena, and in manufacturing processes, among others, which makes them important research objects. When they are not linear, their modeling is complex due to the large number of parameters involved, the existence of exogenous factors, the high volatility and the presence of seasonality. It has been demonstrated that, in these non-linear cases, ANNs perform well both in the apprehension of data behavior and in their prediction within the required accuracy limits. Despite the large number of parameters of an ANN, designs have been used with success, but the high demand for human, computational, financial and time resources are obstacles to their full exploitation. In this work, a new methodology called *Plackett_Burman* Correlation Analysis (PBCA) was implemented in the parameterization of Artificial Neural Networks − ANNs, with the purpose of modeling and predicting univariate non-linear time series. This is a methodology for the selection of significant variables based on the *Plackett_Burman*'s Experimental Design with folding and that proposes the analysis of correlation between the series of residues, considered as signals. This design has the advantage of requiring a smaller number of experiments than the Full Factorial Design and of being able to identify the significant second order interactions among all the variables. In this thesis, this new methodology PBCA was implemented in the construction of predictive ANNs to assist those professionals responsible for strategic-tactical decisions based on time series. The methodology was applied to two cases: (i) number of hours available for research in a world-class professional astronomical observatory and (ii) distribution of electric charge supplied to a Brazilian company, with the aim of providing very short and short-term forecasts to support decisions about the best use of the infrastructures involved. Both series were first studied through the application of the so-called "classical" techniques and models in order to establish benchmarking for comparison. The results of this work suggest the partial-only adequacy of the methodology for these purposes. It is not possible to totally apply the PBCA due to (i) the way of defining the RNAs, (ii) the fact that the process cannot be modeled after all, and (iii) the impossibility of using second-order interactions among the identified significant variables with the software Statistica®. Possible research developments and applications in other areas of knowledge are also presented.

Keywords: Time Series – Artificial Neural Networks – Design of Experiments – *Plackett_Burman* – Correlation Analysis

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACF	*Função de Autocorrelação (Autocorrelation Function)*
ADU	Unidades digitais oriundas da conversão do número de elétrons captados por uma câmera digital *(Analog-to-Digital Units)*
ANN	*Artificial Neural Networks* (v. RNA−Redes Neurais Artificiais*)*
AR	Autoregressiva *(Auto-regressive)*
ARIMA	Modelo Autorregressivo Integrado de Média Móvel (*Autoregressive Integrated Moving Average)*
CCDE	Detector de Carga Acoplada *(Charge-Coupled Device* − sensor de luz; CCD na literatura astronômica)
DOE	Delineamento de Experimentos **(***Design of Experiments*)
FWHM	Largura a Meia Altura do perfil da PSF *(Full Width at Half Maximum)*
IA	Inteligência Artificial
IEPG	Instituto de Engenharia de Produção e Gestão
IID	Idêntica e Independentemente Distribuído (*Independently Identically Distributed)*
LNA	Laboratório Nacional de Astrofísica
MA	Média Móvel (*Moving Average*)
MAD	Desvio Médio Absoluto (*Mean Absolute Deviation*)
MAPE	Erro Porcentual Absoluto Médio (*Mean Absolute Percentage Error*)
MCTIC	Ministério da Ciência, Tecnologia, Inovações e Comunicações
MdAPE	Erro Porcentual Absoluto Mediano (*Median Absolute Percentage Error*)
MLP	Perceptrons em Camadas Múltiplas (*Multilayer Perceptrons)*
MSD	Desvio Padrão Quadrático da Média (*Mean Squared Deviation*)
MVAP	Média mais o Valor de Amplitude de Pico do Coeficiente de Correlação dos Sinais
NOAO	*National Optical Astronomy Observatory* (centro norte-americano de P&D que opera vários telescópios em terra)
OPD	Observatório do Pico dos Dias
P&D	Pesquisa e Desenvolvimento
PACF	Função de Autocorrelação Parcial (Partial Autocorrelation Function)
PB	(Arranjo fatorial de) *Plackett-Burman*

PBCA	Análise de correlação dos sinais de resíduos gerados pelo delineamento de *Plackett-Burman* (*Plackett-Burman Correlation Analysis)*
PBSA	Análise de Sensibilidade utilizando o delineamento de *Plackett-Burman* (*Plackett-Burman Sensitivity Analysis)*
PSE	Pseudo Erro Padrão *(Pseudo Standard Error)*
PSF	*Point Spread Function*
P-*value,* P	Menor nível de significância que conduz à rejeição da hipótese nula (H_0)
QD	Qualidade dos Dados (*Data Quality*)
RNA	Redes Neurais Artificiais (v. ANN)
S/N	Relação Sinal-Ruído (*Signal-to-Noise*)
SARIMA	Modelo de série temporal de Box-Jenkins Sazonal (*Seasonal Box-Jenkins)*
STAR	Autoregressiva de Transição Suave (*Smooth Transition Autoregressive)*
TAR	Autoregressiva de Limiar (*Threshold Autoregressive)*
USD	Dólares Americanos (*United States Dollar*)
VAPP	Valor da Amplitude Pico a Pico
VAPPCC	Valor de Amplitude Pico a Pico do Coeficiente de Correlação
VCC	Valor do Coeficiente de Correlação

LISTA DE SÍMBOLOS

a_{t-1}	Sequência de variáveis aleatórias
E	Esperança matemática ou valor esperado ou média
$f_i(t)$	Série temporal
$I_{\Delta\lambda}$	Intensidade de fonte luminosa em determinado filtro
k	Número de fatores de um DOE
$m_{\Delta\lambda}$	Magnitude medida com determinado filtro
$M_{\Delta\,\lambda}$	Magnitude absoluta de um corpo celeste em determinado filtro
N	Número de experimentos fatoriais
p	Número de geradores de viés independentes
r	Coeficiente de correlação
s	Desvio padrão de uma amostra
t	Tempo
Var	Variância
x_t	Série temporal, tal que $t = 1, \dots, T$
X	Massa de ar
z	Ângulo zenital
α	Nível de significância, máximo nível aceitável para o risco de se rejeitar a hipótese nula quando deveria ocorrer o oposto (Erro do Tipo I)
β	Poder de um teste estatístico
ε	Erro experimental (aleatório ou aleatório + caótico)
λ	Comprimento de onda da radiação eletromagnética (luz)
Δλ	Intervalo de comprimentos de onda da banda passante de um filtro
μ	Média populacional
ν	Graus de liberdade
σ	Desvio padrão de uma população
σ^2	Variância, se a função for Gaussiana
ψ_i	Conjunto de números reais tais que $\psi_0 = 1$

SUMÁRIO

1. Introdução

1.1 . Considerações iniciais

Há inúmeros **fenômenos** que **variam** com o passar do **tempo** (*e.g.* mercadológicos, experimentais, industriais e financeiros). Em diversas áreas do conhecimento, é **necessário** e **fundamental prever** seu **comportamento** em diferentes escalas de tempo, mas como fazer isso de forma adequada e confiável não é um problema de simples solução.

Essa variabilidade temporal pode ser linear ou não, sazonal, caótica; pode depender de um ou mais fatores e/ou ocorrer em regime de vários níveis. As **séries temporais não lineares** são de especial interesse devido à sua **frequente ocorrência** em áreas de vital **importância** para a **sociedade**, o **mercado**, o **ecossistema**, o **tempo** e o **clima**, a distribuição de **carga de potência** e a **manufatura**, entre outros (INMAN *et al.*, 2013).

A **primeira pergunta** que se tem a responder, então, é: como prever eficiente e economicamente uma série temporal? A solução aqui escolhida é: via Redes Neurais Artificiais (RNAs).

Redes Neurais Artificiais são especialmente **hábeis** em lidar com **problemas não lineares** devido à sua capacidade de aprender e serem capazes de fornecer informações sem depender de modelos parametrizados (SRIVASTAVA *et al.*, 2014). São estruturas computacionais **orientadas aos dados**, livres de hipóteses acerca dos fenômenos em estudo, e livres de condições de contorno dos problemas (TIROZZI *et al.*, 2006).

Um **problema**, no entanto, **tem perdurado** por décadas. E a **segunda pergunta** é: como **simular uma RNA** (entenda-se treinar) ao mesmo tempo em que os muitos parâmetros que a caracterizam são identificados como produtores do menor erro? **Uma forma** é a tentativa e erro – entenda-se força bruta −, na qual se testam diversos modelos de RNAs, à exaustão, até que se alcance a precisão e a confiabilidade desejadas.

Com o advento da popularização do Delineamento de Experimentos (DOE), **a segunda forma** de otimização das características das RNAs é aplicar – novamente por tentativa e erro – diversos arranjos experimentais na construção das RNAs a fim de identificar aquela com melhor desempenho (entenda-se, menor erro) dentro da região experimental de interesse e

segundo o grau de qualidade pré-estabelecido como aceitável para cada caso (BALESTRASSI *et al.,* 2009).

Frente a um cenário prático-acadêmico, onde profissionais devem tomar decisões que envolvem custos de toda espécie dentro de um determinado horizonte temporal, **pergunta-se**: haveria uma metodologia menos custosa, mais direta, talvez um novo paradigma na aplicação de DOE a RNAs com a finalidade de propiciar previsões de fenômenos não lineares univariados?

Fornecer uma **resposta afirmativa** a essa pergunta é a **proposta desta pesquisa**. A **terceira forma**, então, é a Análise de Correlação dos resíduos em arranjos fatoriais de *Plackett_Burman* − um tipo especial de DOE que envolve apenas um passo (PBCA, COUTO, 2012) − aplicada à definição da arquitetura interna de RNAs previsoras. Esta pesquisa também segue parte da metodologia descrita em Balestrassi *et al.* (2009). A **Erro! Fonte e referência não encontrada.** apresenta a dicotomia entre os grandes conjuntos de métodos de tratamento e análise de séries temporais e resume o algoritmo da metodologia aqui proposta.

1.1.1. Contribuições científicas e tecnológicas

- Implantação inédita do uso da Análise de Correlação em arranjos fatoriais de *Plackett_Burman* modificados na modelagem da arquitetura interna de RNAs previsoras,
- Contribuição eficaz para a eficiente seleção de variáveis e de suas interações de segunda ordem que sejam significativas em modelagem fatorial, com um mínimo de recursos computacionais, financeiros, de infraestrutura e de recursos humanos,
- Abertura de campo de pesquisa em, principalmente, Engenharia de Produção, Tecnologia da Informação, Engenharia e Ciências da Computação.
- Disponibilização do modelo conceitual/algoritmo para empresas, universidades, centros de pesquisa e inovação tecnológica em diferentes áreas do conhecimento,

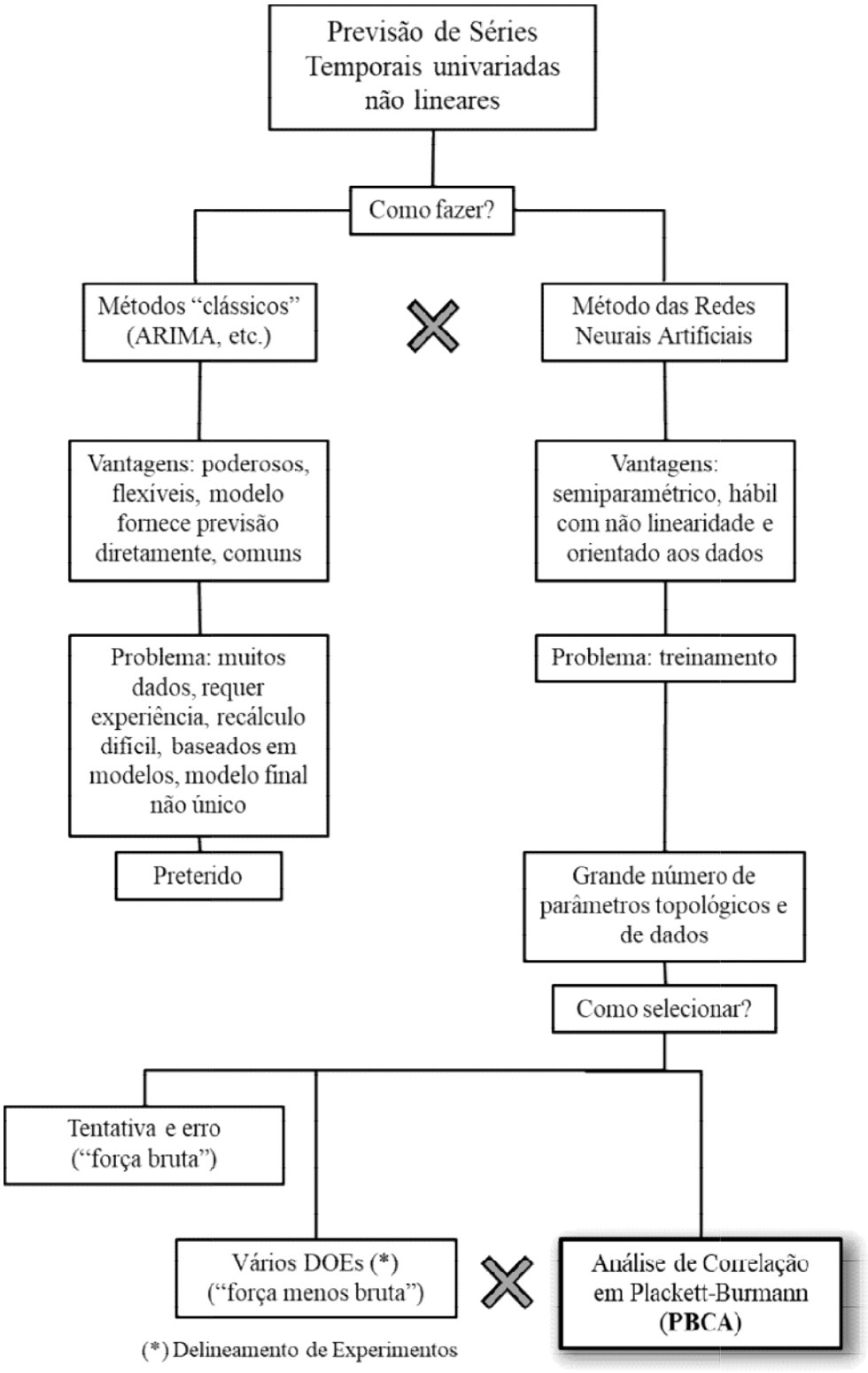

Figura 1.1 - Visão resumida das escolhas para a metodologia desta pesquisa.

1.2 . Objetivo geral

Implantar metodologia que permita a identificação robusta, em apenas **um DOE**, de parâmetros arquitetônicos de RNAs previsoras quando aplicadas a casos de séries temporais univariadas não lineares.

1.3 . Objetivos específicos

- Modelar, de forma bem-sucedida, a arquitetura interna de RNAs previsoras através da metodologia PBCA.
- Verificar a adequação de se avaliar o desempenho das RNAs através do cálculo do Erro Porcentual Absoluto Médio (MAPE) e do Erro Porcentual Mediano (MdAPE) no processo de validação da metodologia.
- Implementar a metodologia a dois casos reais de naturezas diferentes e verificar sua adequação.
- Prever o comportamento dessas duas séries temporais com o objetivo de propiciar subsídios para tomadas de decisões por parte dos *stakeholders*.
- Por último, disponibilizar uma metodologia econômica e técnicas eficientes aos profissionais responsáveis por tomadas de decisões tático-estratégicas da indústria e da academia, entre outros.

1.4 . A pesquisa

Trata-se, então, de uma pesquisa de natureza aplicada, com objetivo normativo, cuja abordagem é quantitativa e o método é modelagem e simulação (MIGUEL, 2012) – v. maiores detalhes da classificação de pesquisas no Apêndice A.

1.5 Delimitações

As principais delimitações desta proposta estão elencadas a seguir.

- Por empregar delineamentos do tipo *Plackett_Burman*, a PBCA de dois níveis requer um número de experimentos que seja múltiplo de 4 a fim de que os efeitos principais sejam determinados com precisão (Plackett; Burman, 1946a).

- No delineamento de experimentos, as variáveis independentes possuem níveis definidos, os quais delimitam a região de operação; a modelagem só pode ser feita dentro destes limites.
- Nem todas as interações de segunda ordem das variáveis podem ser analisadas simultaneamente por motivos de método e de capacidade computacional.
- A pesquisa requer fortemente o uso de computadores com alta velocidade e grande volume de memória.
- O número de graus de liberdade para o estudo das interações de segunda ordem restringe a escolha desses pares, que deve basear-se em critérios estatísticos, mas *extra* PBCA.
- Essa escolha das interações significativas deve seguir metodologia empírica específica que, embora seja estatisticamente justificada, envolve tentativa e erro.
- RNAs necessitam dados em quantidade suficiente para serem treinadas e testadas.
- A precisão das previsões fornecidas por RNAs depende da riqueza de dados.
- Requer pessoal treinado em estatística, experimentação, análise de dados e de resultados.
- Em ambientes empresariais e de manufatura, o uso de experimentação na previsão de séries temporais encontra uma certa resistência por parte das chefias e gerências no que diz respeito à modelagem e DOE.

1.6 . Estrutura da tese

O Capítulo 2 contém a fundamentação teórica e revisão da literatura. Já no Capítulo 3, apresenta-se a metodologia PBCA em detalhe. No Capítulo 4 apresenta-se metodologia de trabalho na aplicação da metodologia PBCA. No Capítulo 5 são apresentadas as aplicações aos casos reais. O Capítulo 6 traz as conclusões da tese e os possíveis trabalhos futuros.

Dada a interdisciplinaridade e a amplidão dos campos envolvidos, optou-se por simplificar o texto do corpo da tese sempre que possível. Destarte, incluíram-se: o Apêndice A, que contém a classificação da pesquisa; o Apêndice B, o qual contém o pseudocódigo para a implantação, passo a passo, da PBCA em três fases; o Apêndice C, que traz o código do cálculo do arranjo experimental e da criação das matrizes para os cálculos das regressões e análise de sinais em Scilab®, e o Apêndice D contém o código em VBA/Excel® para o cálculo das regressões para os fatores e para as interações, das matrizes de correlação entre as séries de resíduos e

dos diversos indicadores da qualidade dos resultados da PBCA. O Anexo A traz a lista das principais produções acadêmicas da autora desta tese.

2. Fundamentação teórica

Dada a importância da previsão de séries temporais, na seção 2.1 apresenta-se a visão geral das publicações sobre o tema e as ferramentas utilizadas nesta pesquisa. Na seção 2.2, discorre-se sobre séries temporais, o que são e quais abordagens têm sido usadas em sua análise e previsão. A seção 2.3 contempla RNAs e o problema do treinamento quando de sua aplicação nesse campo. Na seção 2.4, apresenta-se o caso dos DOEs em simulação.

2.1. Panorama das publicações

O levantamento do material publicado estendeu-se até o ano de 2017. Dado o crescente volume de publicações disponíveis, foi necessário traçar um plano de ação; os três primeiros capítulos de Nisbet *et al.* (*2009*) trazem conceitos e diretrizes interessantes sobre Descoberta do Conhecimento e as formas de expressar a informação coletada. No caso desta tese, e guardadas as devidas proporções, isto envolveu o garimpo dos dados de interesse e como expressá-los de forma adequada à extração e apresentação de informação útil para o trabalho e quem dele se beneficiar. Buscas preliminares mostraram que os diferentes assuntos desta pesquisa não interessam unicamente a pesquisadores nas áreas das engenharias, mas também em física, astronomia e principalmente, em farmácia, biologia, biotecnologia, ecologia e psicologia.

A escolha das bases de periódicos, consequentemente, reflete a preocupação de conhecer como os métodos e técnicas são mais empregados. As bases estão disponíveis através do Portal de Periódicos da CAPES/MEC.

As expressões usadas nas buscas foram: (a) "*Plackett_Burman residuals correlation analysis*", (b) "*Plackett_Burman correlation analysis*", (c) "*Plackett_Burman*" e "*sensitivity analysis*", (d) "*Plackett_Burman sensibility analysis*", (e) "*Plackett_Burman*", (f) "*Neural network*" e "*time series*", (g) "*Neural network*" e "*time series*" e "*Plackett_Burman*", e (h) "*Neural network*" e "*Plackett_Burman*".

Apesar da existência de tradutores *online,* optou-se por incluir apenas o Português, Inglês, Francês, Alemão e Espanhol. Foram selecionados apenas artigos publicados em periódicos com revisão por pares e publicações em anais de congressos, já que desenvolvimento instrumental e análise de desempenho nem sempre terminam em publicações, mesmo de cunho técnico. Embora a existência de fator de impacto seja recomendável, esta condição foi

relaxada porque os artigos exatamente na área deste trabalho foram poucos em comparação com os temas mais abrangentes. Não foi feita restrição quanto à data de início da filtragem ou aos tipos de bases indexadas.

Foram pesquisadas várias bases, mas as que renderam resultados positivos foram: ISI/Web of Science, Scopus/Elsevier, Emerald, IEEE Xplore e PubMed.

Um resultado imediato e interessante dessa busca é que não há publicações propondo qualquer tipo de PBCA. Tampouco há mais que uma centena de artigos que empregam a metodologia PBSA (*Plackett_Burman Sensitivity Analysis* − BERES; HAWKINS, 2001), na qual a PBCA é baseada.

Para ilustrar uma das nuances da complexidade deste tipo de mineração, a Figura 2.1 apresenta uma comparação entre os resultados da busca de artigos na base do IEEE Xplore empregando as palavras-chave fornecidas pelos autores e aquelas atribuídas pelo serviço da base. A diferença pode dever-se à maior liberdade que os autores têm para melhor caracterizar suas publicações, e isso torna a busca mais eficiente porque é mais fácil coadunar os termos.

Figura 2.1 - Comparação entre as palavras-chave mais empregadas pelos autores dos artigos extraídos da base IEEE Xplore (esq) e aquelas designadas pelo próprio serviço da base (dir).

Em cada base, os resultados da busca foram armazenados em planilhas MSExcel® para manipulação posterior. Quando a base não fornecia esse tipo de saída, gerou-se arquivos em formato *Comma-Separated Values* (".csv") que depois foram transformados em planilhas MSExcel®. As duplicatas surgiram por motivo da superposição de termos nas buscas e foram removidas pelo software. Também houve duplicação de resultados quando as bases foram

juntadas numa só planilha, já que o mesmo artigo, por exemplo, pode aparecer em mais de uma base. Uma primeira leitura dos títulos e resumos serviu para filtrar os artigos que não empregaram realmente os métodos e técnicas de interesse. A leitura em diagonal do corpo do texto serviu para selecionar os mais úteis para a tese. Eventualmente, cerca de um par de centenas de publicações terminaram por ser referenciadas.

A título de exemplo, a Figura 2.2 apresenta as distribuições anuais das publicações encontradas pelas frases "*Plackett_Burman* e *Artificial Neural Networks*" e "*Neural Networks* e *Time Series* e DOE" na base Scopus/Elsevier. Um resultado também interessante é a evolução temporal do número de patentes de processos em biotecnologia envolvendo arranjos de *Plackett_Burman* encontrados nesta mesma base (v. Figura 2.3).

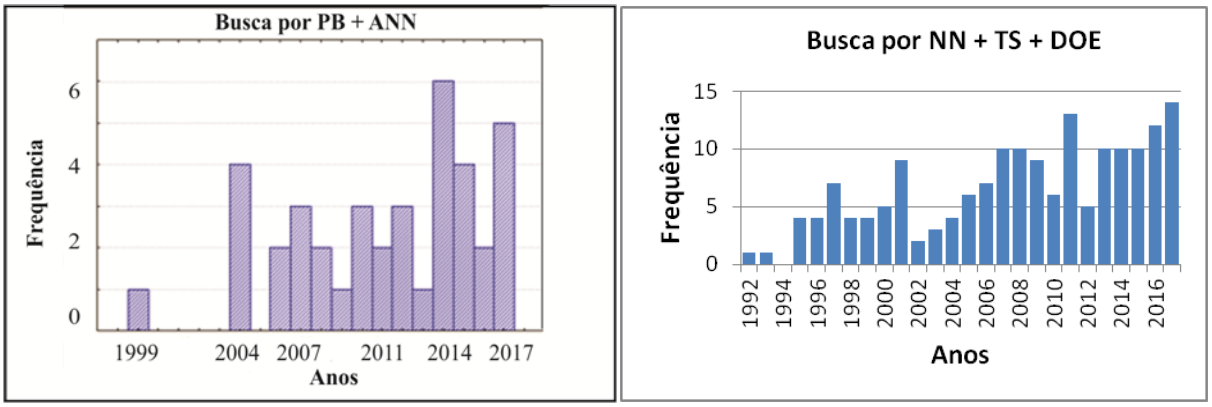

Figura 2.2 - Exemplos de histogramas com resultados de buscas na base Scopus/Elsevier: usando a combinação de palavras-chave *Plackett_Burman* e *Artificial Neural Networks* (esq.), e *Neural Networks* e *Time Series* e DOE (dir.).

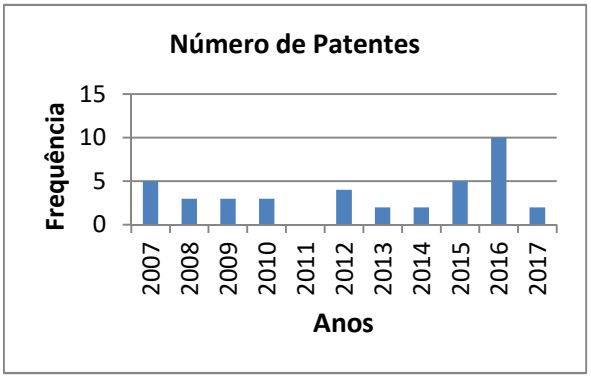

Figura 2.3 – Evolução temporal do número de patentes de processos envolvendo PB segundo a base Scopus/Elsevier.

A metodologia PBCA foi proposta com base no trabalho de Beres e Hawkins em 2001. Na base ISI/Web of Science encontrou-se os resultados apresentados na Figura 2.4, que evidenciam a meia-vida longa dessa proposta. Já o Quadro 2.1 evidencia a diversidade das áreas de pesquisa nas quais trabalhos baseados em Beres e Hawkins (2001) têm sido desenvolvidos através dos diferentes títulos de periódicos.

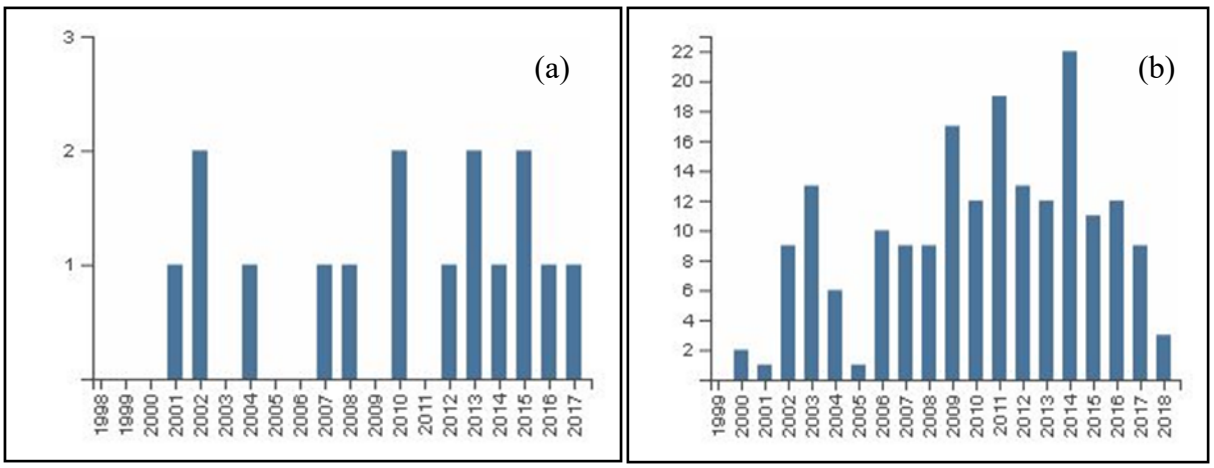

Figura 2.4 – Acompanhamento do uso da metodologia proposta por Beres e Hawkins (2001) (a) e das citações de todos os trabalhos encontrados na base ISI/Web of Science (b).

Quadro 2.1 – Diversidade das áreas de aplicação da PBSA

Periódicos/congressos com publicação/apresentação de trabalhos baseados em Beres e Hawkins (2001)
17th International Conference On Vlsi Design, Proceedings: Design Methodologies For The Gigascale Era
2015 Ieee 9th International Symposium On Embedded Multicore/Manycore Systems-On-Chip (Mcsoc)
Annual Reliability And Maintainability Symposium, 2002 Proceedings
Archiv Fur Tierzucht-Archives Of Animal Breeding
Bioprocess And Biosystems Engineering
Clean-Soil Air Water
Ecological Modelling
Environmental Toxicology And Chemistry
Iconbm: International Conference On Biomass, Pts 1 And 2
International Journal Of Numerical Modelling-Electronic Networks Devices And Fields
International Journal Of Precision Engineering And Manufacturing
Journal Of Mechanical Science And Technology
Mechanics Based Design Of Structures And Machines
Rock Mechanics In Civil And Environmental Engineering
Science Of The Total Environment
Weed Technology

A base de dados criada neste levantamento certamente pode ser explorada de diversas maneiras e com vários outros propósitos, mas essas ações são desnecessárias neste trabalho porque os resultados obtidos até aqui são suficientes para estabelecer-se que: o tema da pesquisa é inexplorado, o cenário no qual a tese se insere é atual e de interesse da comunidade técnico-científica e atende a várias áreas de pesquisa tais como em Engenharias, Ciências Exatas e da Terra e Ciências Biológicas, em Ciências da Saúde, Ciências Agrárias e Ciências Humanas (CNPq/MCTIC, 2018).

Em suma, este é um nicho de pesquisa ainda inexplorado, o qual possui mérito científico para constituir um problema de pesquisa interessante, atual e de aplicabilidade, em particular, em várias subáreas da Engenharia de Produção. Esta tese, então, é uma contribuição para o portfólio de soluções técnicas e metodológicas para este problema de pesquisa, qual seja, a parametrização de RNAs previsoras de séries temporais univariadas não lineares, em particular do tipo *Multilayer Perceptron*, aqui usada, através da implantação da metodologia PBCA.

2.2. Análise de séries temporais

Séries temporais são conjuntos de dados observados, medidos ou calculados em intervalos de tempo uniforme- ou não uniformemente espaçados. Modernamente são consideradas como sinais digitais e sua análise, em termos gerais, é equivalente ao Processamento Estatístico de Sinais Digitais para os profissionais das áreas de Engenharia (DSP, *Digital Signal Processing)*, ao contrário daqueles que trabalham em administração e finanças, entre outros. A análise de sinais fornece informações úteis para o controle de processos, previsão, detecção prematura de ocorrência de falhas, reconhecimento de padrões, descrição de fenômenos e processos, entre outros (TANGIRALA 2015). Os tipos de análises dividem-se em duas grandes famílias: aquelas que pertencem ao domínio temporal (e.g., tendência e correlação) e aquelas que pertencem ao domínio das frequências (ditas espectrais); ambas envolvem modelagem, estimativa, filtragem, previsão a curto, médio e longo prazos, e são usadas em pré-processamento de dados e em casos multivariados, lineares ou não. São muitas as áreas onde ocorrem séries temporais, como, por exemplo, gerenciamento de operações, (macro)econometria, finanças e gerenciamento de riscos, marketing, controle de processos industriais, demografia, geociências, meteorologia, bioestatística, medicina, física, astronomia e negócios (MONTGOMERY, JENNINGS e KULAHCI, 2008; TANGIRALA, 2015; CHATFIELD, 1996).

A literatura está repleta de obras a respeito da análise de sinais e a academia e a indústria têm à sua disposição vários pacotes computacionais, comerciais e/ou gratuitos. Existe farto material especializado sobre os métodos e técnicas, disponível em bibliotecas e bases digitais públicas. O Quadro 2.2, no entanto, traz um cenário mínimo das principais técnicas para fins de ilustração.

Quadro 2.2 – Pequena lista de técnicas em análise de séries temporais relevantes para este trabalho

Nome da Abordagem/Técnica	Sigla	Nome original
Ferramentas em diagnóstico		
Função de Autocorrelação	ACF	*Auto Correlation Function*
Função de Autocorrelação Parcial	PACF	*Partial Auto Correlation*
Análise espectral univariada		
Transformada de Fourier	FT	*Fourier Transform*
Espectro de potência	PS	*Power Spectrum*
Modelagem univariada		
Média Móvel	MA	*Moving Average*
Autoregressiva	AR	*Auto-Regressive*
Média Móvel Auto Regressiva	ARMA	*Auto-Regressive Moving Average*
Média Móvel Auto Regressiva Integrada (abordagem Box-Jenkins)	ARIMA	*Auto-Regressive Integrated Moving Average*
Box-Jenkins Sazonal	SARIMA	*Seasonal Box-Jenkins*
Modelagem Não Linear		
Autoregressiva de Limiar	TAR	*Threshold Autoregressive*
Autoregressiva de Transição Suave	STAR	*Smooth Transition Autoregressive*

Fonte: Pesquisa da autora.

Chatfield (1996) classifica as séries temporais em contínuas ou discretas em função da forma como os dados são colhidos ao longo do tempo. As primeiras são compostas de medidas que podem ser especificadas em qualquer valor de tempo *t*, onde $t \in R_+$; as segundas, a amostragem é feita em momentos específicos que são geralmente igualmente espaçados entre si, ou seja, o intervalo de tempo *n* é discreto ou seja, $n \in N$. Esse autor chama a atenção para o fato de que as medidas devem ser analisadas na ordem cronológica em que foram tomadas e que geralmente não são independentes entre si; acrescenta, ainda, que uma série é denominada determinística se seu comportamento temporal pode ser previsto com exatidão, mas a grande maioria dos sinais (na natureza e no dia a dia industrial, empresarial e/ou de mercado

financeiro[1]) é estocástica e depende do conhecimento da história das séries a fim de permitir alguma previsão, atrelada a uma função distribuição de probabilidades. Tangirala (2015) ressalta que os sinais são aleatórios, gerados por processos aleatórios e que não se consegue identificar uma causa física responsável pela variação desses sinais; tais variações terminam por ser consideradas como efeito de causas desconhecidas quando da fase de modelagem. Esse autor define, ainda, identificação de sistemas na modelagem em engenharia como sendo a busca da relação entre pares de variáveis de entrada e saída. Montgomery, Jennings e Kulahci (2008), ao tratarem sobre modelos para previsão, citam que é comum assumir-se que qualquer série temporal pode ser representada pela soma de duas componentes: uma determinística e outra estocástica; a primeira é modelada em função do tempo e a segunda é representada por uma função de ruído aleatório, a qual resume em si todo o comportamento não determinístico da série. Séries temporais podem requerer modelos univariados ou multivariados em função do número de variáveis mensuráveis, e cada tipo conta com um conjunto de estratégias e técnicas de análise e modelagem.

Ditam as boas práticas, porém, que antes de qualquer tentativa de modelagem, é necessário conduzir um estudo exploratório da série temporal de interesse, no qual características tais como periodicidade, sazonalidade, tendência, pontos discordantes (*outliers*), regressividade e autocorrelação são identificadas. Este reconhecimento e classificação da série fornecem informações básicas para a escolha das técnicas de análise e modelagem subsequentes com o objetivo de fornecer previsões para auxílio a tomadas de decisões estratégicas. Algumas ferramentas de diagnóstico desse estudo inicial estão enumeradas no Quadro 2.2, cujos conjuntos caracterizam-se por serem usados no domínio do tempo (Espaço de Estados) ou no domínio de fase ou frequências (Análise Espectral).

O modelo ARIMA é bastante usado, tendo sido proposto por Box e Jenkins (1976) com base nos modelos autorregressivo (AR, onde a série é descrita por seus valores passados regredidos e pelo ruído aleatório), de médias móveis (MA, que explora a estrutura de autocorrelação dos resíduos da previsão do período atual com aqueles ocorridos em períodos anteriores) e da combinação de ambos (ARMA). ARIMA contempla casos não estacionários (estatísticas básicas, como média, variância e covariância não são constantes) e sazonais (SARIMA). Estas séries geralmente não variam em termos de valor fixo da média devido à presença de autocorrelação. Quando a série é não estacionária, é utilizada a componente de integração I

[1] Parênteses da autora.

em ARIMA. Quaisquer aplicações a séries não estacionárias devem ser precedidas de técnicas como diferenciação e transformação a fim de estabilizá-las.

Como verificar a não linearidade de uma série temporal? Uma série puramente estocástica é linear se puder ser representada da seguinte forma:

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i \ \ a_{t-1} \tag{2.1}$$

Onde $\{x_t \ | t = 1, \dots, T\}$ é a série temporal observada, μ é constante, ψ_i são números reais dos quais $\psi_0 = 1$, e $\{a_t\}$ é uma sequência de variáveis aleatórias do tipo IID com função distribuição contínua definida e $E(a_t) = 0$. É comum assumir-se que a_t é Gaussiana e, nesse caso, $Var(a_t) = \sigma_a^2$.

Séries não lineares são modeladas geralmente assumindo-se a existência de regimes estocásticos diferentes ao longo do tempo, ou seja, heterocedastiscidade. Os regimes são caracterizados por médias, variâncias e até mesmo autocorrelações diferentes; entendendo-se volatilidade como a variabilidade ou variância instantânea, a magnitude das autocorrelações é inversamente proporcional à magnitude da volatilidade. Balestrassi *et al.* (2009) dividem em duas classes os modelos que levam em consideração mudanças de regime: aqueles determinados pelas variáveis observadas (p.ex., TAR, STAR, etc) e aqueles que são determinados por variáveis não observadas (p. ex., cadeias de Markov−MARKOV, 1906). Esses autores, no entanto, tratam apenas de séries que podem ser modeladas por AR para as quais os parâmetros autorregressivos dependem do regime. Seja qual for a modelagem, é imperativa a análise dos resíduos. Li (2004) apresenta um apanhado de testes de diagnóstico para séries temporais em função do tipo de série e de modelo de interesse.

Nesta pesquisa, esta análise é feita dentro da metodologia PBCA e SARIMA é usado para construir a base de comparação (*benchmarking*) para os resultados das RNAs.

2.3. Redes Neurais Artificiais

2.3.1. Visão geral

Redes Neurais Artificiais são estruturas computacionais programáveis usadas na solução de problemas especialmente não lineares multifatoriais e multiníveis. Foram propostas em meados da primeira metade do século XX, no esteio do artigo pioneiro de McCulloch e Pitts (1943), cuja ideia era a de neurônios binários funcionando segundo uma lógica de limiar com o objetivo de processar informações de maneira similar ao que se acredita ocorrer na estrutura cerebral dos seres vivos. Nascia, assim, a neurocomputação; Barreto (2002) apresenta a linha do tempo das pesquisas em Inteligências Artificiais com bom grau de detalhamento. Analogamente, as RNAs são formadas por unidades denominadas neurônios que se distribuem em subestruturas identificadas como camadas internas e de borda; as internas sendo conhecidas como *hidden layers* e as outras, como de entrada e de saída (*input* e *output*, respectivamente). Os neurônios atuam em sinergia com o propósito de processar informações, comunicando-se entre si segundo protocolo pré-estabelecido e com eficiências diferenciadas entre si. RNAs são capazes de aprender, de serem treinadas, de realizarem tarefas diversas como classificar, propor novos grupos, sugerir temas, produtos e serviços de interesse a usuários, tomar decisões sem supervisão, fornecer previsões, e muitas outras aplicações, principalmente nas áreas de Reconhecimento de Padrões, Análise de Sinais, Robótica e Sistemas Especialistas (TAYLOR e SMITH 2006) e séries temporais, como, por exemplo, marés (TIROZZI, PUCA *et al*., 2006)− v. Quadro 2.4 para a visão crítica de RNAs. A função das multicamadas é didaticamente explicada por Nielsen (2015)[2] .

A popularidade das RNAs extrapola os limites da academia e da manufatura; Spyros Makridakis criou uma competição internacional de solução de problemas reais complexos − M-*Forecasting Competitions* – com o objetivo de avaliar e comparar a exatidão de diferentes métodos de previsão; a terceira e última edição ocorreu em 2000 (MAKRIDAKIS e HIBON, 2000) e em todas houve competidores que usaram RNAs.

[2] "Nesta rede, a primeira coluna de perceptrons – a qual chamaremos de primeira camada de perceptrons– está tomando três decisões muito simples, pesando a evidência da entrada. E quanto aos perceptrons na segunda camada? Cada um desses perceptrons está tomando uma decisão ponderando os resultados da primeira camada de tomada de decisão. Desta forma, um perceptron na segunda camada pode tomar uma decisão em um nível mais complexo e mais abstrato do que os perceptrons da primeira camada. E as decisões ainda mais complexas podem ser feitas pelo perceptron na terceira camada. Desta forma, uma rede de perceptrons de várias camadas pode envolver-se em uma tomada de decisão sofisticada." (Tradução livre da autora)

Yeung *et al*. (2010) traçam um paralelo entre a anatomia de um sistema nervoso central vivo e um sistema artificial: soma e função de ativação, dendrito e entrada, axônio e saída, sinapse e peso (v. Figura 2.5). Makridakis *et al*. (1998) traçam um paralelo entre o jargão de DOE e o das RNAs: modelo e rede, parâmetros e pesos, estimativa de parâmetros e treinamento da rede. Segundo estes autores, a especificação de uma RNA requer quatro características: arquitetura, funções de ativação, função-custo e um algoritmo de treinamento. A arquitetura é o número de camadas, unidades e da topologia das conexões. As funções de ativação descrevem como cada unidade combina as informações de entrada e saída. A função-custo mede a exatidão da previsão; neste trabalho serão usadas as estatísticas Erro Porcentual Absoluto Médio e Erro Porcentual Absoluto Mediano (MAPE e MdAPE, respectivamente). O aprendizado de conceitos e o treinamento para dominá-los envolvem a otimização dos valores dos parâmetros a fim de minimizar a função custo. RNAs com várias camadas são conhecidas como *Multilayer Perceptrons* (MLP) e pertencem a uma grande família denominada *feedforward* (unidirecional ou direta). Uma vez que discorrer sobre RNAs detalhadamente foge ao escopo desta proposta, recomenda-se a leitura do trabalho de Nielsen (2015).

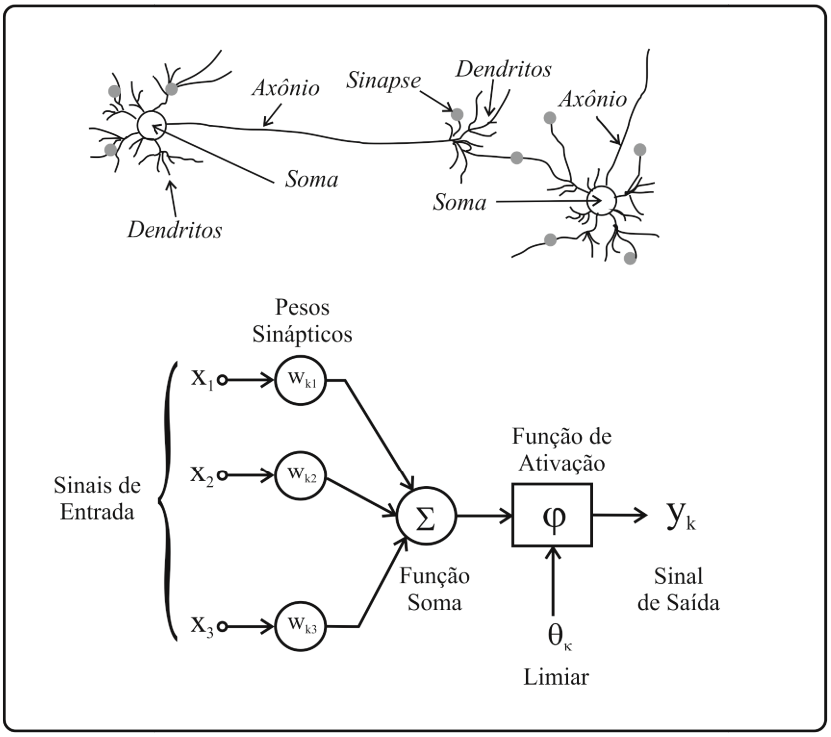

Figura 2.5 - Paralelo entre neuroanatomia de seres vivos e neurotopologia de RNAs.
Adaptado da Figura 1.1 de Yeung *et al.* (2010)

Atualmente, RNAs são classificadas como um tipo de Inteligência Computacional, a qual é um ramo dentro do campo de Inteligências Artificiais; os algoritmos, inspirados na natureza, são heurísticos, possuindo elementos de aprendizagem e adaptação (ŠTĚPNIČKA *et al*., 2013). Vale citar que, além da capacidade de realizar aproximações de quaisquer funções não lineares, RNAs podem oferecer resultados satisfatórios mesmo com uma topologia simples, além disso, são tolerantes a falhas, aprendem rapidamente, podem auto adaptar-se e calculam em paralelo (LI e YANG, 2014).

O tipo de treinamento também é uma escolha do profissional. Nesta pesquisa será usado o algoritmo de retropropagação dos erros (*backpropagation*) juntamente com algum método de otimização do comportamento das sinapses. Este algoritmo é largamente usado na aprendizagem das RNAs devido à sua rapidez (NIELSEN, 2015) e bom desempenho em problemas de classificação e reconhecimento de padrões (SOUZA *et al*., 2011), e capacidade de generalização (HAYKIN, 1999). Há que se tomar o cuidado, no entanto, de evitar o superaprendizado (*overfitting*). O superaprendizado caracteriza-se pelo ajuste de toda e qualquer característica da amostra de treinamento, o que é indesejável porque, embora o desempenho da RNA pareça cada vez melhor, a capacidade de generalização piora cada vez mais. Pode muito bem ocorrer quando o treinamento é realizado por muito tempo, quando a amostra de treinamento é muito pequena e/ou quando a mesma amostra é apresentada tanto no treinamento como na validação. Redes bastante complexas – entenda-se, grandes − também propiciam o superaprendizado (ANGUS, 1991), um problema que se agrava nas RNAs de aprendizado profundo (*deep learning*; BONTEMPI *et al.* 2013) e que novas técnicas tentam de contornar (SRIVASTAVA *et al.*, 2014).

2.3.2. Redes Neurais Artificiais para problemas de Séries Temporais

No campo de **previsões de séries temporais,** RNAs têm-se reveladas satisfatoriamente competentes. Chiang *et al.* (1996) relatam que elas são superiores aos modelos de regressão quando há escassez de dados como, por exemplo, quando fundos mútuos são recém lançados. Tem se tornado comum explorar sua associação a outras inteligências artificiais e abordagens estatísticas, *e.g.* regras linguísticas *fuzzy*, algoritmos genéticos, Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM), caos, análise de sensibilidade e Decomposição por *Wavelets* (por exemplo, LIU, 2017; DOUCOURE *et al.*, 2016; ŠTĚPNIČKA *et al*., 2013;

LECUN *et al.*, 2015; VAHIDNIA *et al*., 2010). Vários desses trabalhos também foram elencados por Zhang (2001) e Balestrassi *et al.* (2009), entre outros.

Independentemente do número de entradas que alimentam uma RNA neste tipo de problema, a saída é única, ou seja, apenas um neurônio fornece a informação desejada. Assim sendo, as séries que serão usadas nesta pesquisa são séries não lineares univariadas.

Uma RNA é caracterizada por vários fatores, o que faz com que os profissionais busquem reduzir a dimensão do problema de interesse através da seleção das variáveis e das interações mais significativas. A metodologia aqui proposta não carece desse recurso, uma vez que o DOE será realizado apenas uma vez.

Algumas das variáveis necessárias à caracterização das RNAs a serem construídas nesta pesquisa estão elencadas no Quadro 2.3, que é uma adaptação da Tabela 2 de Balestrassi *et al.* (2009), a qual contém parte dos parâmetros supracitados; esses autores discorrem detalhadamente sobre cada um deles.

Findo o processo, as RNAs assim construídas serão aplicadas a casos reais. O Quadro 2.4 contém a Visão Crítica do método.

No caso específico de aplicação em séries temporais reais, tratamentos tais como estabilização, remoção de tendências e sazonalidade não são necessários, pois que as RNAs são especialistas em captação de características não lineares. No entanto, como em qualquer análise de dados experimentais, deve-se proceder à avaliação preliminar através de ferramentas tais como histogramas (verificação de distribuição multimodal), estatísticas gerais como média, variância, mediana, existência de *outliers*, periodograma e diagrama de densidade espectral (para verificação da existência de sazonalidade) e inspeção visual da distribuição dos dados; eventualmente, estas informações são úteis às modelagens de séries temporais tradicionais (ARIMA, etc.). A metodologia PBCA aqui empregada indicará a existência ou não de interação entre as diversas variáveis e quão forte é essa correlação através de regressão linear, análise ANOVA e análise de sensibilidade.

É importante ter em mente que esta tese modela a arquitetura das RNAs e não as séries temporais em si.

Quadro 2.3 – Visão geral dos parâmetros relacionados ao uso de RNAs

Variável	**Sigla**	**Parâmetro no Statística®**
Tipo de análise: personalizada	-	*Custom Network Designer*
Arquitetura da RNA	archit	*ANN architecture*
Número de unidades de entrada,	-	*Number of input units*
Número de unidades de saída	-	*Number of output units*
Número de camadas internas,	HL	*Number of hidden layers*
Número de unidades neuronais por camada interna,	UL	*Number of units per hidden layer*
Tipo de função de saída da regressão,	OF	*Regression output function*
Tipo de problema	PT	*Network type*
Número de instâncias a serem previstas,	-	*Steps ahead to predict*
Número de instâncias	SP	*Steps used to make the prediction*
Método de amostragem,	SM	*Sampling method*
Algoritmo de treinamento da fase 1	P1	*Phase 1 training algorithm*
Algoritmo de treinamento da fase 2	P2	*Phase 2 training algorithm*
Número de vezes que o treinamento passa por toda a amostra tanto na fase 1 como na fase 2	Ep	*Epoch*
Modelos de inicialização e de ativação de cada neurônio,	IM	*Initialization method*
Condições de parada dos cálculos (um objetivo é alcançar o alvo, que é um erro, e o outro é alcançar uma situação na qual a mudança do erro é menor que um determinado valor),	SC	*Stopping conditions: training and selection target error*
Condição para evitar o superajuste (*overfitting*) e começar a perder precisão e qualidade,	ET	*Minimum improvement in error in training/selection*
Valor das mínimas melhorias nos erros de treinamento,	EE	*Minimum improvement in training and selection ,error for a window of a certain number of epochs*
Critério de adormecimento de neurônios quase inativos,	PU	*Prune inputs/units with small fan-out wights (prune input variables and prune hidden units as well)*
Critério de descarte de variáveis de entrada menos significativas e pesos das sinapses.	PI	*Prune inputs with low sensitivity after training*
Taxa de aprendizagem	LR	*Learning rate*
Regularização do decaimento dos pesos na fase 1	W1	*Weight decay regularization in Phase 1*
Regularização Weigend do decaimento dos pesos na fase 2	W2	*Weigend decay regularization in Phase 2*
Ajuste da taxa de aprendizagem e momento a cada época no caso de *Back Propagation*	LM	*Adjust learning rate and momentum each epoch*
Reordenar o uso das medidas a cada nova época no caso de *Back Propagation*	SO	*Shuffle presentation order of cases each epoch*
Adicionar ruído gaussiano no caso de *Back Propagation*	AG	*Add Gaussian noise*
Reamostragem: número de amostras	NR	*Resampling: Number of Samples*
Formação de *ensemble*	FE	*Form an ensemble*
No caso de amostragem aleatória, optar por reamostrar todos os conjuntos, *i. e.*, treinamento, seleção e validação	-	*Resample all subsets*
Tamanho de cada conjunto no caso de reamostragem aleatória é empregado o critério 2:1:1, respectivamente	-	*Subset sizes for Training, Selection and Test.*

Fonte: Adaptado de Balestrassi *et al.* (2009)

Quadro 2.4 - Visão crítica da aplicação de RNAs na análise de séries temporais

Características	Uso	Vantagens	Limitações	Justificativa da escolha
Estrutura básica: neurônios, sinapses, camadas internas, de entrada e de saída Permite medida estatística de desempenho Todas as unidades de processamento compartilham as informações. Redundância generalizada Modelo básico: Rummelhart (camadas)	Previsão Classificação Associação de dados Conceitualização de dados Filtragem Proposição de novos conjuntos Aconselhamento Tomada de decisões	Multi-propósito Preveem de forma não linear Lidam bem com problemas complexos demais para métodos "convencionais" Substituem abordagens que necessitam modelos matemáticos Podem ser associadas a outras abordagens (fuzzy, etc.) Sistema tolerante a falhas pelo paralelismo dos neurônios	"caixas pretas" Requerem grande número de dados O aumento do número de camadas internas implica em maior tempo de processamento Capacidade computacional Perigo de superaprendizado	Capacidade de generalização Não requer modelo conceitual do problema de interesse Lida bem com grande número de variáveis de entrada e saída Lida bem com situações não lineares Tem sido usada com sucesso na previsão temporal Pode ser associada a outras técnicas para melhores resultados

2.4. Delineamento de Experimentos para simulação

O uso de DOEs é largamente difundido na manufatura, mas as aplicações fora dos processos industriais, do desenvolvimento e aperfeiçoamento de produtos e serviços, longe do chão de fábrica e dos laboratórios, estas demoraram para ser descobertas, conquistarem profissionais de outras áreas e firmarem-se como estratégia vantajosa − Kleijnen *et al.* (2005) levantam hipóteses sobre essa dicotomia.

Segundo Balestrassi *et al.* (2009), o treinamento de uma RNA (que envolve a experimentação com diversos parâmetros de entrada em algoritmos computacionais inteligentes, a execução seguida desses algoritmos e a análise dos resultados fornecidos a cada instância) pode ser encarado como um estudo simulatório do problema de RNAs. Kleijnen *et al.* (2005) apresentam a correspondência entre os termos do jargão de simulação e os de DOE; um aspecto importante nessa relação é a manutenção do conceito de réplicas Distribuídas Idêntica e Independentemente (*Independently Identically Distributed* – IID) dos experimentos "clássicos". Em simulação, isso é conseguido através do fato das simulações estocásticas

usarem sequências de números pseudoaleatórios (*Pseudo-Random Numbers* – PRNs) que não se sobrepõem para cada cenário. Caso contrário, os resultados de uma simulação são imutáveis. Em DOE, os testes de hipóteses acerca dos efeitos dos fatores e suas interações, o ajuste de modelos matemáticos, as estatísticas, a análise dos resíduos, tudo visa a obtenção de um modelo ótimo; não há, na simulação, um objetivo paralelo e sim diferente: a obtenção de um modelo robusto, de uma tomada de decisão a melhor possível no quesito de satisfazer a todas as exigências do problema da melhor e mais balanceada forma possível, e de uma política ou protocolo robusto. De acordo com Giesbrecht e Gumpertz (2004), um experimento computacional permite lidar com problemas excepcionalmente grandes e complexos, o que os torna difíceis de apreender. Simulações têm por objetivo calcular uma quantidade, e se esta for função de variáveis aleatórias, o objetivo é calcular sua esperança matemática.

O uso de DOEs em simulação traz consigo preocupações e cuidados adicionais, às vezes pouco óbvios de serem traduzidos. É necessário avaliar a escolha das hipóteses, das variáveis de entrada, as condições de contorno, as variáveis de ruído, a própria escolha das métricas para avaliar os DOEs e comparar os resultados entre as réplicas, e confrontá-los com os oriundos de metodologias "tradicionais", que constituem *benchmarks* (v. o Capítulo 0).

Como já adiantado na Introdução, uma forma de buscar o conjunto de parâmetros de uma RNA é por tentativa e erro; uma de suas deficiências é que dificilmente chega-se ao final da busca porque o número de possíveis combinações de todos os parâmetros é muito grande. Isso faz com que a busca seja longa, cara e sem garantia de identificação de uma solução ótima.

Outra forma, sem dúvida superior, é a aplicação sucessiva de diversos DOEs em subconjuntos de parâmetros. Começa-se realizando a seleção das variáveis mais significativas e estabelecendo-se as relações de causa e efeito. Após, seleciona-se um novo arranjo, mais indicado para as variáveis acima e segue-se com o processo até que um determinado resultado seja confirmado pelo DOE seguinte, o que indica que as variáveis (e seus níveis) que mais influenciam o resultado do processo, foram identificadas. Deste ponto em diante, a RNA está pronta para ser usada na reprodução da série e fornecer previsões. Estes são os passos seguidos por Balestrassi *et al.* (2009). Esta metodologia não é plenamente satisfatória porque ainda remete à força bruta, requer a dedicação de um profissional especialista em escolher DOEs, em aplicá-los e compreender os resultados para realizar sua análise.

2.4.1. Ações preparatórias para um DOE bem-sucedido: uma estratégia

Em 1935, Sir Ronald Fisher publicou a primeira edição do livro intitulado "*The Design of Experiments*", cujos princípios fundamentais para a experimentação computacional são os mesmos da experimentação na indústria e na empresa, por exemplo. Antes de dar início a qualquer experimento, há uma série de passos a serem seguidos a fim de garantir a sua boa condução. Outro ponto importante a ser levado em conta é que mesmo um experimento bem planejado depende grandemente da forma como os dados são coletados. Abaixo estão elencadas algumas das principais ações da fase de pré-experimentação (COLEMAN e MONTGOMERY, 1993; MONTGOMERY, 2009; JIJU, 2012) , já adaptadas a este trabalho:

- Reconhecimento e verbalização do problema → leva à clara visão do seu universo
- Escolha dos fatores, seus níveis e intervalos de valores → leva ao conjunto de parâmetros
- Classificação dos fatores em controláveis ou não → leva a decisões sobre tratamento
- Listagem das interações de interesse → leva à clara compreensão das relações
- Seleção das variáveis de resposta → em séries temporais, uma só: estatística de qualidade
- Escolha do arranjo experimental → a metodologia PBCA se baseia em *Plackett_Burman*
- Condução do experimento em si → é uma simulação e leva à mais robusta RNA
- Análise estatística dos dados → leva à confiabilidade do modelo
- Conclusões e recomendações → comparação com *benchmarks* e avaliação por *stakeholders* levam à validação, permitindo a aplicação a casos ainda não estudados.

Latgé e Simon (2007) apresentam a estratégia sob a forma de perguntas características da área clínica, mas que são válidas fora dela também, tais como "Quantas unidades devem ser colhidas para o experimento de forma que o mesmo seja generalizável e tenha poder suficiente?". Estes autores, ademais, abordam o problema de evitar "falsos positivos" (Erros do Tipo I[4]), decorrentes de vários fatores de pressão, sejam reais, do meio ambiente acadêmico ou empresarial, inconscientes, psicológicos, financeiros, etc. Exemplos corriqueiros são: a pressão por publicar resultados "bons" rapidamente, e a busca inescrupulosa ou simplória pelo P-*value* < 0,05, entre outros. O número de graus de liberdade

[4] Erros do Tipo I ocorrem quando o a hipótese nula é erroneamente rejeitada. Em contraposição, Erros do Tipo II ocorrem quando, erroneamente, a hipótese nula não é rejeitada ("falso negativo").Exemplo de Erro do Tipo I: o alarme de incêndio dispara quando não há incêndio algum. Exemplo de Erro do Tipo II: o alarme não dispara e o incêndio está ocorrendo.

e a proposta de análise devem estar bem declarados no começo dos trabalhos. A documentação da metodologia experimental deve ser clara e completa a fim de permitir a replicação dos resultados.

Maiores detalhes sobre DOEs encontram-se no Capítulo 3.

3. Delineamento de experimentos

3.1. Visão geral

O uso de arranjos fatoriais remonta a meados do século XIX, quando John Bennet Lawes fundou o que mais tarde viria a ser a Estação Experimental de Rothamsted, onde, com a ajuda de Joseph Henry Gilbert, experimentou em fertilizantes e alimentos para animais (Cressie 2015). Ronald A. Fisher publicou dois livros pioneiros, "*The Arrangement of Field Experiments*" em 1926 e "*The Design of Experiments*" em 1935, onde introduziu uma metodologia para a concepção de experimentos. Grande parte de seu trabalho foi dedicada às aplicações de métodos estatísticos em questões agrícolas (Jayabal 2010) - veja também Fisher e Wishart (1930) e Fisher, R. A. Sir (1974).

Foram propostos por Plackett e Burman em 1946 com base em trabalhos muito anteriores, tais como os de R. C. Bose , Sir R. A. Fisher, K. Kishen, K. R. Nair, R. E. A. C. Payley e W. L. Stevens, entre outros. Uma leitura cuidadosa do trabalho original do trabalho de Plackett e Burman (1946) leva ao profundo conhecimento de como as matrizes dos arranjos são construídas, facilitando subsequentes adaptações a casos de interesse e de como proceder ao rebatimento (*fold-over*).

O Planejamento de Experimentos requer o uso de arranjos fatoriais, que são matrizes bidimensionais, numéricas ou não, que contém informação estratégica suficiente para a realização de experimentos de forma estatisticamente adequada dentro do universo do processo ou fenômeno a ser modelado (veja FISHER, 1974, e COX e REID, 2000, para uma visão geral e crítica do que são experimentos, o que são estudos observacionais, quais são os requisitos para seu planejamento e execução, qual é sua relação com a análise dos resultados, qual é o papel da Estatística, o que são inferência e indução, entre outros temas introdutórios gerais). Jiju (2012) apresenta o cenário da aplicação do DOE no âmbito industrial, com suas dificuldades, necessidades e vantagens; a **Erro! Fonte de referência não encontrada.** apresenta um resumo pictórico dessa apresentação:

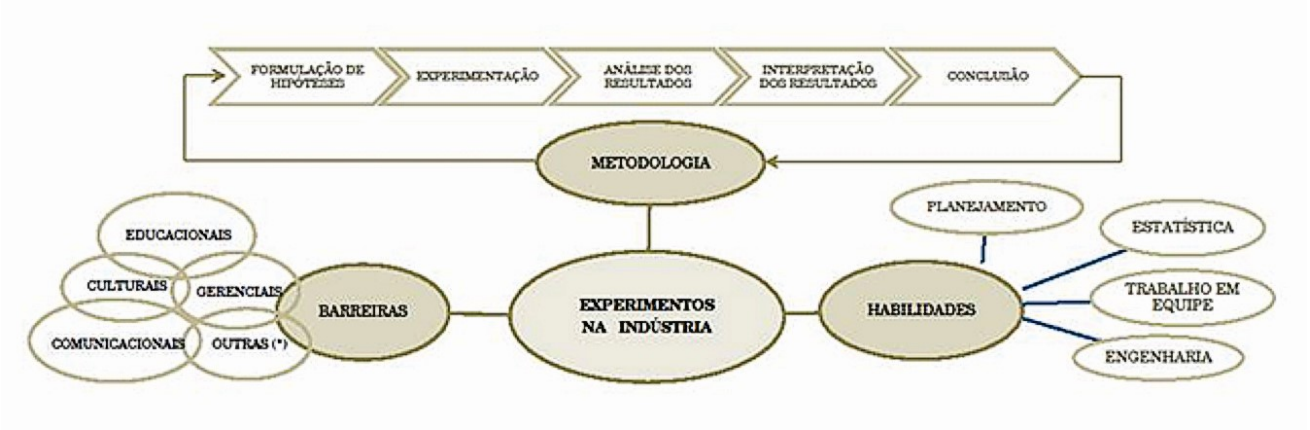

Figura 3. 1- Resumo pictórico do cenário da experimentação na indústria de acordo com a descrição dada por Jiju (2012). Desenho da autora.

(*) Outras barreiras: falta de experiência, falta de orientação de alguém com experiência, falta de tutoriais e de manuais e falta de estatísticos nos grupos de trabalho.

Este planejamento é usado basicamente em três situações: seleção de variáveis, otimização e/ou teste de robustez (LINDBERG, 2010). Os arranjos são construídos com base em regras bem estabelecidas − ver, por exemplo, Montgomery (2013), Plackett e Burman (1946) e Shen e Morris (2016) −, e geralmente contêm os dois valores extremos de cada variável do problema, seja quantitativa ou seja qualitativa, os quais podem estar codificados ou não. Existem, também, arranjos que contemplam casos em que as variáveis podem ter mais de dois níveis de valores (BOX e BEHNKEN, 1960) e arranjos mais complexos, que contemplam a existência de medidas adicionais com o fito de introduzir replicações (para permitir a estimativa de erros aleatórios), blocagem (para aumentar a precisão e reduzir a variabilidade do erro experimental) e/ou avaliação do comportamento do modelo nas regiões intermediárias do espaço experimental. Os experimentos, por sua vez, devem ser realizados de forma aleatória a fim de não se introduzir tendenciosidade ou viés (*bias*) inadvertidamente. Sivarao, Anand e Ammar (2010) apresentam uma visão geral sobre três abordagens compreendidas na metolodogia DOE: Modelagem por Superfície de Resposta com Arranjo Centralmente Composto, Método de Taguchi e Arranjos Fatoriais.

Em arranjos fatoriais de resolução III, os efeitos principais estão confundidos com interações de segunda ordem e estas, por sua vez, estão confundidas entre si. O valor da influência de um fator sobre as quantidades medidas é a soma de seu próprio efeito e do valor de cada interação com a qual é confundida. Se há alguma interação significativa entre os fatores, esse viés afeta

diretamente o valor do efeito calculado para esse fator que é confundido com uma interação significativa. Os arranjos fatoriais fracionários podem ser classificados de acordo com os padrões de viés que eles produzem. Esses padrões determinam o que se convencionou chamar de Resolução do arranjo. Há várias classes diferentes, mas os tipos III, IV e V são particularmente importantes (MONTGOMERY e RUNGER, 2011).

Os delineamentos de Resolução IV normalmente fornecem informações úteis sobre os principais efeitos e não há nenhum viés entre os efeitos principais ou entre os efeitos principais e as interações de dois fatores. Há, no entanto, um viés entre as interações de dois fatores, mas informações úteis também podem ser fornecidas sobre as mesmas (MONTGOMERY e RUNGER, 2011; MYERS, MONTGOMERY e ANDERSON-COOK, 2009). Um método para obter um delineamento de resolução IV para dois fatores a partir do correspondente III é aplicar a técnica de rebatimento (BOX, HUNTER e HUNTER, 1978) em (k-1) fatores (BOX e WILSON, 1951). A idéia é recriar cada combinação original com pequenas mudanças (HINKELMANN e KEMPTHORNE, 2012), ou seja, aumentar o arranjo original com uma fração ligeiramente modificada de si mesmo. Deve-se lembrar, no entanto, que a replicação fracionária implica enviezamento. Esta técnica tem a vantagem de eliminar o confundimento entre os principais fatores e as interações de segunda ordem (MONTGOMERY, 2013). De acordo com Li e Mee (2002), Li e Lin (2003) e Ye e Li (2003), o rebatimento é mais frequentemente usado em projetos ortogonais de resolução III para a obtenção de arranjos ortogonais de resolução IV.

Essa técnica também pode ser usada com efeitos principais ortogonais, como é o caso dos projetos PB (DIAMOND, 1995; MILLER e SITTER, 2001 e 2005), resultando em um arranjo IV de resolução de viés complexa de segunda ordem chamado super saturado porque todos os graus de liberdade disponíveis (ν) são utilizados na estimativa dos efeitos principais. O PB assim construído é uma boa alternativa para examinar todos os parâmetros simultaneamente (BERES e HAWKINS, 2001), já que o reconhecimento apropriado dos efeitos principais não é afetado pelas interações de segunda ordem (MILLER e SITTER, 2001).

Contudo, existe a possibilidade de haver pelo menos uma interação de ordem superior, digamos, de terceira ordem, que pode não ser devidamente levada em consideração (CAI *et al.*, 2014) devido ao Princípio da Escassez de Efeitos, que é frequentemente escolhido para ser aplicado por aqueles profissionais que não possuem tempo e/ou recursos para realizar experimentos fatoriais completos. Isso pode dificultar a interpretação apropriada dos efeitos principais (LIN, MILLER e SITTER, 2008).

Não obstante, Beres e Hawkins (2001) propuseram o uso de PB com rebatimento na análise de sensibilidade de modelos multiparamétricos e demonstraram que esse tipo de arranjo produz resultados consistentes além de ser capaz de identificar interações de segunda ordem. Análise de sensibilidade é uma técnica utilizada para compreender melhor o comportamento dos modelos simulados em função de alterações aplicadas aos fatores de interesse Newham (2002). Este tipo de análise é importante no processo de validação dos modelos (CONFALONIERI, BELLOCCHI e BREGAGLIO, 2010). Para uma visão crítica resumida do Delineamento de Experimentos, veja Quadro 3. 1.

Quadro 3. 1- Visão crítica resumida do Delineamento de Experimentos

Características	**Uso**	**Vantagens**	**Limitações**	**Justificativa da escolha**
Aleatoriedade Replicação Blocagem Insensibilidade a alguns fatores subjetivos e ao fator humano em geral Rigor Método pré-produção e pré-serviços Útil em melhoria de processos e produtos	Modelagem Seleção de variáveis significativas (*screening*) Identificação sistemática de interações entre as variáveis independentes Quando aliado a outras técnicas estatísticas: Análise de sensibilidade Otimização de modelos Ajuste fino dos limites experimentais	Permite o estudo de problemas não lineares Permite variação simultânea de todos os fatores Controle sobre as variáveis significativas e as irrelevantes Exploração de todos o espaço experimental Economia de recursos Evidencia relação causa-efeito	Dependendo do arranjo, interações significativas de segunda ordem e maiores podem não ser identificadas. Pode haver falta de recursos para total exploração do conceito Impossibilidade de generalização Deixa a desejar em previsões com modelos não lineares	Útil como auxílio à decisão As limitações não afetam significativamente a aplicação desejada aos casos aqui selecionados

3.1.1. Arranjos fatoriais fracionados

Se fôr possível supor que os efeitos das interações de ordem superior são desprezíveis, então pode-se obter informações sobre os efeitos principais e interações de menor ordem através de um planejamento fatorial fracionário. Isso é feito tipicamente nas fases exploratórias de um projeto, quando a seleção de fatores é necessária devido ao seu grande número (k), que dificulta o uso de um projeto fatorial completo (2^k). Vale ressaltar que o delineamento de um fatorial completo segue um modo binário lógico, mas o fatorial fracionário possui apenas $Q_t = (k - p)$ colunas binárias logicamente construídas, onde p é o número de geradores de

viés independentes; cada uma das colunas restantes recebe um gerador de viés que é definido pelo profissional. Em geral, o planejamento completo pode ser estudado com uma fração 2^{k-p}; este arranjo é chamado de arranjo Fatorial Fracionário 2^{k-p} e é classificado como geométrico. O número total de experimentos a serem executados neste caso é $N = (2^{k-p})$.

Nos cálculos dos efeitos principais, é usual estimar os coeficientes através de técnicas de regressão seguida de análise de ANOVA, que gera os erros residuais. Por outro lado, quando se está usando um fatorial fracionário, não há um grau de liberdade dedicado a esses erros residuais. Isso dificulta os cálculos dos testes *t* de Student e *F* de Snedecor e da correlação de Pearson. Uma alternativa para contornar este inconveniente é usar o método *PSE* de Lenth (*Pseudo Standard Error*; LENTH, 1989), que assume o Princípio da Escassez e que qualquer variação que cause os efeitos menores é puramente aleatória.

3.1.2. Arranjos fatoriais de Plackett-Burman

Os arranjos fracionários multivariados de PB para dois níveis requerem que o número de parâmetros de interesse seja dado por $k = (N - 1)$, onde N é o número total de experimentos (medidas ou observações), de tal forma que N deve ser divisível por 4. Nos casos em que $k \leq (N - 1)$, é possível planejar um experimento de resolução III, o qual se caracteriza pelo fato de que os efeitos principais estão confundidos com as interações de segunda ordem ao mesmo tempo em que algumas destas interações podem estar confundidas entre si (MYERS, MONTGOMERY e ANDERSON-COOK, 2009). O valor da influência de um fator nas quantidades medidas é a soma de seu próprio efeito e do valor de cada interação com a qual ele é confundido; Se houver qualquer interação significativa entre os fatores, este viés afeta diretamente o valor do efeito calculado para esse fator que é confundido com a interação significativa. Desconhecendo o valor real da importância de um determinado fator, o profissional pode ser levado a cometer um erro de Tipo II, assumindo que este não é importante, quando ele realmente é.

Plackett e Burman (1946) afirmam que "Se $N = 2^h(p^4 + 1) = 4K$, onde p é um número primo ímpar ou zero, uma matriz ortogonal pode ser construída com números 1 positivos e negativos. As de ordem 2^r são estruturalmente as mesmas que o arranjo fatorial completo de *r* fatores se forem obtidas por duplicações sucessivas. Estas serão chamadas de arranjos geométricos devido à sua estreita conexão com geometrias finitas." Montgomery (2013), no entanto, observa que alguns projetos de PB (por exemplo, N = 12, 20, 24, 28 e 36) são

classificados como não geométricos porque não podem ser representados por cubos e sua estrutura de confundimento é muito complexa. O leitor interessado encontra em Montgomery e Runger (2011) um bom texto introdutório sobre confundimento, incluindo exemplos didáticos numéricos.

Devido à sua capacidade de destacar os fatores principais, PB é descrito como um projeto de seleção sempre que o Princípio da Escassez é assumido. Isso tem sido motivo de debate frequente, mas Magallanes e Olivieri (2010) argumentam a favor da busca de estratégias que permitam a validação deste método devido à sua vantagem inerente de exigir um número menor de experimentos quando comparado, por exemplo, ao Fatorial Completo.

Todo PB é um arranjo fatorial de resolução III não regular devido a essa dependência entre os efeitos − em oposição ao Fatorial Completo (MONTGOMERY, 2013). Para alguns exemplos da aplicação de PB em outras áreas do conhecimento, o leitor pode referir-se a Chapin III *et al.* (2003, mudança da interação humana com o fogo), Faulkner *et al.*, (2003, lixiviação de vírus), Wu, Hall e Scatena (2007, impacto de mudanças recentes na cobertura do solo nos fluxos), Periago *et al.*, (2007, extração e quantificação do licopeno do tomate). Para uma visão crítica resumida de PB, veja o Quadro 3. 2.

3.1.3. Complexidade de confundimento

Em um arranjo PB que tenha sido construído pela técnica do rebatimento, cada coluna de efeitos principais é ortogonal a todas as outras e também àquelas das interações de segunda ordem. Ao mesmo tempo, pelo menos uma destas colunas de interações é ortogonal a cada outra com a qual compartilha um fator comum, e isto forma os blocos de interações. Colunas de pares de interações de segunda ordem que estejam completamente confundidas entre si possuem um coeficiente de confundimento unitário, ao passo que as outras, que não possuem interações em comum e, portanto, estão apenas parcialmente confundidas, possuem um coeficiente de correlação de cerca de $1/3$. Montgomery (2013) esclarece que cada arranjo fatorial de Resolução IV pode ser expresso pela notação 2_{IV}^{k-p}, onde p é o número de geradores de confundimento independentes e que pelo menos $2k$ experimentos devem ser realizados. Se for realizada exatamente essa quantidade, o arranjo é denominado Arranjo de Resolução IV Mínimo.

Quadro 3. 2- Visão crítica do arranjo fatorial de *Plackett_Burman*

Características	Uso	Vantagens	Limitações	Justificativa da escolha
Arranjo fatorial fracionado Multivariado Número de experimentos = número de fatores + 1 Número de experimentos deve ser múltiplo de 4 Colunas são ortogonais entre si Resolução III Pode ser geométrico ou não Assume-se modelos de primeira ordem, ou seja, detecta efeitos lineares Experimentos de dois níveis	Seleção de variáveis independentes Aplicação logo no início dos experimentos, quando pouco se conhece acerca do problema Quando é adequado desconsiderar interações Começa a valer a pena se houver mais de quatro variáveis	Boa ferramenta para seleção de variáveis Requer menos experimentos que, por exemplo, o Fatorial Completo Pode ser usado com número de variáveis alto Após seleção inicial, o mesmo arranjo e experimentos podem ser usados como fatorial completo com réplicas e permitir o estudo das interações de ordem dois significativas Se acrescido de pontos centrais, permite o estudo da curvatura, ou seja, efeitos de segunda ordem Forma econômica de detectar efeitos principais notáveis	Efeitos principais estão confundidos com as interações de segunda ordem Pode levar ao cometimento de Erros do Tipo II Não dedica um grau de liberdade para o cálculo do erro experimental Estrutura de confundimento que pode ser complexa Não verifica o efeito interfatores	Usa-se a técnica de rebatimento na nova metodologia (PBCA) de modo que as limitações deixam de sê-lo

A estrutura de confundimento das interações de segunda ordem varia em função do número de graus de liberdade e, neste sentido, PB com rebatimento tem a mesma estrutura de um arranjo fatorial fracionado. Este novo PB é expresso pela notação $2_{IV}^{((k+1)-p)}$. Pode-se criar um arranjo fatorial do tipo IV de duas maneiras: atendo-se ao arranjo não geométrico mínimo ou através de um arranjo geométrico expresso por $2_{IV}^{((k+1)-4)}$. Para fins de ilustração, o Quadro 3. 3- Exemplo de uma matriz PB obtida pela técnica do rebatimento mostra uma matriz PB de Resolução IV criada por rebatimento. A metade superior é constituída pelo arranjo PB original, de Resolução III para $N = 8$ e $k = 7$; a primeira linha é padrão deste arranjo, considerado geométrico. Note que esta primeira linha é aquela fornecida por Plackett

e Burman em 1946 (pág. 323). A metade inferior é resultado do rebatimento, na qual todas as linhas, exceto a última, são construídas com base na primeira através de deslocamento helicoidal, todos os elementos da última devem ser -1.

Quadro 3. 3- Exemplo de uma matriz PB obtida pela técnica do rebatimento

Matriz final *Plackett_Burman* de Resolução IV	Matriz *Plackett_Burman* original de Resolução III	+	+	+	-	+	-	-
		-	+	+	+	-	+	-
		-	-	+	+	+	-	+
		+	-	-	+	+	+	-
		-	+	-	-	+	+	+
		+	-	+	-	-	+	+
		+	+	-	+	-	-	+
		-	-	-	-	-	-	-
	Rebatimento	-	-	-	+	-	+	+
		+	-	-	-	+	-	+
		+	+	-	-	-	+	-
		-	+	+	-	-	-	+
		+	-	+	+	-	-	-
		-	+	-	+	+	-	-
		-	-	+	-	+	+	-
		+	+	+	+	+	+	+

Há pelo menos duas diferenças entre os dois arranjos acima: (i) como já adiantado, o coeficiente de confundimento das interações de segunda ordem é unitário apenas nos arranjos geométricos e (ii) o erro padrão usado no cálculo dos efeitos principais e das interações de segunda ordem no modelo de regressão é de 0.24σ para os arranjos não geométricos e de 0.18σ para os geométricos. A falta de geometria causa correlação entre os coeficientes do modelo, aumentando o erro padrão e perda da precisão na estimativa dos parâmetros; este é o balanço custo-benefício que deve ser feito quando os recursos experimentais são escassos. É interessante ter-se em mente, também, que um arranjo fatorial de Resolução IV apresenta aberração mínima, ou seja, produz o menor número possível de pares de interações (da ordem em questão) confundidas entre si, neste caso, a segunda ordem (Fries e Hunter, 1980).

3.2 Plackett-Burman *Sensitivity Analysis* - PBSA

Há dois tipos de análise de sensibilidade: local e global (MCRAE, TILDEN e SEINFELD, 1982; SALTELLI, CHAN e SCOTT, 2009), este último sendo o mais amplamente utilizado (XU e GERTNER, 2008b). Os métodos de análise global, por sua vez, podem ser

subdivididos em três grandes classes: aquelas baseadas na seleção, regressão e variância (CONFALONIERI, BELLOCCHI e BREGAGLIO 2010). Há uma certa preferência pela regressão quando se trata de explorar a relação entre duas ou mais variáveis (MONTGOMERY e RUNGER, 2011) e PB é baseado na regressão. Por outro lado, Xu e Gertner (2007) argumentam que o Teste de Sensibilidade de Amplitude de Fourier - FAST - é mais popular e, nessa linha, Gevrey, Dimopoulos e Lek (2006) recomendam o uso de derivadas parciais modificadas na análise de todas as possíveis combinações de variáveis em pares, em vez de modelos baseados em RNAs. Para um estudo comparativo das diferentes análises de sensibilidade, veja Confalonieri, Bellocchi e Bragaglio (2010).

Beres e Hawkins (2001) descreveram um método de análise de sensibilidade de modelos para ser usada juntamente com arranjos fatoriais PB completos e rebatidos – PBSA. Embora haja supersaturação, o método fornece informação sobre as interações de segunda ordem ao mesmo tempo em que não requer um grande número de experimentos. A nova matriz destina-se à análise de modelos multivariados. Assim sendo, além de indicar as variáveis de maior influência, também permite o estudo do impacto que suas variações exercem sobre a resposta.

A quantidade de cenários necessários para um dado conjunto de experimentos é aproximadamente o dobro do número de parâmetros considerados. Eles também notaram que o arranjo de PB não era amplamente conhecido por ecologistas e alegaram que, na época de sua pesquisa, não havia obras ecológicas ou biológicas sobre PB na literatura científica. É interessante que a situação tenha mudado completamente ao longo dos anos, como pode ser visto nos trabalhos de, por exemplo, Banerjee, Sarkar e Banerjee (2016), Hassan *et al.* (2016), El Ati-Hellal, Hellal e Hedhili (2014) e El Aty, Wehaidy e Mostafa (2014), entre outros, e o grande número de periódicos nas áreas correlatas e que existem.

Para discussões a respeito deste método, veja Chapin III *et al*. (2003), Faulkner *et al.* (2003), Dowd (2005), Scheurer (2006), Romero *et al.* (2007), WU, Hall e Scatena (2007), Xu e Gertner (2007), Xu e Gertner (2008a) Xu e Gertner (2008b), Xu e Gertner (2011) e Confalonieri i(2010). Esta metodologia serviu de base para aquela usada neste projeto, denominada PBCA.

3.3 . Análise de Correlação em Plackett-Burman - PBCA

3.3.1. Análise do coeficiente de correlação entre os sinais de resíduos

Smith (2003) define sinal como sendo a descrição de como um parâmetro está relacionado a outro. Quando ambos os fatores (x, y) podem assumir quaisquer valores continuamente, o sinal é classificado como contínuo. Quando este sinal é processado digitalmente, torna-se quantizado e o sinal resultante é discreto. Séries temporais apresentam características de sinais.

Se os resíduos são tratados como sinais discretos e sua amplitude é a variável dependente, então cada grupo de resíduos que são gerados no processo de regressão também são sinais discretos (veja também a Seção 2.1). É possível, então, analisar o coeficiente de correlação de cada sinal. No trabalho de Couto (2012) foi escolhido o coeficiente de correlação de Pearson com o propósito de quantificar as diferenças entre os sinais. Zhang *et al*. (2006) expressam o coeficiente de correlação para sinais ondulatórios da seguinte forma:

$$r = \left(\frac{\sum_n (R_n - \bar{R})(A_n - \bar{A})}{\sqrt{(\sum_n (R_n - \bar{R})^2)\ (\sum_n (A_n - \bar{A})^2)}} \right) \tag{3.1}$$

onde R_n é o sinal de referência, $\bar{R}$ é o valor médio de R_n, A_n é o sinal a ser comparado com R_n e $\bar{A}$ é o valor médio de A_n.

3.3.2. . O algoritmo PBCA

Quando fatores e interações significativos não são levados em conta na análise regressão (Erro do Tipo II), a variância aumenta e o gráfico de resíduos não é confiável. Este método permite que se detecte esta falha e, consequentemente, que o cálculo do modelo ótimo seja robusto. A manutenção de fatores e interações pouco significativas durante os cálculos pouco altera a variância, de sorte que qualquer redução no valor da mesma deve-se primariamente à presença daqueles termos significativos. Esta é a importância do método: o padrão com o qual os resíduos se distribuem está associado à ausência dos principais efeitos e interações de segunda ordem no modelo de regressão. Isto influi tanto na magnitude quanto na distribuição dos resíduos. O fluxograma da implantação da metodologia está apresentado na Figura 3.2. O Apêndice C contém o pseudocódigo para a implantação, passo a passo e com exemplos, da PBCA em três fases. O Apêndice D traz o código do cálculo do arranjo experimental e da criação das matrizes para os cálculos das regressões e análise de sinais em Scilab®, e o

Apêndice E contém o código para o cálculo das regressões, correlações e medidas de erro em VBA/Excel®.

As três grandes etapas da implantação da PBCA, para qualquer problema, são: (i) domínio do DOE de PB, cuja função é entregar as matrizes para as próximas etapas de cálculo apenas computacional, sem mais experimentação, (ii) emprego de toda a PBCA apenas para os fatores (aqui denominada Fase 1) e (iii) emprego de toda a PBCA para as interações de segunda ordem e presença todos os fatores simultaneamente (aqui denominada Fase 2).

Ao cabo dessas três etapas, os fatores mais significativos e as interações de hierarquias fortes e fracas terão sido identificados. Como resultado, o melhor modelo que representar o fenômeno de interesse terá sido fornecido.

3.3.3. . Graus de liberdade na PBCA

Pelo menos um grau de liberdade deve ser dedicado ao cálculo do erro residual. Por exemplo, se um processo tem 11 fatores e escolhe-se realizar 12 experimentos, após o rebatimento haverá 24 experimentos e $\nu=23$. Neste caso, os graus de liberdade estão distribuídos da seguinte forma: 11 para os fatores, 1 para a blocagem, 1 para o erro residual e 10 para as interações de segunda ordem.

Três passos são seguidos: (i) usar todos os fatores e calcular o P-*value* > α, onde α é o máximo nível aceitável para o risco de se rejeitar a hipótese nula quando deveria ocorrer o oposto (Erro do Tipo I); (ii) transferir o grau de liberdade da blocagem para as interações de segunda ordem da seguinte forma: 11 para os fatores, 1 para o erro residual, 11 para as interações de segunda ordem e notar que $\nu=23$ ainda – notar que isto permite que o PB adotado não apenas forneça o erro residual como também o valor de P para cada coeficiente da regressão como as séries de resíduos para cada regressão estudada; e (iii) considerar as séries de resíduos como sendo sinais e analisar suas características individualmente e par a par de forma a identificar os efeitos principais e interações significativas.

3.3.4. O impacto de interações de ordem mais alta

A metodologia PBCA requer que todas as interações de ordem igual e maior que três não sejam significativas a fim de que a detecção dos efeitos principais e das interações de segunda ordem não seja severamente afetada. Miller and Sitter (2001) descrevem o impacto da

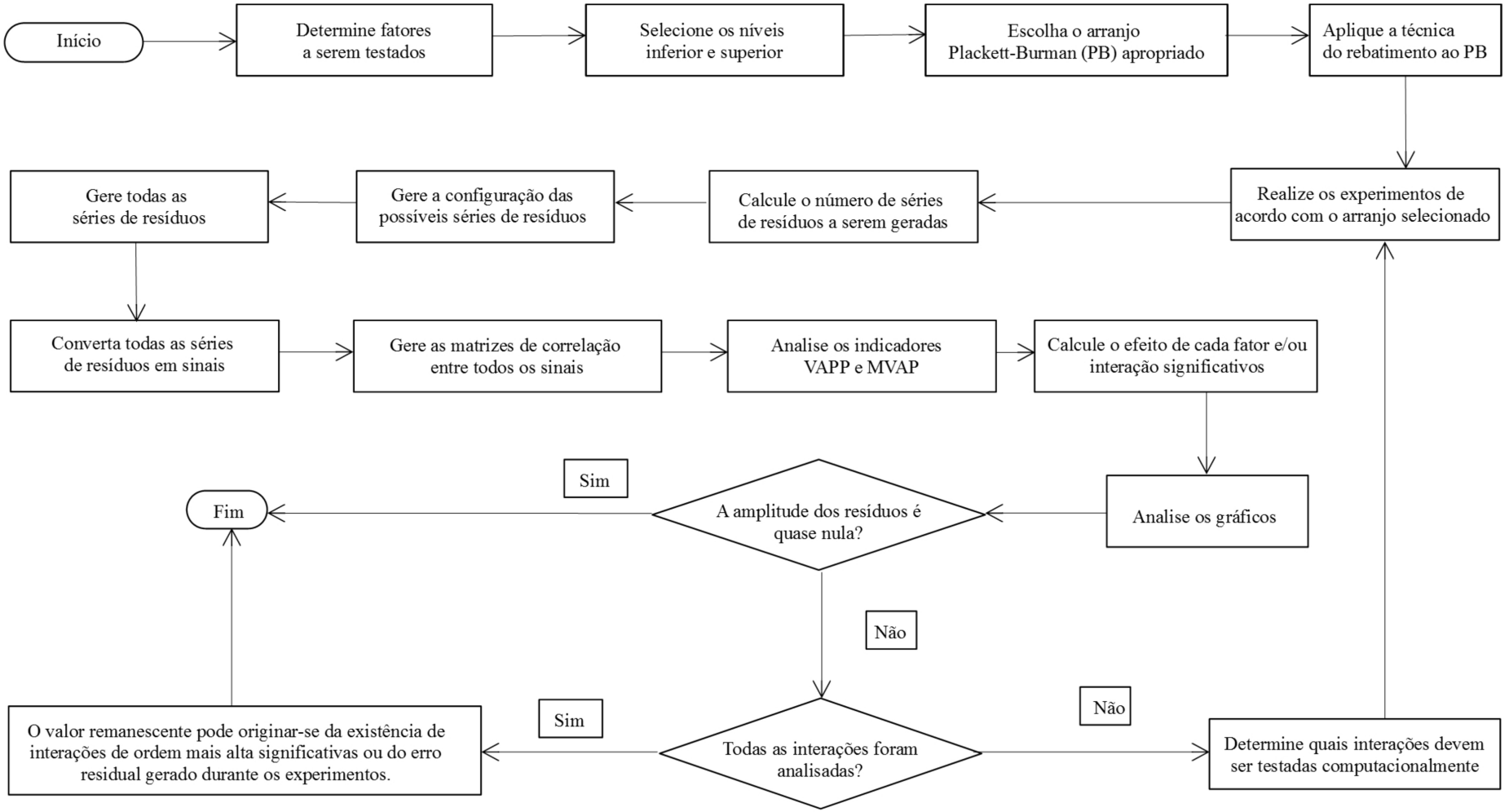

Figura 3.2 - Diagrama de blocos da metodologia PBCA. Adaptado de Couto (2012).

52

presença de interações de altas ordens sobre a dificuldade de identificação dos efeitos principais no caso de interações de ordem ímpar e das interações de segunda ordem no caso de interações de ordem par.

Durante o processo de identificação do PB correto e do número de experimentos, deve-se tomar o cuidado de verificar a existência de ortogonalidade a fim de compreender a estrutura de confundimento. Somente a seguir é que se realiza o rebatimento.

Cada linha da matriz representa um modelo de cenário específico e cada coluna está associada a um fator. Se o número de fatores é menor que o número de colunas, isto é, $k < (N-1)$, descartam-se as colunas excedentes. A seguir, pode-se proceder às medidas para cada experimento, respeitando-se as combinações de valores máximos e mínimos das variáveis em ordem aleatória. Os resíduos advindos da diferença entre os valores medidos e os ajustados a cada regressão compõem as séries, cujos valores não devem ser padronizados. Notar que o número de series de fatores, $NRESI_{fatores} = (2^k - 1)$, não é o mesmo que o de experimentos (N) porque não há resíduos quando todos os fatores estão ausentes $(RESI^0_{fatores})$; cada série possui tantos elementos quantos forem os fatores.

Quando se trata dos resíduos das interações, são dois os casos: (a) arranjo geométrico: a estrutura de confundimento determina o número de blocos de interações de segunda ordem e o número de séries de resíduos $NRESI^{geom}_{bloco} = \left(2^{número\ de\ blocos} - 1\right)$; neste momento, e novamente, não há resíduos sem a presença de interações; (b) arranjo não geométrico: $NRESI^{nãogeom}_{blocos} = (2^{\nu} - 1)$, onde ν é o número de graus de liberdade que podem ser usados para estudar as interações de segunda ordem, não existindo $RESI^0_{blocos}$. A notação usada define $RESI^1_{fatores}$ como sendo a série que contém apenas o primeiro fator, $RESI^2_{fatores}$ apenas o segundo, $RESI^3_{fatores}$, apenas o primeiro e o segundo e assim por diante, de acordo com a matriz PB em questão.

É interessante notar que:

(a) $RESI^0_{blocos} = RESI^{(2^k-1)}_{fatores}$, e

(b) Todos os fatores estão presentes em toda e qualquer série de resíduos das interações de segunda ordem. Se, por conveniência, decide-se numerar as séries relativas às interações em sequência às dos fatores, então o primeiro bloco é identificado como $RESI^{NRESI_{fatores}}_{blocos}$.

Seguindo com o mesmo exemplo acima, no caso de um arranjo geométrico com $k = 6$, há 16 cenários possíveis, 15 interações de segunda ordem, 63 séries de resíduos fatoriais, 7 blocos de interações e 127 séries de resíduos de blocos; a última série de resíduos fatoriais é $RESI_{fatores}^{63}$, a primeira série dos blocos é $RESI_{blocos}^{64}$ e a última é $RESI_{blocos}^{190}$. Já no caso de um arranjo fatorial completo não geométrico com $k = 11$, por exemplo, há 55 interações de segunda ordem, $N = 2^{11}$ experimentos, 2047 séries de resíduos fatoriais, e a última série de resíduos das interações denomina-se $RESI_{blocos}^{4094}$.

Após a geração das séries de resíduos dos fatores, é interessante gerar dois gráficos: (a) um *Two-Sample t* para os efeitos padronizados como teste de hipótese e (b) outro dos resíduos. O mesmo deve ser feito ao cabo da aplicação da PBCA, o que permite que se verifique sua eficiência.

3.3.5. . Conversão das séries de resíduos em sinais

Seguindo a lógica do Planejamento Fatorial Completo, as séries de resíduos contemplam todas as possíveis combinações de fatores e interações de segunda ordem. Os modelos que contêm variáveis significativas são aqueles que apresentam os menores resíduos. Se cada série é considerada como um sinal discreto, os valores residuais são agora valores de amplitude, a serem dispostos em gráficos *versus* o número de ordem seq**u**encial dos respectivos experimentos.

Este novo paradigma permite definir três indicadores nesta fase da análise: (1) o Valor da Amplitude Pico a Pico (VAPP), (2) o Valor de Amplitude Pico a Pico do Coeficiente de Correlação (VAPPCC) e (3) a Média Mais metade do Valor de Amplitude de Pico do Coeficiente de Correlação dos Sinais (MVAP). Estes indicadores são calculados tanto para os fatores como para as interações de segunda ordem. As explicações abaixo se aplicam a ambos os casos.

O primeiro indicador, VAPP, é a diferença entre os valores máximo e mínimo dos resíduos de cada experimento. Para permitir uma visão da adequação do modelo, deve ser disposto sequencialmente em um gráfico em função da ordem de realização dos experimentos.

O segundo indicador, VAPPCC, é calculado a partir da Matriz de Correlação dos Coeficientes, que expressa o grau de associação entre cada par de sinais. Trata-se de uma matriz triangular cuja diagonal é deixada vazia porque se refere à correlação de um sinal com ele mesmo. A diferença entre os valores máximo e mínimo da correlação para cada experimento define o VAPPCC, que também deve estar num gráfico em ordem de

experimento. Por construção, o último elemento é o único que se correlaciona com todos os outros.

O terceiro indicador, MVAP, é simplesmente a soma do valor médio de correlação de cada sinal e a metade do valor de pico da amplitude correspondente.

3.3.6. . Análise dos indicadores

Na busca dos fatores e blocos de interações de segunda ordem significativos (se houver) que descrevem um processo, deve-se ter em mente que: (a) quando a amplitude do sinal é pequena, VAPP não é tão sensível quanto MVAP, (b) a aceitação de um modelo baseada unicamente no VAPP leva ao cometimento de um Erro do Tipo II, (c) o critério final é escolher o modelo que produz o menor MVAP, a ser previamente escolhido dentre aqueles que produziram o menor VAPP, e (d) não é incomum encontrar valores residuais muito semelhantes entre si e/ou semelhantes aos menores encontrados (devido ao fato de que esses resíduos de ordem superior também contêm os fatores significativos).

A estratégia a ser seguida é a seguinte: (a) tratar os fatores e (blocos de) matrizes de interação separadamente, (b) classificar os modelos em ordem crescente do VAPP, (c) escolher analisar um certo número dos primeiros modelos, (d) comparar os MVAP mais baixos tanto para os fatores como para os blocos/interações pois eles contêm os descritores significativos do processo, (e) ter em mente que a consideração dos blocos/interações de segunda ordem diminui o valor do VAPP, o que é precisamente a inovação trazida pelo PBCA, (f) verificar que a não inclusão de fatores e/ou interações é prontamente indicada pela amplitude dos resíduos.

3.3.7. Ajuste fino do modelo

Uma vez que os melhores modelos candidatos foram identificados através das etapas acima mencionadas, uma nova análise de regressão deve ser feita. Desta vez, apenas com os fatores e interações significativas. Posteriormente, traçam-se gráficos destes resultados a fim de compará-los com aqueles obtidos antes da aplicação da metodologia PBCA.

No caso de um planejamento geométrico, deve-se usar apenas uma das interações pertencentes a um bloco significativo. No entanto, a existência de interações do tipo com-hereditariedade dentro dos blocos deve ser adequadamente analisada para não deixar de lado quaisquer fatores. Basear decisões sobre o princípio da escassez de efeitos pode levar o praticante a cometer um erro do Tipo II (BERES e HAWKINS, 2001). Uma maneira possível

de examinar corretamente os efeitos de cada interação é dessaturar o projeto através da adição de mais experimentos, de modo a eliminar o confundimento das interações de segunda ordem de interesse.

O coeficiente de correlação entre os sinais é um bom termômetro de quão bem o modelo representa o processo. Quanto mais fatores e interações significativos são levados em consideração, mais esse coeficiente aproxima-se de zero; consequentemente, quanto mais próximo de zero é o coeficiente, mais preciso é o modelo porque considera as interações importantes e não deve haver erro residualque mereça ser estudado com mais cuidado.

Examinando-se mais de perto esta propriedade dessa correlação, verifica-se que, à medida que mais e mais termos significativos são adicionados ao modelo, tanto o VAPP do sinal quanto a forma mudam em função dessa adição. A correlação entre os sinais cai abruptamente quando o sinal que contém os fatores significativos e interações é encontrado. Os sinais podem então ser divididos em dois grupos: aqueles que contêm fatores e interações que são significativos, e aqueles que não os possuem. Os membros dentro de cada grupo estão correlacionados uns com os outros, mas ambos os grupos não estão correlacionados entre si. Estas características são intrínsecas à metodologia PBCA (COUTO, 2012). Para uma visão crítica resumida de PBCA, veja o Quadro 3.4 - Visão crítica de PBCA

Quadro 3.4 - Visão crítica de PBCA

Características	**Uso**	**Vantagens**	**Limitações**	**Justificativa da escolha**
Arranjo PB com rebatimento Cálculo da influência de todos os fatores e suas interações de segunda ordem Cada série de resíduos da regressão é um sinal Usa técnicas de análise de sinais Analisa as correlações entre todos os sinais	Seleção de variáveis e interações de segunda ordem significativas Obtenção de modelos prontos para otimização	Identifica os fatores significativos Identifica as interações significativas, mesmo se os fatores que as compões não o sejam Identifica a falta de fatores e/ou interações no modelo Permite que qualquer planejamento de experimentos seja feito apenas uma vez	Resultados válidos apenas no universo do experimento	Pode ser aplicada tanto na academia como na indústria e comércio Não há necessidade de recorrer a outros experimentos ou métodos para identificação do modelo

4. Método de pesquisa

4.1. Considerações iniciais

O método de pesquisa desta tese dividiu-se em sete partes: (i) seleção dos pacotes computacionais necessários a cada etapa do trabalho, (ii) confecção dos algoritmos nas respectivas linguagens (iii) seguida da programação propriamente dita, (iv) seleção dos casos para implantação inédita da metodologia, (v) análise exploratória dos dados, (vi) aplicação da metodologia PBCA, (vii) análise crítica dos resultados e extração de conclusões. Com base nestas, e mesmo durante o trabalho, foram sendo divisadas as propostas de trabalhos futuros.

Os algoritmos e pseudocódigos confeccionados para esta tese estão apresentados nos Apêndices B, C e D e podem serem utilizados por quaisquer pesquisadores, desde que a fonte seja citada.

Durante a etapa de análise exploratória dos dados ficou clara a necessidade de estabelecer um protocolo de tratamento dos dados passo a passo, que terminou por poder ser aplicado a qualquer série temporal e que fica como contribuição desta tese àqueles que necessitam um guia para iniciar suas pesquisas com séries temporais. Este protocolo encontra-se na Seção 4.2.

Na etapa de aplicação da PBCA, devido às delimitações apresentadas na Seção 1.5, foi necessário padronizar as variáveis independentes, dependentes e as de ruído. Maiores detalhes são apresentados na Seção 4.4.

4.2. Protocolo exploratório das séries temporais

O uso de ambas as séries teve como primeiro passo a exploração de suas características estatísticas e comportamentais ao longo do tempo. Deve-se ressaltar, no entanto, que o estudo das séries em si não é o objetivo desta tese e ensejaria a inclusão de várias outras variáveis de entrada e de contorno. Este passo pode ser subdividido nas seguintes ações:

- Inspeção da amostra buscando valores faltantes.
- Inspeção visual do diagrama de dispersão da amostra de dados em função do tempo buscando apreender o comportamento histórico e identificar aquelas medidas claramente atípicas.

- Construção e análise do histograma para avaliação da simetria e achatamento, teste de Anderson-Darling (para, por exemplo, normalidade), média, desvio padrão e mediana.
- Se necessário, divisão da série empiricamente e recálculo da média e seu desvio para cada parte, geralmente em patamares.
- Identificação visual tentativa da existência de volatilidade (ou seja, heterocedasticidade: variância não constante) e estacionariedade da média.
- Levantamento de informações complementares extra série, mas no mesmo período estudado, as quais podem explicar determinadas características comportamentais da série. Estas variáveis podem classificadas como ruído (são incontroláveis) às vezes previsível. Exemplos destas ocorrências são: partidas de futebol em grandes estádios, feriados, férias coletivas e ocorrência do El Niño.
- Se necessário, substituição de valores atípicos ou faltantes pelo valor da mediana das medidas adjacentes no tempo.
- Análise espectral: construção de periodograma e de diagrama de densidade espectral para a identificação dos principais períodos que se combinam na composição da série temporal (picos com os valores mais altos da densidade espectral) e confirmados no periodograma correspondente.
- Inspeção visual do diagrama de densidade espectral para melhor escolha do número de medidas que as RNAs utilizam para o cálculo das previsões.
- Verificação da existência ou não de sazonalidade e tendenciosidade através do cálculo da ACF e PACF com *lag* definido pelos resultados do diagrama de densidade espectral.
- A correção por tendenciosidade é comumente realizada se ela for linear. Para tanto, experimentam-se os modelos linear, quadrático, exponencial e curvo S-*curve* − Pearl-Reed *logistic*− e observa-se a qualidade dos ajustes em função das estatísticas, por exemplo, MAPE, MAD e/ou MSD.
- Identificação da natureza do modelo: se aditiva ou multiplicativa, com base na distribuição sazonal dos resíduos de cada modelo. É necessário verificar se essa distribuição, em função dos valores ajustados, não apresenta agrupações ou tendências e, por fim, comparar o P-*value* de ambos os modelos. Novamente, as estatísticas MAPE, MAD e/ou MSD auxiliam nesta decisão. Se a sazonalidade não varia com o tempo ou com o nível do sinal, isto é indicativo de aditividade.
- Produção de série estacionária: cálculo das diferenças cuja ordem é baseada nos resultados da curva de densidade espectral seguido ou não por transformação dos dados via

diferenciação em *lags* e/ou aplicação do operador log_{10} . O cálculo de diferenças em *lags* dos valores logarítmicos também é um recurso válido de e eficiente na busca pela estacionariedade. A decisão de uso (ou não) do operador é feita em função, por exemplo, do aumento da amplitude ao longo do tempo na série alterada. Verifica-se esta premissa dividindo-se a série em partes e calculando as respectivas variâncias.

- Confirmação da estacionariedade: comparar a ACF e a PACF dos resíduos de cada operação acima e identificar aquela cuja distribuição esteja majoritariamente dentro dos limites de confiabilidade e não apresente tendências ou padrões.
- Especificação da classe de modelos de séries temporais a ser empregada (SARIMA).
- Especificação do modelo com base na análise de ACF e PACF.
- Modelagem da série e obtenção de estimativas dos parâmetros do modelo.
- Análise dos resíduos padronizados da modelagem. Sua distribuição deve ser estacionária, sem tendência e de variância constante.
- (opcional) Simulação por Monte Carlo para identificação do melhor modelo SARIMA e previsões.
- As previsões acima prestam-se à comparação (como *benchmarks*) com as previsões fornecidas pelas RNAs.

De posse de todas as informações acima, já é possível proceder à aplicação da PBCA na definição das RNAs de melhor desempenho na reprodução das séries e que tenham a capacidade de generalização suficiente para fornecer previsões estatisticamente confiáveis dentro dos limites experimentais de cada caso. A comparação com os resultados da análise "clássica" das séries temporais valida a metodologia aqui empregada.

4.3. Visão geral do algoritmo

A Figura 4.1 apresenta a algoritmo da metodologia em linhas gerais. Na fase de escolha do melhor arranjo, programou-se a montagem da matriz de PB no *software* Scilab® v.6.0.1 (64-bit). De posse da matriz de delineamento, dos parâmetros da RNA, seus níveis e da série temporal já pré-analisada e bem conhecida, cada experimento é realizado com o módulo de Redes Neurais do *software* Statistica® v7.0. Os resultados são armazenados para serem introduzidos no módulo PBCA, que vai então realizar a análise das correlações entre os sinais (séries de resíduos) através do *software* MSExcel®. Com o objetivo de verificar os ajustes,

lançou-se mão dos suplementos para Excel Crystal Ball ®, XLSTAT® e OLSRegression (BARRETO e HOWLAND, 2006), e do *software* Minitab®. Ao cabo do método PBCA, terá sido identificada a melhor arquitetura da RNA específica para aquela série temporal de interesse, podendo ser usada para previsões através, novamente, do *software* Statistica®. Uma vantagem adicional deste software é a possibilidade de transcrever qualquer algoritmo que seja executado com o mesmo em linguagem C de forma otimizada, o que permite que qualquer profissional e/ou estudante possa executar, ou mesmo adaptar, a RNA já treinada sem ter o Statistica® instalado em sua máquina.

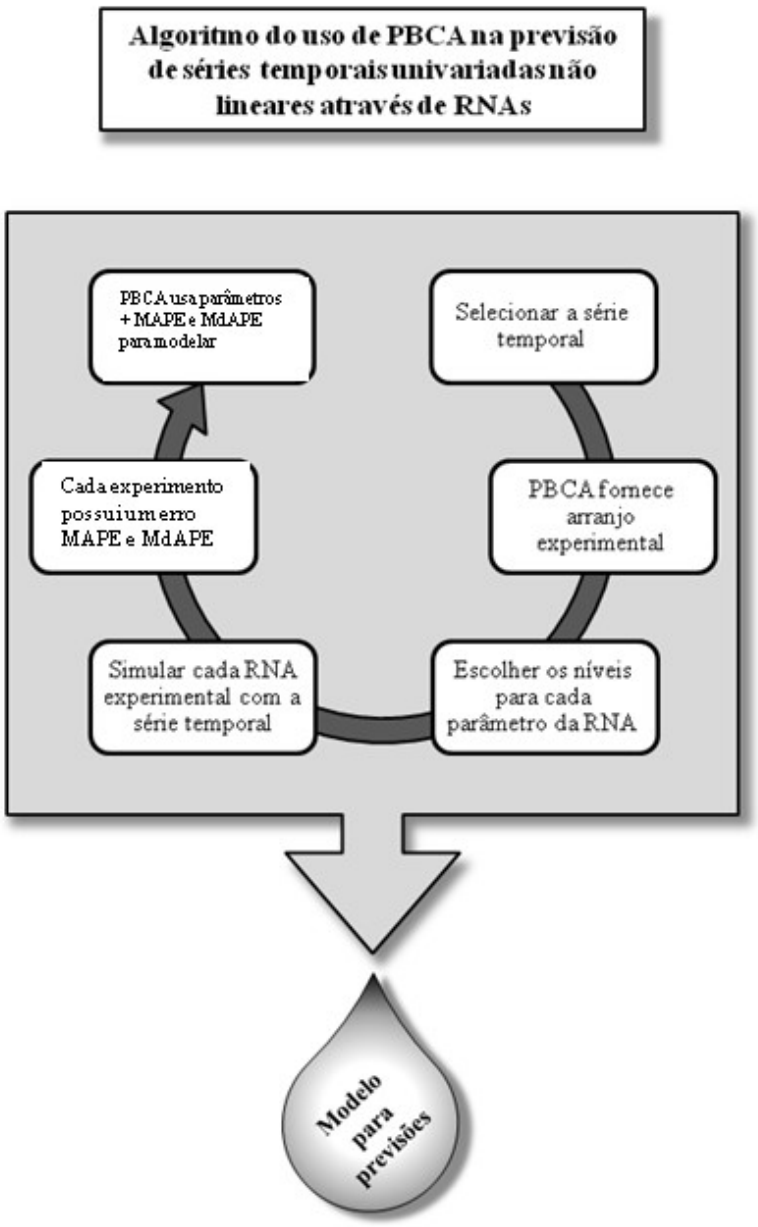

Figura 4.1 - Algoritmo da metodologia da pesquisa proposta.

A Figura 4.2 representa esta sequência para os parâmetros x_j, a série temporal $f_i(t)$ onde t é o tempo, e a variável dependente $y_i(t) = MAPE(i)$ou $MdAPE(i)$, complementando, assim, a Figura 4.1.

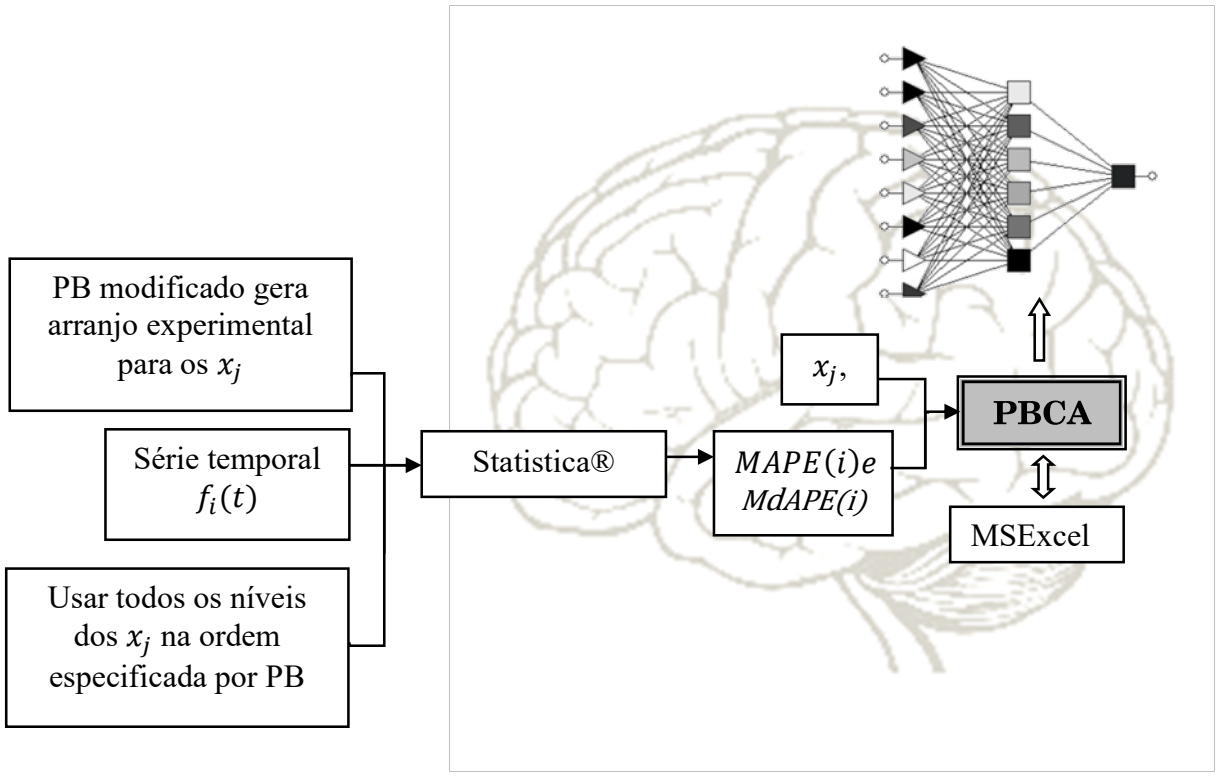

Figura 4.2 - Algoritmo relativo às RNAs

A estratégia de ataque aos dados reais a ser seguida neste projeto, com o objetivo de validar a metodologia servindo de *benchmark* é a seguinte: após conhecer e caracterizar os dados brutos, encontrar o melhor modelo SARIMA, para o que se assume que os erros são aleatórios, Independentes e Identicamente Distribuídos (IID) com média zero e variância constante. Ora, isto implica na independência dos erros, que, na prática, não é regra geral. Um exemplo de mercado é o fenômeno da alta volatilidade ou de variância que aumenta (MONTGOMERY *et al.* 2008) . Este comportamento não é detectado por modelagem ARIMA. É necessário, então, testar a hipótese dos erros IID através de um modelo que represente os erros como sendo não correlacionados, com média de ruído nula e variância variável ao longo do tempo. Neste modelo, chamado de Autorregressivo de Variância Variável Condicional, o ruído branco original de ARIMA é um termo a mais no modelo AR deste novo erro.

4.4. Parâmetros da PBCA comuns a ambos os casos

A ideia inicial era estudar k = 19 variáveis/fatores/parâmetros arquitetônicos das RNAs, em um arranjo PB modificado por *foldover* com 64 experimentos (geométrico) e 43 graus de liberdade, a fim de testar a eficiência da PBCA. Esta escolha resultaria em 524.287 resíduos de fatores e de pares de interações de segunda ordem. Uma tentativa resultou em falta de memória disponível para o cálculo das planilhas, tanto do notebook como da memória reservada pelos programas, no notebook e em nuvem, e alguns cálculos nem mesmo chegaram a iniciar.

Diante da impossibilidade de realização dos cálculos, procedeu-se a uma análise exploratória das diferentes combinações de números de variáveis e tipos de arranjos. Sabendo que o número máximo de resíduos dos fatores calculados a que se conseguiu chegar foi quase 40.000, e buscando trabalhar com um arranjo geométrico, pela menor complexidade das interações de segunda ordem, decidiu-se adotar k = 14 (16.383 resíduos) em vez de 15 (32.767 resíduos), configuração esta que permite estudar até 15 blocos de interações de segunda ordem. Esta escolha também não foi levada a cabo devido ao fato de algumas redes demorarem de mais de 13 horas a até mesmo 26 horas para finalizarem.

Foi necessário, novamente, rever o espaço experimental a fim da aplicação do método ser factível com os recursos humanos e de infraestrutura computacional à disposição. Decidiu-se pelos parâmetros contidos no Quadro 4.2 (considerados fixos) e Quadro 4.2 (valores máximos e mínimos para os parâmetros restantes)**.**

Quadro 4. 1 – Parâmetros com valores fixos da arquitetura das RNAs

	Valores fixos	No.
Análise	Planejamento personalizado de rede neural (*Custom Network Designer)*	1
PT	Série temporal univariada(*Univariate Time Series*)	2
archit	Rede multicamadas de Perceptrons (*Multiple Layer Perceptron* – MLP)	3
EP	Épocas (*Phase One AND two*) = 400	4
P1	Retropropagação (*Back Propagation*)	5
NR	Número de amostras pára reamostragem (*Number of Samples Resampling)* = 4	6
FE	Formar conjunto pára reamostragem (*Form Ensemble for Resampling)*	7
IM	Gaussiano de distribuição aleatória (*Random Gaussian*) N(0,1)	8
OF	Logístico (*Logistic)*	9
AG	Adicionar ruído gaussiano (*Add Gaussian Noise)* = 0,10	10
SP	Número de pontos usados pára a previsão (*Steps used to predict*) = (depende do caso)	11
EE	Melhoria mínima no erros em número em épocas = janela (*Minimum improvement in error #epochs = window*)	12
PU	Amortecer a entrada/neurÔnios internos com pequenos pesos decrescentes (*Prune Input /Units with small fan-out weights)*	13
PI	Amortecer a entrada pela diminuição da sensibilidade após treinamento (*Prune input with low sensitivity after training)*	14
LM	Ajustar taxa de aprendizado e momento a cada época (*Adjust learning rate and Momentun each epoch)*	15

Fonte: Adaptado de Balestrassi *et al.* (2009)

Quadro 4.2 – Variáveis consideradas para a matriz DOE. BFGS representa o método Quase-Newton, Lev-Marq, o método Levenberg-Marquardt e Crosval, o método de amostragem da Validação Cruzada.

No.	Sigla	Parâmetro	Valor mín.	Valor máx.
1	HL	Número de camadas internas (*Number of Hidden Layers)*	1	2
2	UL	Número de neurônios por camada interna (*Number of Units per Layer)*	1	2
3	P2	Algoritmo de treinamento da Fase 2 (*Phase 2 training algorithm)*	BFGS	Lev-Marq
4	LR	Taxa de aprendizado na Fase 1 - *Learning Rate (Phase One*)	0,01	0,1
5	SC	Condições de parada (*Stopping conditions)*	0	0,1
6	ET	Mínima melhora no erro de treinamento/seleção (*Minimum improvements error training/selection)*	-0,1	0
7	W1	Pesos/fatores de decaimento na Fase 1 com valores-padrão (*Phase one decay factors*)	Yes	No
8	W2	Pesos/fatores de decaimento na Fase 2 com valores-padrão (*Phase two decay factors with default values)*	Yes	No
9	SM	Método de amostragem (*Sampling Method)*	Crosval	Random

Fonte: Adaptado de Balestrassi *et al.* (2009)

Com o objetivo de fornecer uma visão geral do uso de PB e das RNAs neste trabalho aos leitores pouco familiarizados com estes temas, construiu-se o Quadro 4. 3:

Quadro 4. 3 – Classificação empírica dos parâmetros arquitetônicos das RNAs.

Classificação das variáveis relativas às RNAs		
I	Topológicas	HL, UL, a variável de entrada, a variada de saída
II	De comportamento neuronal	LR, W1, W2 e LM
III	Metodológicas	P2, PU, PI, SO e SM
IV	De controle	SC, ET e EE

A Tabela 4. 1 apresenta o arranjo experimental final. Esta configuração de arranjo não é geométrica, portanto, há 36 pares de interações de segunda ordem, com confusão de efeitos de 1/3 por par. As linhas numeradas 12 e 24 obedecem ao padrão de PB com *foldover*: a linha 12 é constituída apenas por -1 e a 24, por +1.

Tabela 4. 1– Arranjo experimental do Delineamento de Experimentos PB com *foldover*

Fatores:	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	**Ordem aleatória**
No. experim.	**HL**	**UL**	**P2**	**LR**	**SC**	**ET**	**W1**	**W2**	**SM**	
1	-1	-1	-1	1	1	1	-1	1	1	9
2	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**12**
3	1	1	1	-1	-1	-1	1	-1	-1	21
4	-1	1	1	1	-1	1	1	-1	1	7
5	-1	1	-1	1	1	1	-1	-1	-1	13
6	-1	-1	-1	1	-1	-1	1	-1	1	18
7	1	1	-1	1	-1	-1	-1	1	1	2
8	-1	1	-1	-1	-1	1	1	1	-1	11
9	1	-1	1	-1	-1	-1	1	1	1	1
10	1	1	-1	1	1	-1	1	-1	-1	5
11	1	-1	1	1	1	-1	-1	-1	1	23
12	-1	-1	1	-1	1	1	1	-1	-1	14
13	1	1	1	-1	1	1	-1	1	-1	6
14	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**24**
15	1	-1	-1	-1	1	1	1	-1	1	10
16	**1**	**-1**	**-1**	**1**	**-1**	**1**	**1**	**1**	**-1**	**15**
17	-1	-1	1	-1	-1	1	-1	1	1	17
18	-1	-1	1	1	1	-1	1	1	-1	8
19	-1	1	-1	-1	1	-1	1	1	1	16
20	-1	1	1	1	-1	-1	-1	1	-1	22
21	1	-1	-1	-1	1	-1	-1	1	-1	19
22	1	-1	1	1	-1	1	-1	-1	-1	4
23	-1	1	1	-1	1	-1	-1	-1	1	3
24	1	1	-1	-1	-1	1	-1	-1	1	20

A escolha dos fatores fixos e dos valores extremos das variáveis consideradas na PBCA foi feita com base (i) no trabalho de Balestrassi *et al.* (2009) e (ii) empiricamente, através de vários experimentos exploratórios para cada série temporal deste trabalho.

Em resumo, as condições dos experimentos são: número de fatores (k) = 9, número de pares de interação = 36, número de experimentos (N) = 24, número de graus de liberdade (ν) = 13, número de resíduos dos fatores = 511, número de resíduos das interações de ordem dois = 8.191.

Ocorre, no entanto, que o número de graus de liberdade restringe a análise dos pares de forma que uma matriz de 36 colunas por 8.191 linhas não comporta todas as combinações de pares de interações de segunda ordem para as regressões lineares, tendo por nulos todos os elementos a partir da coluna 14. Isto impossibilita o estudo da influência de 22 pares. Não há porquê aumentar o número de linhas artificialmente porque esta é uma limitação do arranjo experimental e qualquer pesquisador deve conviver com essa restrição. A fim de resolver este

impasse, decidiu-se pela estratégia descrita mais abaixo. Na busca de situações semelhantes na literatura recente, verificou-se que, em outro campo da Engenharia de Produção, o da manufatura, de Oliveira (2018) realizou extenso levantamento das publicações sobre a Metodologia da Superfície de Resposta e verificou que a grande maioria dos trabalhos emprega apenas três ou quatro variáveis e raramente cinco ou seis; em Engenharia Elétrica, no estudo do custo tarifário de energia elétrica, Ribeiro Júnior *et al.* (2013) empregaram oito; em Biotecnologia, nos últimos dois anos, o número de variáveis chega a dezesseis (ZHAO *et al*., 2017). Isto reforça o fato de que outros recursos computacionais, de IA e de software (Java, R e Python, por exemplo) deverão ser empregados nas extensões deste trabalho.

Realizaram-se alguns testes aplicando-se a PBCA apenas aos fatores ("Matrizes 1 e 2"), para ambas as séries e para ambas as medidas de qualidade. Os resultados OPD/MAPE, Duke8/MAPE e Duke8/MdAPE apontaram para o mesmo conjunto de fatores significativos: HL, UL, SC, W1 e W2. A combinação OPD/MdAPE indicou os fatores HL, P2, LR, SC, ET, W1, W2 e SM. Por simplicidade e por repetição do resultado, escolheu-se o primeiro conjunto como um primeiro passo na análise. Sempre se teve presente que esta não é uma solução definitiva para o estudo das interações e dos confundimentos; outras combinações devem ser utilizadas, ou seja, interações entre estes fatores significativos e fatores não significativos (interações de hereditariedade fraca) e então das interações de fatores não significativos com os fatores principais para ver o confundimento. Estes cinco fatores produzem 10 pares de interação, o que é comportado pelos graus de liberdade.

As novas matrizes para a parte da PBCA relativa aos pares de interação, pela própria construção, terminam então com 17 colunas. Ao ser executada, porém, a macro parou na linha 254 e emitiu mensagem de erro porque a função de regressão do MSExcel® aceita apenas 16 variáveis.

Assim sendo, a solução seguinte foi eliminar um dos cinco fatores e trabalhar com seis pares de interações. Como escolher qual eliminar? Cada série temporal foi tratada separadamente porque não há razão, *a priori*, de se supor que ambas são descritas pelo mesmo conjunto de fatores. Foi realizada a análise fatorial das duas séries temporais, com o cálculo ANOVA e regressão completa, com todos os fatores e as interações de ordem dois dos cinco candidatos e, com base nesses resultados, eliminou-se a quinta variável, UL, e a Matriz 2 PBCA ficou com 15 colunas (Tabelas Tabela 4. 2 e Tabela 4.3).

Tabela 4. 2 – Parte superior da Matriz1 usada no cálculo da PBCA referente às interações entre os quatro fatores escolhidos e, por exemplo, Y = MdAPE. A Matriz 1 possui 24 linhas/experimentos.

Fatores									Interações						Y
X1	X2	X3	X4	X5	X6	X7	X8	X9	X1X5	X1X7	X1X8	X5X7	X5X8	X7X8	MdAPE
-1	-1	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	26,51125
-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	23,99102
1	1	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	1	-1	24,48272
-1	1	1	1	-1	1	1	-1	1	1	-1	1	-1	1	-1	25,12224
-1	1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	24,03596

Tabela 4.3 – Parte superior da Matriz2 usada no cálculo da PBCA referente às interações entre os quatro fatores escolhidos. A Matriz 2 possui 63 linhas.

Todos os fatores presentes									Interações de apenas quatro deles					
X1	X2	X3	X4	X5	X6	X7	X8	X9	X1X5	X1X7	X1X8	X5X7	X5X8	X7X8
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
1	1	1	1	1	1	1	1	1	0	1	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1	1	0	0	1	0	0	0
1	1	1	1	1	1	1	1	1	1	0	1	0	0	0

Os cálculos todos foram realizados em dois notebooks:

- um com processador Intel® Core™2 Duo CPU T6600 @2,20 GHz, 4 GB RAM e Windows 10 Pro®
- e outro com processador Intel® Core™2 i7-7500U (2,7 GHz, Cache de 4 MB), 16 GB de memória RAM, DDR4, 2400 MHz, drive primário SSD de 128 GB e disco rígido de 1 TB, e Windows 10 Pro®.

Aqui cabe um parênteses com o seguinte exercício: um experimento desse tipo demora cerca de 60 segundos, no máximo, para ser realizado pelo Statistica® no primeiro notebook; um fatorial completo de apenas 2 níveis requer 2^{24} experimentos, portanto, $(2^{24} \times 60\,seg \div 3.600\,seg \div 24\,h \div 365\,dias) \cong 31{,}9\,anos$ de tempo de processamento! Sem a metodologia aqui implementada, é muito difícil e trabalhoso identificar a RNA adequada, mesmo através da aplicação de métodos e técnicas de redução de variáveis.

5. Implementação do método proposto em dois casos reais

A metodologia PBCA foi proposta por Couto (2012) através da aplicação a um modelo sintético criado especialmente para testar seu desempenho sob diferentes aspectos. A mesma foi implementada no estudo de duas séries temporais reais. A primeira, de número de horas noturnas úteis de um observatório astronômico profissional em terra e a segunda, de distribuição de carga de potência elétrica de uma distribuidora no Brasil.

Para introduzir e fundamentar estes temas, a seção 5.1 discorre sobre o conceito de qualidade em Astrofísica Observacional e a seção 5.2 apresenta alguns aspectos gerais do problema da distribuição de carga elétrica.

A seção 5.3 contém a redução dos dados do observatório e sua análise, e a seção 5.4 contém a redução dos dados de carga e sua análise. Na seção 5.5 é apresentada uma discussão dos resultados para ambos casos. As considerações finais deste capítulo encontram-se na seção 5.6.

5.1 Qualidade em Astrofísica observacional

5.1.1. Visão geral de sítios astronômicos

A Astronomia é uma ciência básica interdisciplinar. As pesquisas de ponta na área requerem infraestrutura de obras civis, de Tecnologia da Informação e de instrumental complexas e de alto custo. Os recursos humanos em determinados grupos de trabalho podem ser escassos devido ao alto nível de especialização e ao grande número de anos necessários para sua formação (“notório saber”).

As vertentes de pesquisa inicial e simplistamente divididas em “teórica” e “observacional” (experimental), deram lugar a métodos de trabalho combinados, que envolvem grupos com vários profissionais frequentemente de especialidades diferentes. Atualmente há grande interesse em Exobiologia/Astrobiologia (vida no Universo), Cosmologia e busca de planetas extrasolares, por exemplo. Os astrônomos trabalham em estreita colaboração com, por exemplo, químicos, engenheiros ópticos, engenheiros eletrônicos, engenheiros mecânicos e biólogos marinhos. Para viabilizar as pesquisas de interesse, comunidades internacionais criam parcerias e consórcios para suprir a demanda por construção de novos equipamentos, laboratórios e oficinas de alta precisão, observatórios, telescópios em terra e no espaço, entre outros. Surgem, assim, os astrônomos-administradores, os astrônomos-desenvolvedores de

software, os astrônomos-desenvolvedores de instrumentos, trabalhando ao lado de gestores administrativos, financeiros, de compras, etc. em projetos de grande porte e valor científico.

Assim sendo, a Astronomia, enquanto ciência básica, no seu desenvolvimento possibilita a oferta de produtos e serviços variados, empregos, divisas, troca de *know-how*, interação com a indústria, subprodutos que terminam por fazer parte do cotidiano do público em geral e garantia de Tecnologia e Inovação no país.

Neste momento, o Brasil conta com mais de 300 doutores contabilizados junto à Sociedade Astronômica Brasileira (SAB, www.sab-astro.org.br), possui assento no Comitê Assessor de Física e Astronomia (CA-FA) do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), participa de grandes consórcios internacionais para implantação de novos observatórios no exterior, possui sólida formação de novos pesquisadores e desenvolve instrumentos de vanguarda. Em solo brasileiro possui apenas um observatório aberto a todo e qualquer pesquisador de instituições brasileiras de ensino superior e pesquisa, caracterizando-se, pois, como sendo de uso nacional: o Observatório do Pico dos Dias (OPD, Brazópolis, MG) gerenciado pelo Laboratório Nacional de Astrofísica/MCTIC. A comunidade astronômica brasileira cresce a uma taxa aproximada de 10% ao ano e a grande maioria das propostas solicitando tempo de observação possui mérito científico.

Devido aos altos custos contábeis, financeiros, de tempo e recursos humanos envolvidos no projeto, implantação, operação, manutenção e melhorias dos observatórios, telescópios e instrumental periférico, espera-se que um sítio astronômico seja explorado por, no mínimo, 20 anos. Sua vida útil depende de fatores climáticos, humanos, políticos, tecnológicos e científicos. Este trabalho restringe-se aos aspectos técnico-científicos e climáticos.

O LNA mantém um histórico de horas úteis desde o começo das operações do OPD. Até o ano de 2001, considerou-se como hora útil aquela que permitiu a obtenção de dados com a qualidade requerida pelo pesquisador junto ao telescópio de 1,60 m. Depois, considerou-se também o uso dos outros telescópios do OPD, e a denominação passou a ser “horas adequadas para observação”. Em 01/09/2006, teve início o registro dos dados de uma estação meteorológica, colhidos a cada meia hora. A equipe de engenharia de desenvolvimento instrumental e o pessoal de manutenção também utilizam os telescópios, porém boa parte dos trabalhos não requer noites de qualidade astronômica. Contrariamente ao senso comum, imagens astronômicas de alta qualidade não são, necessariamente, aquelas obtidas em condições excepcionais de céu, com nitidez extrema. O objetivo último de uma imagem

tomada exatamente com o arranjo (*setup*) instrumental solicitado pelo pesquisador que recebeu tempo de telescópio é ser útil à pesquisa que ele pretende realizar.

Por muito tempo, a definição de qualidade foi "adequação ao uso" Juran (1974), onde o uso de um produto ou serviço era definido pelo cliente. Com o crescente emprego da abordagem estratégica do gerenciamento da qualidade, houve uma mudança de paradigma e qualidade passou a ser definida como "adequação aos propósitos" (JURAN e DE FEO, 2010). Strong e Lee (1997) apresentam 4 categorias de Qualidade dos Dados (QD): intrínseca, de acessibilidade, contextual e de representação; as definições das diferentes dimensões encontram-se no Quadro 4.4. Strong e Lee (1997) definem, ademais, Problema de QD como sendo "qualquer dificuldade encontrada em uma ou mais dimensões que fazem com que os dados sejam completa ou largamente inadequados para o uso". Neste trabalho, a QD caracteriza-se como intrínseca e contextual. O Quadro 4.4 traz o detalhamento das dimensões da QD segundo Pipino, Lee e Yang (2002).

Quadro 4.4 - Classificação da qualidade de dados

Categorias de QD	**Dimensões da QD**
Intrínseca	Exatidão, Objetividade, Credibilidade, Reputação
de Acessibilidade	Acessibilidade, Segurança de Acesso
Contextual	Relevância, possuir Valor Agregado, Completeza, Quantidade de dados e ser Oportuna
Representacional	Interpretabilidade, Facilidade de Compreensão, representação Concisa, representação Consistente

Fonte: Adaptado de Strong e Lee (1997)

Nesta nova acepção, o astrônomo recebe os dados obtidos com a configuração instrumental, condições meteorológicas e de céu solicitadas e com a qualidade apropriada para sua pesquisa. Este é o resultado da fase chamada de validação dos dados. A validação é um processo cuja entrada é constituída por fótons e cuja saída é um conjunto de dados científicos e de calibração cientificamente úteis e de qualidade (veja Figura 5.1).

É interessante notar que, embora as medidas sejam frequentemente repetidas em sequência ou ao longo de dias e/ou anos, cada observação astronômica é única. Trata-se de um ensaio destrutivo no sentido de que, apesar do *setup* instrumental ser o mesmo, as condições de céu,

o comportamento físico e dinâmico dos próprios objetos celestes e as decisões tomadas pelo observador no momento da coleta mudam.

Quadro 4.5 - Detalhamento das diversas dimensões da Qualidade de Dados

Dimensões da QD	Definições
Acessibilidade	Extensão dos dados disponíveis, ou fácil e rapidamente coletáveis
Completeza	Até onde dados não são faltantes e possuem fôlego e profundidade para a tarefa em questão
Compreensão	Até onde os dados são facilmente compreensíveis
Credibilidade	Até onde os dados são vistos como verdadeiros e críveis
Facilidade de manipulação	Até onde os dados são facilmente manipuláveis e aplicáveis a diferentes tarefas
Interpretabilidade	Até onde os dados estão nas linguagens apropriadas, símbolos, unidades e as definições são claras
Livre de erros	Até onde os dados estão corretos e são confiáveis
Objetividade	Até onde os dados não possuem viés, são livres de preconceitos e parcialidade
Oportunidade	Até onde os dados estão suficientemente atualizados para a tarefa em questão
Possuir valor agregado	Até onde os dados oferecem benefícios e vantagens com seu uso
Quantidade apropriada de dados	Extensão do volume de dados apropriados para a tarefa em questão
Relevância	Até onde os dados são aplicáveis e úteis para a tarefa em questão
Representação concisa	Até onde os dados podem ser representados compactamente
Representação consistente	Até onde os dados são apresentados no mesmo formato
Reputação	Até onde os dados são bem conceituados em termos de fonte ou conteúdo
Segurança	Até onde os dados são de acesso restrito de forma apropriada a fim de manter sua segurança

Fonte: adaptado de Pipino, Lee e Yang (2002)

Com o advento das câmeras digitais (*Charge-Coupled Device* na língua inglesa – "CCD"; no entanto, e somente neste trabalho, criou-se a sigla CCDE a fim de não causar confusão com aquela do Arranjo Centralmente Composto − CCD), os dados experimentais, ditos observacionais, são arquivos constituídos de uma parte binária (os dados em si) e outra, alfanumérica (metadata: cabeçalho com detalhes da medida). A parte binária é proporcional ao número de fótons coletados por cada elemento espacial do detector, expressa em contagens/pixel. Um CCDE é um detector de estado sólido e pode ter, por exemplo, 19 milhões de pixels; suas características principais são: estabilidade, linearidade e sensibilidade;

um pixel possui dimensões da ordem de uma dezena de mícrons. As próximas gerações de CCDEs produzirão mais de 1 TB de informação por minuto. Os dados são imagens bidimensionais em tons de cinza, guardadas individualmente ou em cubos com várias imagens "empilhadas". Mesmo com telescópios de grande porte (espelho principal de, pelo menos, 8 m de diâmetro), o sinal gerado deve ser amplificado (fator denominado ganho) a fim de gerar corrente elétrica mensurável. Para maiores detalhes acerca da física e da engenharia envolvidas na construção deste tipo de detector, detalhes sobre quantidades características de um CCDE (*bias*, ruído de leitura, corrente de escuro – *dark current* −, ADU, tamanho de pixel, ganho, taxa de aquisição de quadro – *frame rate*), veja Ahmed (2015), Evagora *et al.* (2012), Salvador *et al.* (2012), Holton, Nielsen e Frankel (2012).

A duração de uma noite de telescópio depende da latitude do observador e do dia do ano, variando tipicamente de 8 a 11 horas úteis, ou seja, sem o brilho do Sol nas camadas atmosféricas mais altas. Há três critérios de classificação do crepúsculo: o civil, o náutico e o astronômico. O primeiro compreende o período entre o pôr do sol, quando o disco solar aparente está abaixo da linha do horizonte , e o momento que em seu centro geométrico atinge uma altura de -6 graus. No segundo, o centro atinge a altura de -12 graus e, no terceiro caso, -18 graus, o que garante a mínima interferência da luz solar nas observações astronômicas.

Um projeto que é agraciado com tempo de telescópio passou pelo crivo de pesquisadores especialistas da área, que verificaram seu mérito científico, e de pesquisadores com profundo conhecimento do instrumental a ser usado, que emitiram parecer sobre a viabilidade e adequação das observações à ciência pretendida. O resultado dessa análise gera uma lista com projetos merecedores de tempo, estratificados por prioridade.

O número de projetos aceito geralmente é menor que o número de propostas submetidas. A fim de mensurar o interesse por determinado telescópio, define-se fator de pressão como sendo a razão entre o número de horas solicitadas por todos os projetos submetidos em um determinado semestre e o número de horas a serem efetivamente dedicadas à pesquisa (sem contar as noites de engenharia, testes de equipamento, realuminização dos espelhos, etc.). O valor médio para o OPD é de ≈ 2. Em outras palavras, é pedido o dobro das horas disponíveis.

Embora em uma proposta de pedido de tempo o pesquisador deva justificar o que pretende medir quase que cada minuto; na prática – e isto está incluído na sua previsão de tempo para seu projeto −, vários segundos e minutos úteis são gastos com movimentação do telescópio ao mudar de alvo, com medidas de calibração, troca de instrumento, leitura do CCDE, mudança

de filtros e redes de difração, abertura e fechamento do obturador que protege o detector para permitir ou não a incidência de luz, tomada de decisões sobre esperar as condições meteorológicas melhorarem ou mudar de projeto, etc. Isto reduz a quantidade de minutos efetivamente gastos em aquisição de dados para a ciência pretendida.

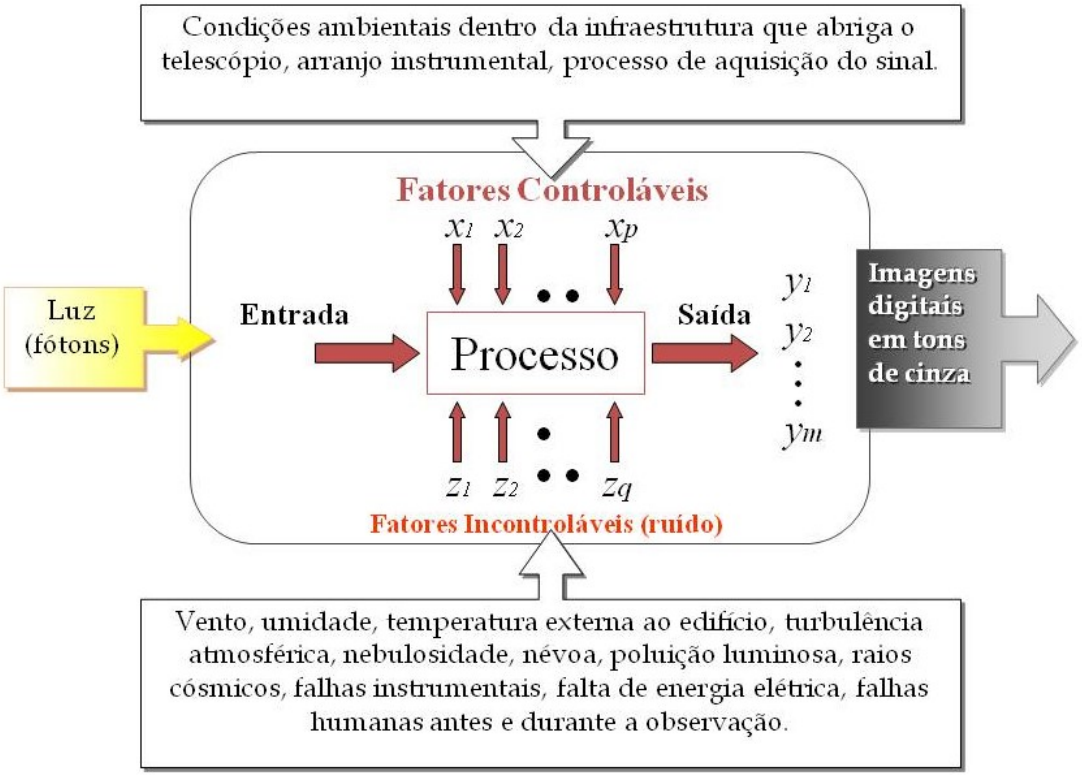

Figura 5.1 - Mapa conceitual resumido do processo de aquisição de dados astronômicos com telescópios em terra (Diagrama de Parâmetros).

Uma peculiaridade deste tipo de experimento é que os alvos não são visíveis durante todo o ano. Se não são medidos em um determinado período, somente poderão ser observados em uma janela até seis meses depois. Repita-se, aqui, cada medida é única, não existindo a rigor repetição.

Como calcular o custo de uma noite de observação no OPD? Este valor é calculado levando-se em conta três tipos de gastos: salários, operações e pesquisa. A título de exemplo, em um ano em que esses gastos somem R$ 5.200.000,00 e 330 noites estejam disponíveis, o custo de uma noite é de R$ 15.757,58. No entanto, o aproveitamento do sítio é de cerca de 50%, ou seja 165 noites são, efetivamente, usadas para observação, o que implica em R$ 31.515,15 por noite. Há quatro telescópios operando no OPD, cujos espelhos principais medem 160 cm, 60 e

60 cm e 40 cm, num total de 27.017,70 cm^2 de área coletora de luz, ou seja, cada cm^2 custa R$ 1,17. Uma vez que a taxa de câmbio do dólar vem flutuando de forma considerável nos últimos meses, e para efeito de cálculo, adotemos o valor de aproximadamente $ 10.000 dólares norte-americanos (USD) por noite no OPD. Para efeito de comparação e colocação do problema na perspectiva correta, em uma noite de 10 horas de duração, o procedimento de apontamento do telescópio, que demora 5 minutos, custa cerca de $ 83 USD; o procedimento de focalização, que é realizado no começo da noite e às vezes repetido sempre que o telescópio perde a referência de apontamento e/ou as condições de céu mudam, ao demorar de 12 a 20 minutos, custa quase $ 330 USD; o posicionamento da imagem no local correto sobre a superfície do CCDE leva de 2 a 10 minutos, ou seja, custa até $ 167 USD. A título de exemplo, uma noite no Canada-France-Hawaii Telescope (3,6 m de diâmetro) custa $ 25.000 USD, uma noite em um telescópio Keck (10 m de diâmetro, Havaí) custa $ 53.700 USD (NOAO, 2012) e uma noite em um telescópio do Gemini Observatory, $ 93.000 USD.

Dito todo o acima, compreende-se porque é de suma importância que o uso do tempo de telescópio seja o mais eficiente possível, através da otimização dos processos, do bom funcionamento dos equipamentos e da infraestrutura computacional, execução por profissionais altamente treinados e experientes e submissão de projetos bem dimensionados e que exploram ao máximo a capacidade do sítio, do telescópio e do instrumental. Solicitar a aquisição de imagens cuja qualidade não permitirá seu uso para a ciência proposta é inadmissível e passível da aplicação de penalidades aos pesquisadores proponentes quando das próximas submissões de pedidos de tempo.

É neste aspecto que a presente tese mais contribui. A diminuição dos riscos de perda de tempo, o melhor aproveitamento das condições de céu ao longo do ano, a melhor relação custo-benefício, tudo isto pode ser otimizado através da melhor compreensão do comportamento das condições físicas do sítio e a previsão do número de horas efetivamente úteis à observação astronômica.

5.1.2. Qualidade de dados obtidos em terra no visível e infravermelho

A radiação eletromagnética advinda dos corpos celestes sofre alterações ao longo de sua trajetória até o detector, causadas pela óptica dos instrumentos, dos telescópios, pela atmosfera terrestre, pelo meio interestelar, pelo meio intergaláctico e pela geometria do espaço-tempo, afinal. O efeito conjunto da atmosfera terrestre-meio interestelar recebe o nome genérico de extinção, e a natureza desses meios tem como efeito líquido o

avermelhamento do pincel de luz, não por acréscimo, mas por espalhamento da componente azul, desviando-a do feixe incidente. O sinal recebido, ademais, deve ser devidamente corrigido por técnicas de processamento antes de poder ser analisado.

Denomina-se Fluxo Específico a Intensidade luminosa por unidade de área (ou ângulo sólido), tempo, comprimento de onda (ou frequência) coletada por um telescópio, a qual tem por unidade [erg/cm^2/s/A]. A medida de Intensidade integrada em uma banda passante finita (filtro) mais usada é a magnitude, que pode ser aparente ou absoluta. A primeira é definida como $m_{\Delta\lambda} = [-2{,}5\ \log_{10}(I_{\Delta\lambda}) + constante\ zero\ de\ escala]$, onde $\Delta\lambda$ refere-se ao intervalo de comprimento de onda do filtro usado na observação e $I_{\Delta\lambda}$ é a intensidade da fonte. A segunda, por convenção, é a magnitude aparente de um objeto situado a uma distância de 10 parsecs (1 parsec equivale a 3,26 anos-luz, que, por sua vez, equivale a $3{,}26 \times 9{,}5 \times 10^{12}$ km) e é expressa por $M_{\Delta\lambda} = m_{\Delta\lambda} - 5[\log_{10}(dist\hat{a}ncia\ do\ objeto) - 1]$. Este moderno conceito foi proposto por Norman R. Pogson em 1856 e sua argumentação está reproduzida em Jones (1967). Crumey (2014) discorre sobre a visão humana, com detalhes sobre as células fotossensíveis e o limite de detecção do olho, estimado em 18,9 mag/seg arco^2 em $\lambda \cong$ 550 nm por Puschnig, Posch e Uttenhaler (2014), que equivale à luminância de 3x10^{-3} candelas / m^2; para efeito de comparação, o brilho do fundo de céu no visível é dado mais adiante.

As medidas de fluxo e de brilho superficial de um objeto qualquer envolvem o conceito de ângulo sólido, que em Astronomia pode ser aproximado pela razão entre o elemento de área na superfície de uma esfera e o seu raio ao quadrado; em coordenadas esféricas, a unidade de ângulo sólido é o esferorradiano, às vezes também denominado esterorradiano. Há que se ter em conta de que quaisquer medidas lineares e angulares são sempre feitas no plano do céu, ou seja, são sempre projetadas; a determinação dos ângulos de inclinação depende de modelos.

Dois conceitos importantes são comumente confundidos entre si: *seeing* e *Point Spread Function* (PSF).

O primeiro refere-se às alterações na inclinação das frentes de onda planas com relação à direção de incidência causadas pela turbulência atmosférica em várias alturas, as quais fazem com que o feixe passe por camadas de diferentes índices de refração. O resultado visual é uma imagem que "dança" no plano focal do detector, subdividindo-se, reagrupando-se e ocupando áreas de diferentes tamanhos em curtas escalas de tempo. A cintilação a olho nu está relacionada ao seeing, mas refere-se às variações de frentes de onda curvas, que resultam no

efeito visual descrito no linguajar leigo como “piscar” (DRAVINS *et al.*, 1997, termos entre aspas introduzidos pela autora). Leinert *et al.*, (1998, Fig. 1) apresentam uma relação dos principais componentes atmosféricos e extraterrestres que contribuem para o brilho de fundo de céu sem Lua, que nesse trabalho e na direção zenital era tipicamente cerca de 22 mag/seg arco^2; com os CCDEs hoje ultrapassa-se 25 mag/seg arco^2 . Mallmith (2004) apresenta dados específicos sobre o OPD de 1981-1994 e Caetano *et al.* (2010), de 2008 a 2009.

O segundo diz respeito ao sistema focalizado de coleta da radiação eletromagnética em si e caracteriza-o como um todo, podendo ser entendido como sendo a parte espacial da função de transferência do sistema de imageamento. A PSF caracteriza a distribuição bidimensional da luz de objetos admitidos como puntuais (estrelas que não o Sol) e é fortemente influenciada pelo seeing. Tanto engenheiros ópticos como os assistentes noturnos e pesquisadores esforçam-se por minimizar a PSF. O caminho óptico do feixe de luz (já alterado pela atmosfera) contém superfícies refletoras e refratoras, que causam perdas por absorção e espalhamento, por exemplo. A imagem de uma estrela acaba sendo composta por anéis de difração (“discos de Airy”) que são o resultado da convolução da imagem real com a PSF. Este fenômeno, chamado aberração esférica, é causado por imperfeições em espelhos e lentes.. Para sistemas ópticos bem ajustados e calibrados, a PSF é invariante por deslocamentos no plano focal; se a PSF de estrelas de campo em uma imagem variar significativamente, o instrumental apresenta problemas ou, se a exposição for relativamente longa em comparação com a escala de variabilidade atmosférica instantânea, a qualidade de céu sofre degradação. Num projeto de telescópio, principalmente os gigantes, busca-se o desempenho próximo ao limite de difração atmosférica. É comum atribuir-se um perfil gaussiano à PSF e avaliar o comportamento do instrumental através de medidas da Largura a Meia Altura de seu perfil ao longo do tempo (FWHM – *Full Width at Half Maximum*). Nesta aproximação, FWHM = $2\,\sigma\sqrt{2\ln(2)}$, onde σ é o desvio padrão (CRUMEY 2014). Figura 5.2 ilustra dois momentos do processo de focalização do telescópio: a aquisição de imagens para amostragem do perfil estelar e da medida da FHWM; o processo todo pode demorar vários minutos em função da habilidade dos operadores, dos pesquisadores e do comportamento da atmosfera.

Devido à extinção atmosférica, as observações geralmente são realizadas enquanto o objeto de interesse encontra-se a menos de 3 horas em torno de sua máxima altura no céu, que ocorre quando ele cruza o meridiano local – um círculo imaginário que contém o Polo Celeste do hemisfério terrestre da observação e o zênite do observador. Define-se Massa de Ar como

sendo a indicação-padrão da extinção atmosférica terrestre: $X = \sec(z)$, onde z é o ângulo zenital, medido na vertical a partir do zênite em direção ao objeto celeste; enquanto z varia de 0^o a 90^o, $1 \leq X \leq \infty$ (ALMEIDA, 2014; DE OLIVEIRA e SARAIVA, 2013). Quanto maior a massa de ar, pior é a qualidade dos dados.

Na análise da qualidade de um sítio astronômico e seu instrumental, além da qualidade do céu, do telescópio e instrumentos periféricos, são levadas em conta as horas que foram cientificamente úteis (que proporcionaram dados validados), as horas dedicadas à manutenção corretiva, às de engenharia e testes, às horas gastas em medidas de calibração e àquelas quando o telescópio esteve parado por motivos externos ao observatório como, por exemplo, falta de energia elétrica ou nevasca. Maiores detalhes são apresentados mais adiante, na caracterização do estudo de caso. Até aqui, apenas os aspectos ligados às observações astronômicas foram considerados; outros fatores tais como gerenciamento, finanças, recursos humanos, planejamento de ações nas áreas técnicas e científicas, entre outros, não serão abordados neste trabalho.

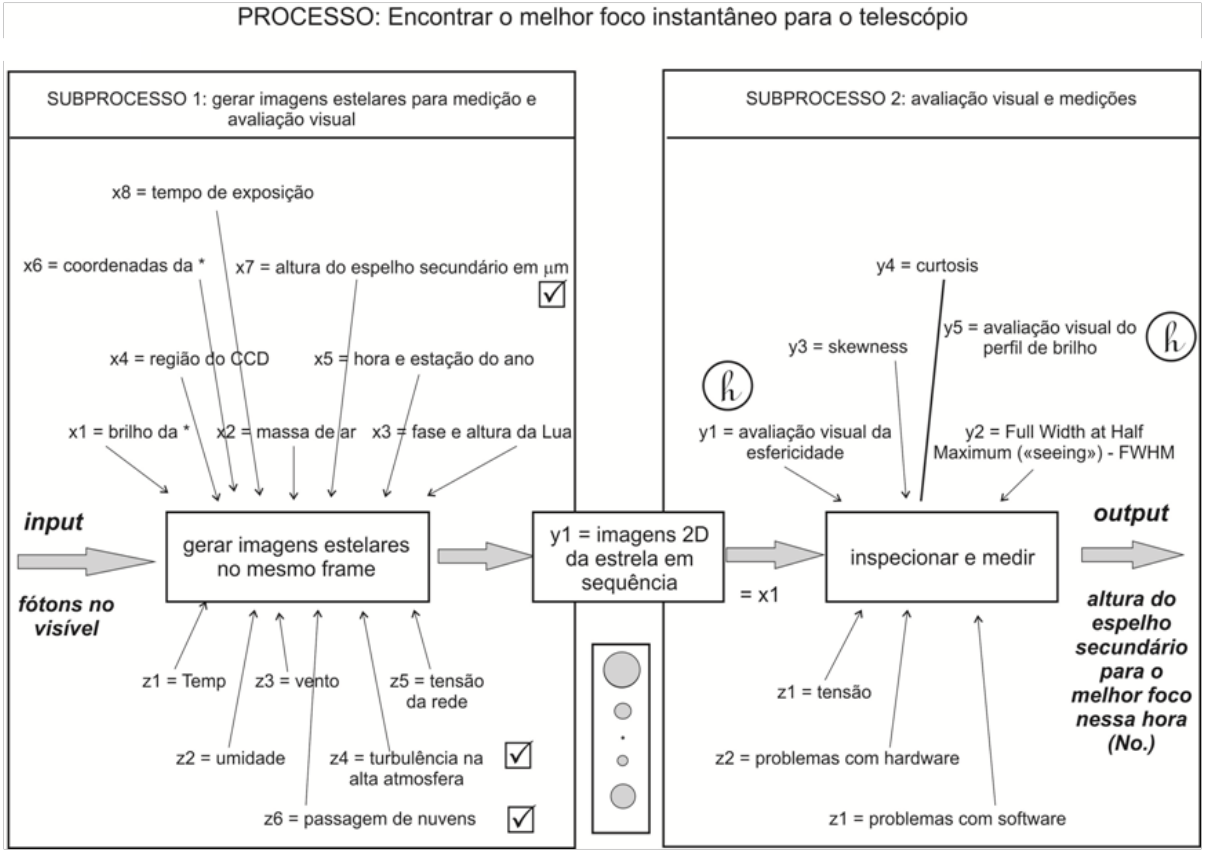

Figura 5.2 - Ilustração do processo de focalização de um telescópio ao longo da noite de observação em terra (Diagrama de Parâmetros). A letra *h* indica a necessidade de decisão por humanos.

5.2 Distribuição de carga elétrica

A aplicação da metodologia proposta a dados reais de um sistema elétrico de potência é bastante atraente devido à importância da capacidade de prever corretamente a carga a ser distribuída por operadores de sistemas de potência. É importante a previsão da demanda (variável) de carga devido ao fato de que, se em algum momento, o fluxo de potência reativa nas linhas não atende adequada e pontualmente os saltos de solicitação, o risco de colapso de tensão aumenta. Além disso, a previsão correta permite determinar os valores da geração pela unidades geradoras, o que, por sua vez, permite a adoção de ações de controle com objetivo de ajustar a tensão e a potência reativa. Esse balanço permite operar o sistema nas condições nominais. Balestrassi *et al.* (2009) apresentam vários aspectos do problema de previsão sob o prisma de riscos, preços e retorno financeiro, problema este enfrentado pelas empresas geradoras e/ou distribuidoras de energia elétrica.

Diversos fatores influenciam a necessidade de uso (ou não) da energia elétrica fornecida a residências, indústria, empresas, etc., por parte das empresas geradoras e distribuidoras, aqui entendidas como "clássicas". Fatores imprevisíveis e incontroláveis nesse complexo processo são, por exemplo, chuva, vento e temperatura ambiente, as quais são consideradas como variáveis exógenas ao processo. Existem outros fatores que são esperados, tais como aqueles de cunho social: a ocorrência de partidas de futebol, feriados locais e nacionais, festas e shows de grande porte, entre outros. Há vários tipos de análise para fins de planejamento, mas todos dependem do comportamento das cargas elétricas (MARUJO *et al.* 2015). Cada vez mais se lança mão de recursos de inteligência artificial na análise da estabilidade da tensão e na previsão do comportamento mercadológico, e RNAs não são exceção (MONTGOMERY *et al.* 2008; MAKRIDAKIS *et al.* 1998**).** Abordagens do tipo neuro-fuzzy têm sido usadas com sucesso nos casos em que há sensibilidade ao preço, quando o consumidor tende a reagir a mudanças de preço (GAILLARD *et al.* 2016; CORRADI *et al.* 2013; KHOTANZAD *et al.* 2002).

A perspectiva de mudança no paradigma de sistema de geração/distribuição de energia, no qual as relações entre preço e demanda, entre consumo, geração e distribuição de energia elétrica por cidadãos, empresas, universidades e outros, acrescida da preocupação com fontes de energia renováveis e da implantação de *smart grids* (malhas inteligentes de transmissão e distribuição de carga orientadas pela demanda) traz para a academia e para os operadores o

desafio de enfrentar um problema bidirecional, volátil e descentralizado (v., por exemplo, MOTAMEDI *et al.* 2012).

Neste novo paradigma, os consumidores finais, que antes eram exclusivamente passivos, passam a ser proativos e se empoderam através do controle tanto de seu consumo como de sua produção e armazenagem de energia. Recebem, agora, o nome de "*prosumers*". Morstyn *et al*. (2018) discutem a criação de plantas federadas de *prosumers,* as quais funcionariam em base a um balanço energético regulado por pares. Souza (2017) discute detalhadamente a interação entre os diferentes elementos geradores de energia, potência, redes de distribuição e de transmissão, com foco em microrredes; nestas, é crescente a introdução de fontes renováveis, o que "tende a mudar a matriz energética, introduzindo novos conceitos, como redes ativas e microrredes" (SOUZA, 2017). Existe, ademais, todo um esforço no sentido de prever e detectar em tempo real quaisquer mudanças dinâmicas no sistema elétrico de potência; recentemente, Mateos *et al.* (2017) basearam-se na Teoria da Informação e propuseram metodologia para determinar o momento em que séries temporais passam de um regime caótico para um aleatório. O aprofundamento nestes temas, no entanto, foge ao escopo deste trabalho.

A série histórica escolhida é não linear, compreende intervalo de tempo de um ano e possui medidas a cada hora (v. seção 5.3.1).

Esta etapa do projeto contou com o apoio de pesquisador e doutorandos do Grupo de Estudo de Operação de Sistemas Elétricos do Instituto de Engenharia Elétrica/UNIFEI – GPO/IEEL.

5.3 Aplicação I: Observatório do Pico dos Dias

O OPD vem sofrendo a perda gradativa da possibilidade de observação dos objetos celestes de baixo brilho por causa da poluição luminosa causada pelas cidades do seu entorno (GARGAGLIONI 2007 e 2009; DOMINICI 2013; DOMINICI E RANGEL 2017). A constante modernização do seu parque instrumental tem compensado essa perda, mas é evidente que o planejamento das observações em função do tipo de alvo e das condições de céu, mesmo fadado a atingir um limite, é a melhor estratégia de sobrevivência. Se a metodologia deste projeto se mostrar adequada a este planejamento, será uma ferramenta de assistência ao pesquisador e técnicos noturnos, podendo aumentar a vida útil do sítio. Segundo Gargaglioni (2012), "... a vida útil do OPD como laboratório científico está sendo comprometida pelo aumento descontrolado da poluição luminosa nos seus arredores."

Nesta seção, é apresentada a análise dos dados históricos do número de horas noturnas aproveitadas pelos telescópios do Observatório do Pico dos Dias (OPD) tendo em vista que se trata de uma série temporal. Limitações, aproximações e exploração de diversas formas de representação e análise dos dados são apresentadas. Este estudo serve, em última instância, como fundamentação estatística para uma visão de futuro do OPD, quando será necessário dedicar o sítio a apenas determinados tipos de observações e objetos celestes, a fim de continuar obtendo boa relação custo-benefício para a comunidade astronômica brasileira. Este estudo seminal evidencia, ademais, sua importância no suprimento da necessidade de embasar estatisticamente qualquer tomada de decisão a este respeito.

5.3.1. Os dados

Neste trabalho, usa-se a base de dados da distribuição anual das horas úteis no período de 38 anos de existência do OPD. A série histórica de dados deste trabalho é pública e está disponível em http://www.lna.br/opd/info_obs/historico_obs/todos.html. Os dados encontram-se em formato texto ASCII e, mês a mês, de abril de 1980 a dezembro de 2017. Para efeito desta pesquisa; as medidas do ano de 2017 foram reservadas para comparação com as previsões.

Uma vez que houve meses nos quais os dados não foram coletados, foi necessário adotar uma estratégia para estimar os *missing values* (valores faltantes e/ou inexistentes) para que o programa deixasse de acusar erros. Uma inspeção da série temporal como um todo (Figura 5.3) e de todos os anos sobrepostos em um só intervalo (Figura 5.4 e Figura 5.5) sugere que adotar a mediana correspondente aos meses em questão é aceitável. A abordagem de tratamento dos dados faltantes tem sido objeto de muitos trabalhos e artigos, *e.g*. Howell (2015a,b), Higgins e Green (2011), Horton e Kleinman (2007), Donders *et al*. (2006), Sterne *et al*. (2009) e Acock (2005). Há medidas faltantes em: dezembro de 2000, e novembro e dezembro de 2012. Decidiu-se, portanto, inserir as medianas em seu lugar (v. Figura 5.5). O mesmo foi feito nos meses de (i) fevereiro e março de 1983 e (ii) janeiro de 2010, 2011, 2012, 2013, 2014 e 2015, os quais possuíam valores nulos.

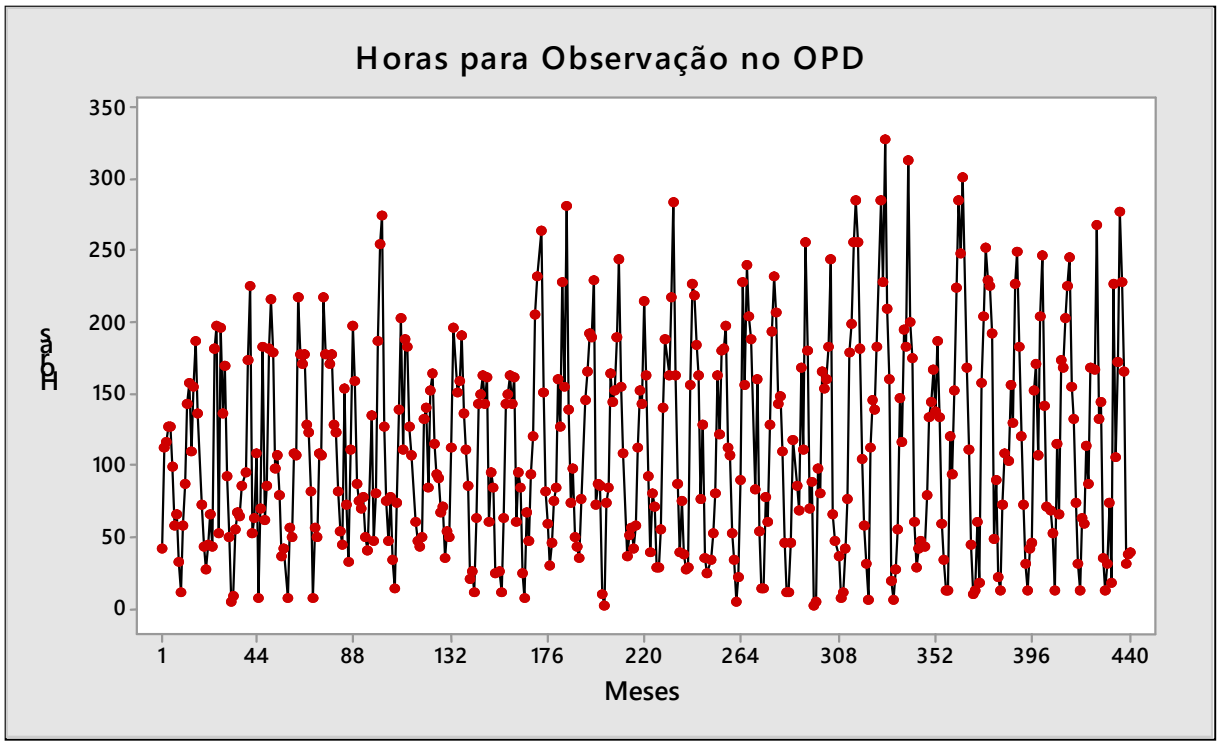

Figura 5.3- Série temporal completa com a distribuição do número de horas de céu adequado para observação no OPD.

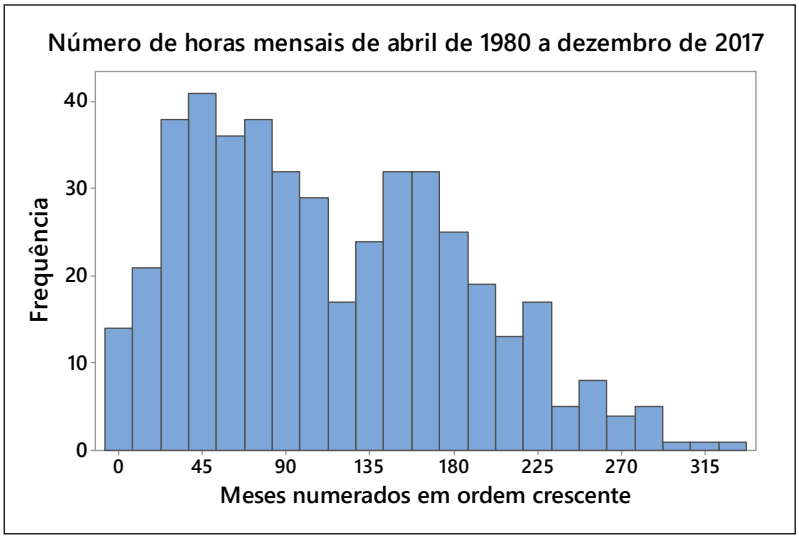

Figura 5.4 - Distribuição bimodal do número de horas úteis mensais no OPD. Prevalecem os meses de menor rendimento e maior dispersão nas medidas, marcadamente no verão, quando há maior ocorrência de chuvas na região. No outono e inverno, as condições atmosféricas estabilizam-se e permitem o melhor aproveitamento do sítio e instrumental. A distribuição bimodal reflete essa característica.

Mês	Mediana	Mês	Mediana
Janeiro	11,4	Julho	189,3
Fevereiro	54,2	Agosto	202,7
Março	69,0	Setembro	137,8
Abril	139,8	Outubro	91,3
Maio	154,8	Novembro	59,5
Junho	170,1	Dezembro	30,8

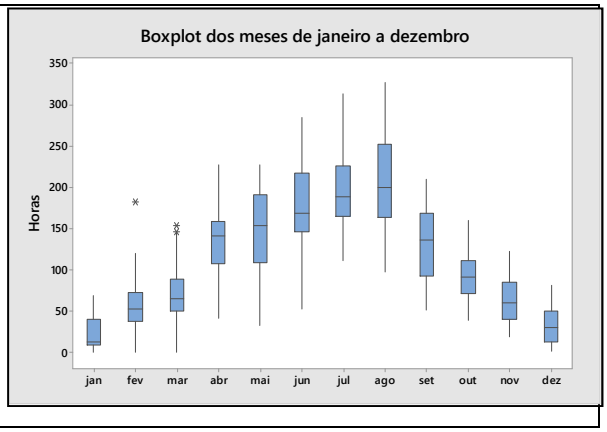

Figura 5.5- Valores medianos mensais das horas úteis no OPD de abril de 1980 a dezembro de 2017
Fonte: http://www.lna.br/opd/info_obs/historico_obs/todos.html

O início da série se dá em abril de 1980 pelo fato do início das operações (“primeira luz”) ter sido em abril, e (ii) é explicada pela concessão de férias coletivas aos colaboradores que trabalham no OPD por motivo de ainda ser a estação chuvosa e poucos dados serem obtidos. A baixa produtividade nessa época não justificava o funcionamento do observatório.

Foi realizado um estudo do comportamento da série temporal desde dois meses antecedentes até sete meses subsequentes aos meses de janeiro acima, com o objetivo de determinar (1) se havia constância no comportamento temporal dentro da mesma estação do ano (novembro e dezembro, e de janeiro a junho) no período compreendido por 14 anos antes e dois depois desse intervalo, e (2) se seria válido adotar um mesmo valor mediano para todos os meses de janeiro desses anos. Cada mês foi tratado separadamente e curvas lineares e quadráticas foram ajustadas segundo cada distribuição. Os melhores ajustes foram selecionados e colocados num mesmo gráfico (Figura 5.6), onde se pode notar que nos meses de novembro a fevereiro as curvas são menos acentuadas que nos restantes, o que faz com que o uso do valor mediano das horas de janeiro nos meses sem essa medida seja razoável aproximação.

A precisão dos dados varia de 0,05 h a 0,5 h; assim sendo, é cauteloso adotar um erro comum de 0,5 h para todas as medidas.

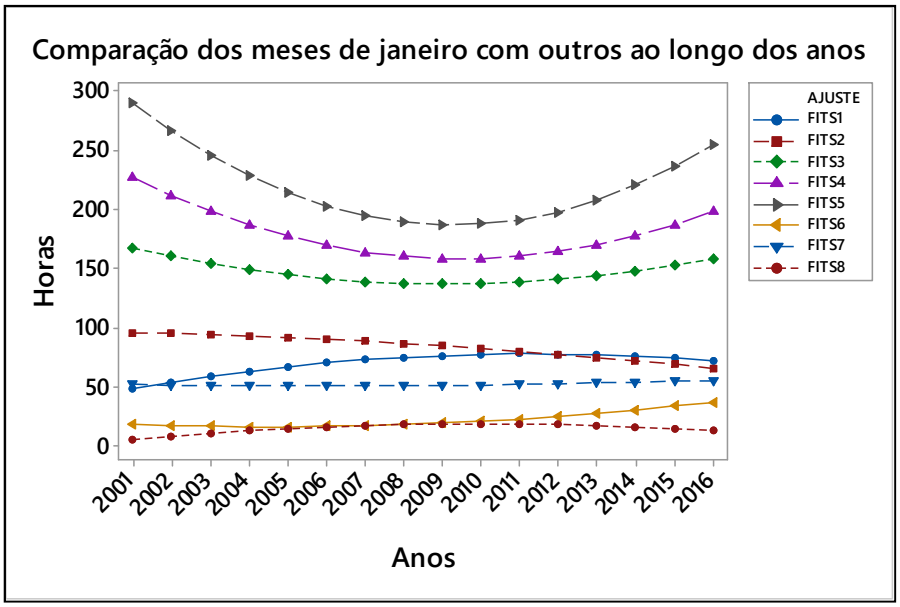

Figura 5.6 - Valores das horas ajustadas por diferentes modelos para os meses de fev(1), mar(2), abr(3), mai(4), jun(5), dez(6), nov(7) e jan(8). Com exceção dos ajustes 3,4 e 5, os restantes indicam que a adoção de valores medianos é razoável devido à pequena curvatura no período considerado.

5.3.2. Análise espectral

Procedeu-s à análise espectral com o objetivo de identificar os principais períodos que compõem a série temporal. A Tabela 4.4 contém os quatro valores mais altos do periodograma e do diagrama de densidade espectral, onde se verifica que a sazonalidade é de 12 meses. Existem outros picos, mas bem menos pronunciados, cuja existência deve ser explorada no futuro (v. Figura 5.7).

Tabela 4.4– Valores de maior potência na análise espectral. A série possui sazonalidade de 12 meses.

No.	Frequência	Período	Periodograma	Densidade espectral
37	0,082589	12,10811	989292,0	553545,2
38	0,084821	11,78947	380257,2	421372,8
36	0,080357	12,44444	75207,0	289935,0
39	0,087054	11,48718	41193,1	148964,6

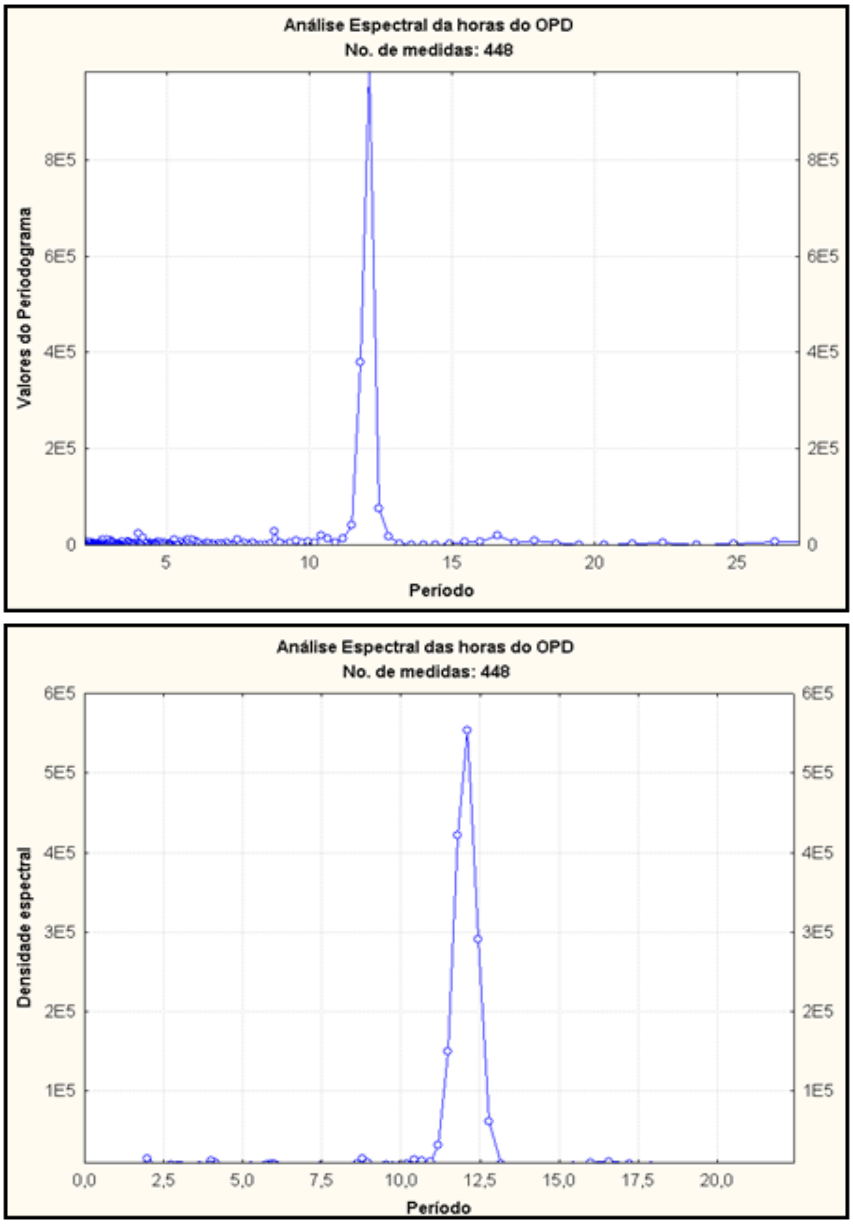

Figura 5.7– Identificação de pico no período correspondente a 12 meses. Este valor foi, então, adotado como parâmetro fixo na construção das RNAs.

5.3.3. Adequação do caso

Uma primeira inspeção visual da série poderia levar a crer que ela é estacionária na média, mas não na variância. Primeiramente calcula-se a ACF e a PACF com *lag* (retardo e/ou intervalo de tempo) *lag*=12 (Figura 5.8). Com base nos comportamentos das curvas, pode-se afirmar que esta não é uma série estacionária e há sazonalidade.

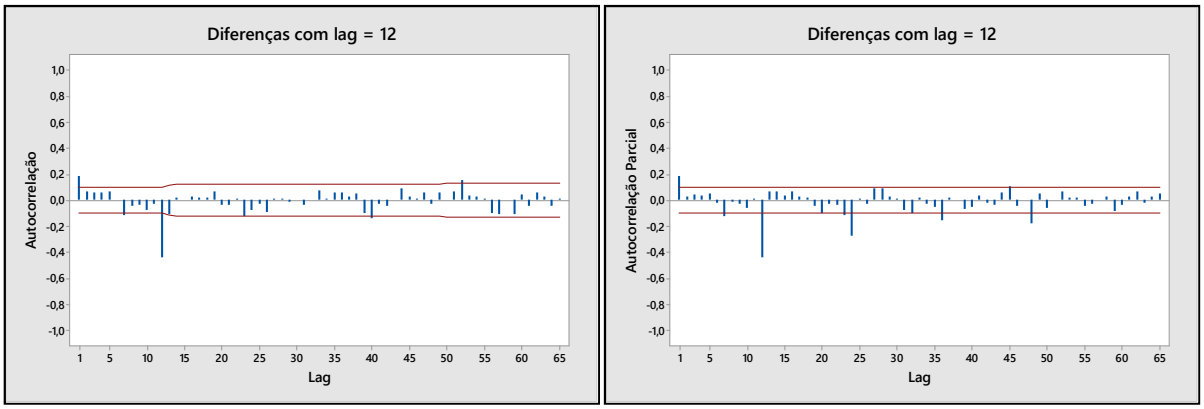

Figura 5.8- ACF e PACF da diferença de ordem 12.

Para concluir se há ou não alguma tendência (*trend*) e/ou sazonalidade e determinar valores característicos, há duas verificações que devem ser feitas:

(a) Estudo da tendenciosidade − deve-se primeiro verificar a condição de existência ou não de tendência e, se sim, se a mesma é linear, caso contrário, o modelo não se aplica. A remoção da tendência foi feita com as opções de ajuste linear, quadrático, exponencial e curvo (S-*curve* − *Pearl-Reed logistic*). Com base nos resultados obtidos, descartaram-se os dois últimos modelos. As medidas estatísticas para o modelo linear e o quadrático são, MAPE = 189,96, MAD = 60,15, MSD = 5053,81. Devido à semelhança da qualidade entre os ajustes linear e quadrático, por simplicidade, adotou-se o modelo linear. Consequentemente, é possível prosseguir com a análise.

(b) E em segundo lugar, deve-se decidir se o modelo é aditivo ou multiplicativo. Esta decisão baseia-se na inspeção visual da série como um todo e na análise da distribuição sazonal (de comprimento = 12) dos resíduos de cada modelo. Uma inspeção visual da série sugere que a sazonalidade não varia com o passar do tempo ou com o nível do sinal, o que é indicação de aditividade. A distribuição dos resíduos versus valores ajustados do modelo aditivo não apresenta agrupações destacadas. As medidas estatísticas para o modelo multiplicativo e o linear são, respectivamente: MAPE = 58,4, MAD = 29,8, MSD = 1513,3 e MAPE = 59,4, MAD = 29,7, MSD = 1478,5. As diferenças de qualidade entre ambos (+1,6%, -0,1%, -3,6%, em ordem) indicam ser possível decidir-se pelo ajuste do modelo aditivo (Figura 5.9).

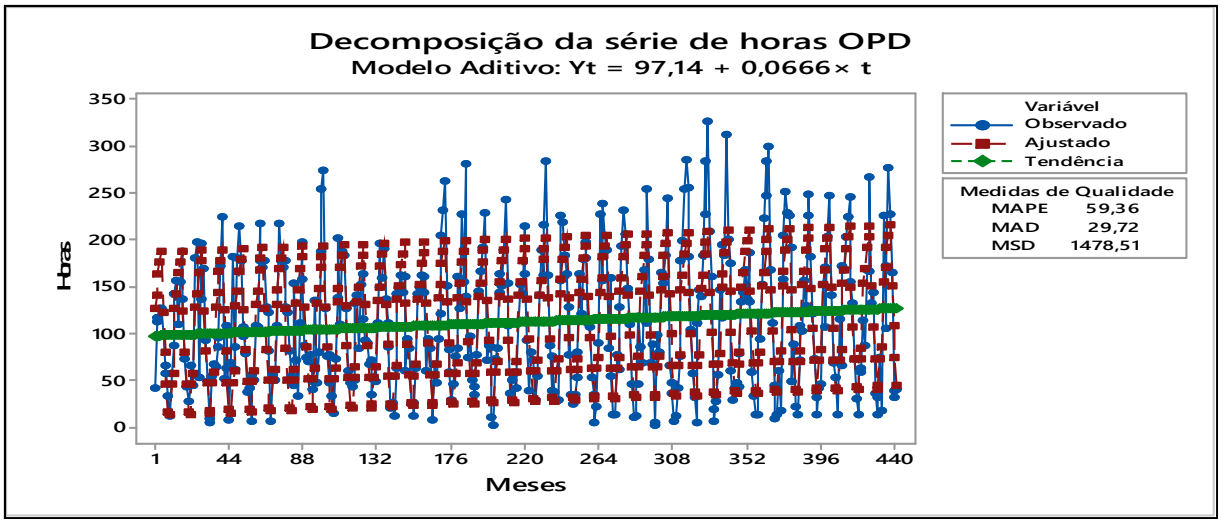

Figura 5.9– Ajuste de modelo aditivo à série do OPD. A linha verde representa a tendência da amplitude média. Veja comentário na seção 5.6.

5.3.4. Abordagens "clássicas"

Dando prosseguimento à busca de uma série estacionária, além do cálculo das diferenças de ordem 12 (v. seção 5.3.3), aplicaram-se algumas outras transformações (Quadro 4.6). A que melhor resultado (Figura 5.10) forneceu foi a transformação via aplicação do operador log_{10} e, em seguida, o cálculo da diferença de ordem 12 sobre esse logaritmo (MAKRIDAKIS *et al,*. 1998).

Quadro 4.6– Transformações aplicadas à série do OPD na busca de estacionariedade

Série OPD Número de horas mensais	
Escolha	Dif=diferença, O=Ordem
	dif O(1)
	dif O(12)
	log_{10}(horas)
	dif O(1) do log_{10}(horas)
X	dif O(12) do log_{10}(horas)
	dif O(12) da dif O(1) do log_{10}(horas)
	dif O(12) da dif O(12) do log_{10}(horas)

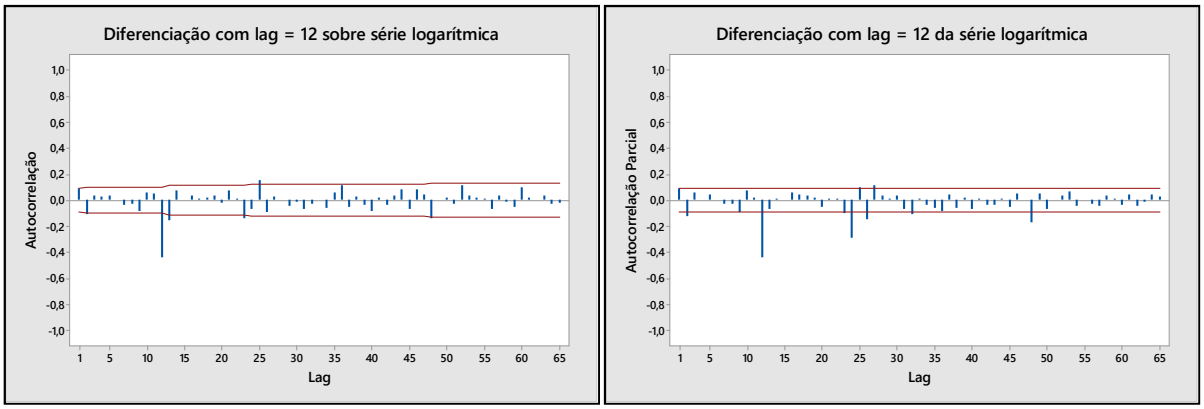

Figura 5.10- ACF e PACF da 12ª. diferença em log10.

A transformação logarítmica na base 10 justificar-se-ia pelo aparente aumento da amplitude ao longo do tempo na série alterada. A fim de testar esta hipótese, procedeu-se à análise da variância ao longo de alguns anos.

Consta que em 1990 entrou em operação um novo tipo de detector, o CCDE, detector de estado sólido de tecnologia de ponta então, e que viria a permitir melhores dados e o melhor aproveitamento do sítio. Sendo equipamento novo no meio científico, houve um intervalo de tempo ("aquecimento") até que os pesquisadores aprendessem a coletar os dados, reduzi-los (torná-los cientificamente úteis) e analisá-los. Vários problemas de ordem técnica também se fizeram presentes. Adota-se aqui o ano de 1994 como sendo o primeiro ano em que o uso do CCDE tornou-se corriqueiro. Houve, certamente, outros fatores de influência no aproveitamento do sítio, mas essa análise mais detalhada foge ao escopo deste projeto. Dividindo-se, então, de forma arbitrária, a amostra em duas partes, de 1980 a 1993, e de 1994 até outubro de 2012, obtém-se a Figura 5.11, a qual mostra o comportamento da amplitude das medidas e da variância em dois intervalos de tempo, e na qual se nota que ambas não são constantes e seguem o mesmo padrão. Estes resultados sugerem a adoção da transformada logarítmica. As Figura 5.12 e Figura 5.13 corroboram esta decisão.

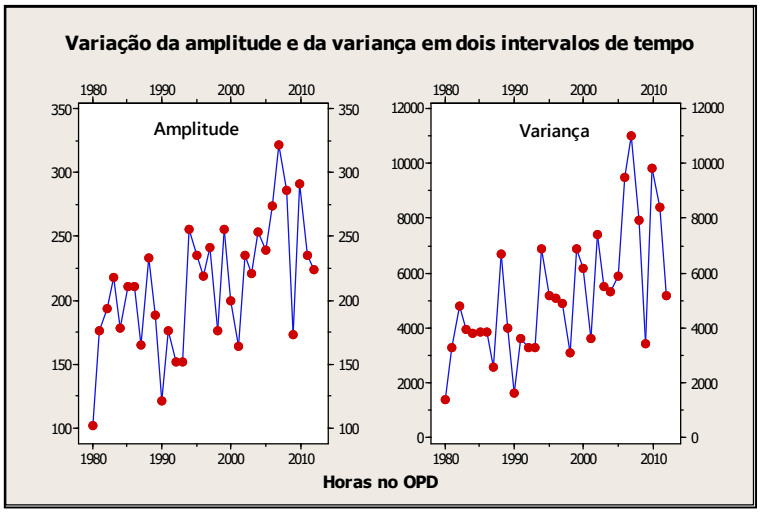

Figura 5.11– Resultados da análise de amplitude e variância

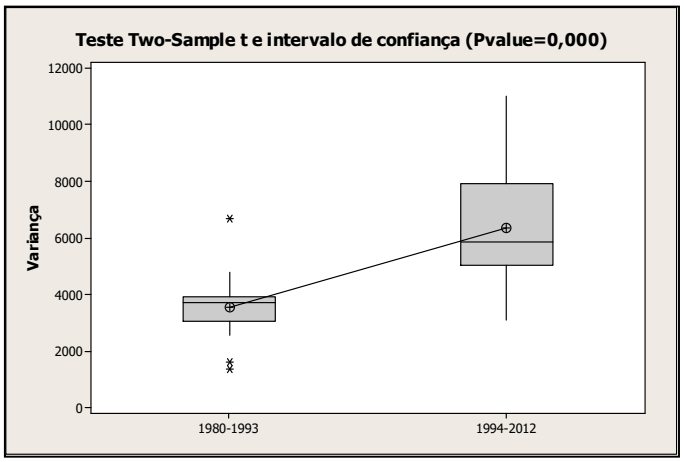

Figura 5.12 - Mudança na variância entre dois períodos de vida do OPD: de 1980 a 1993 e de 1994 até outubro de 2012. Os pontos fora dos quartis referem-se aos anos de 1980, 1988 e 1990.

```
Two-sample T for vartill1993 vs varrest

              N  Mean  StDev  SE Mean
vartill1993  14  3557   1289      344
varrest      19  6368   2202      505


Difference = mu (vartill1993) - mu (varrest)
Estimate for difference:  -2812
95% upper bound for difference:  -1773
T-Test of difference = 0 (vs <): T-Value = -4,60  P-Value = 0,000  DF = 29
```

Figura 5.13 - Resultado do teste de hipótese para verificar se a variância do período de vida do OPD compreendido entre 1980 e 1993 é estatisticamente menor que o período de 1994 até outubro de 2012.

5.3.5. Ajuste de modelo SARIMA

De posse da série estacionária, o passo seguinte é identificar o melhor modelo SARIMA para fins de previsão e *benchmarking* para as RNAs. Um bom modelo proporciona uma base de comparação bem estabelecida para a metodologia proposta neste trabalho. O modelo que apresenta ACF e PACF dos resíduos sob a forma estacionária é o melhor.

Procedeu-se à modelagem SARIMA com base simulação de Monte Carlo sazonal com medida de erro MAPE. O melhor modelo foi SARIMA(2,0,2)(1,0,1)$_{12}$ com MAPE = 52% e parâmetros de ajuste apresentados na

Tabela 4.5. A componente sazonal (1,0,1) já era esperada pelas características da ACF e PACF. A Figura 5. 14 traz os dados originais, o ajuste e a previsão para 12 meses.

Tabela 4.5– Coeficientes do modelo SARIMA obtidos por simulação de Monte Carlo

Variável	Coeficiente	Erro Padrão
AR(1)	1,7300	0,0022
AR(2)	-0,9980	0,0031
MA(1)	1,7000	0,0033
MA(2)	-0,9849	0,0061
Sazonal AR(1)	0,9952	0,0072
Sazonal MA(1)	0,9033	0,0236
Constante	0,1444	

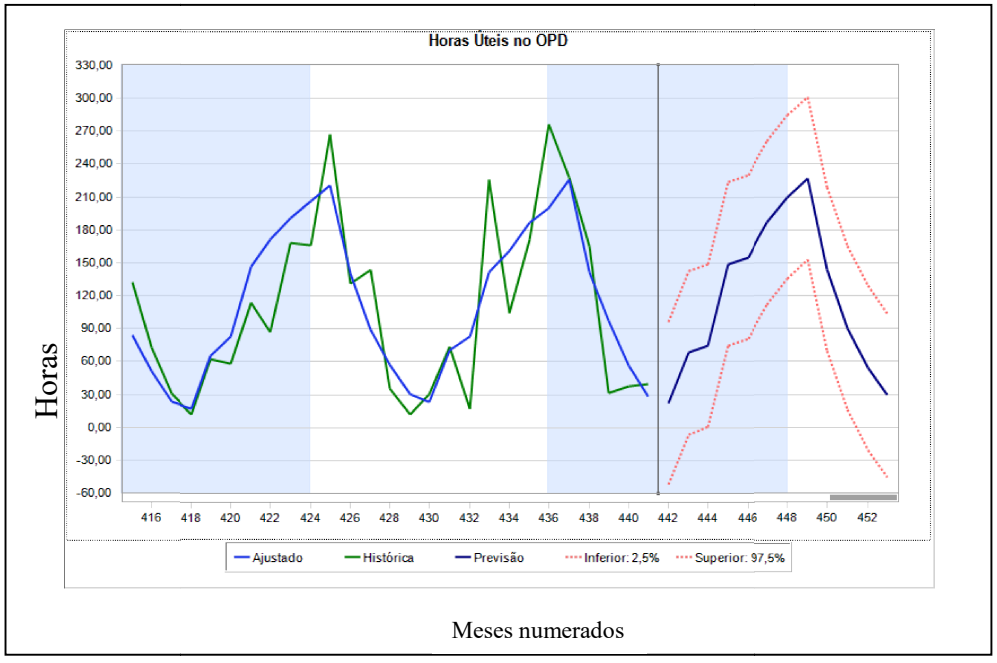

Figura 5. 14 – Dados originais, ajuste e previsão para as horas úteis do OPD por SARIMA via Monte Carlo.

As previsões para o ano de 2017 fornecidas por este modelo foram comparadas via MAPE com as medidas disponibilizadas pelo OPD, mês a mês, com média de -11% e mediana de +7%. Considera-se que este modelo serve ao propósito de *benchmarking*.

5.3.6. Aplicação da metodologia PBCA

Devido à sazonalidade de 12 meses, escolheu-se o número de medidas usadas para previsão também de 12 (SP).

A Tabela 4.6 contém as medidas do erro de cada RNA empregando os fatores fixos e a combinação respectiva dos variáveis. A ordem dos experimentos é aleatória. O último experimento, No. 20, é discordante dos demais e não consta da Figura 5.16. Nesta figura, aparentemente, os valores de MdAPE não estão correlacionados com os de MAPE no intervalo 60% ≤ MAPE ≤ 80%. Uma vez que a mediana é pouco sensível a valores extremos quando estes não são representativos da amostra, uma possível razão é a presença de grande amplitude nos valores que foram usados em cada experimento, ou seja, houve combinações de parâmetros de RNAs que produziram variações notáveis e influenciaram o cálculo de MAPE. Os modelos que forneceram os menores valores de MAPE e MdAPE foram, respectivamente, os de Nos. 6 e 24.

Os valores de MAPE são geralmente maiores que os de MdAPE porque consideram os valores discordantes. Servem, no entanto, como medida de dispersão.

Tabela 4.6– Resultados das RNAs para uso com a PBCA. Os valores em negrito e fundo cinza referem-se aos melhores modelos, de Nos. 6 (MAPE) e 24 (MdAPE).

Ordem	MAPE	MdAPE
9	90,8	26,5
12	78,4	24,0
21	69,1	24,5
7	69,4	25,1
13	77,1	24,0
18	84,2	25,1
2	75,5	26,7
11	77,4	24,8
1	79,7	25,1
5	70,2	23,4
23	81,1	24,0
14	93,9	27,6
6	**68,3**	24,5
24	81,6	**22,4**
10	140,3	34,7
15	104,1	28,6
17	73,5	25,5
8	76,8	25,5
16	71,3	24,6
22	69,7	24,9
19	71,6	25,7
4	160,2	38,3
3	110,6	28,9
20	609,6	189,7

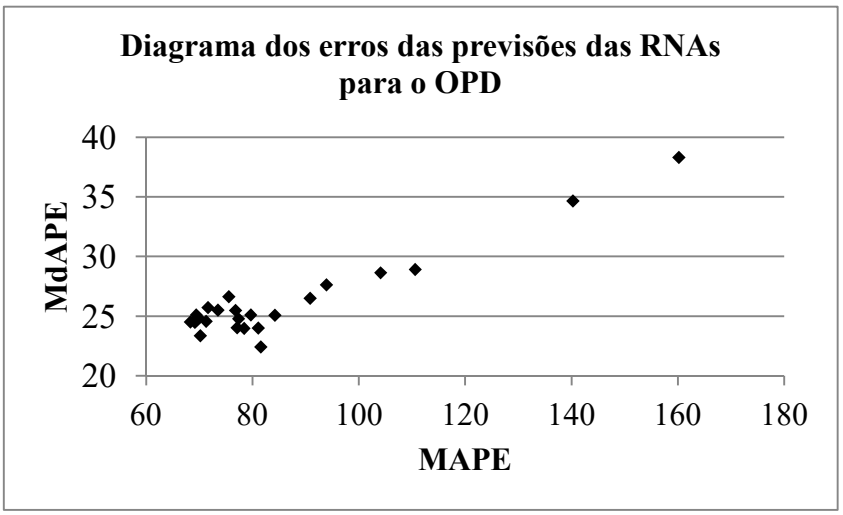

Figura 5.15 – Diagrama de dispersão entre as medidas de erro do desempenho das RNAs

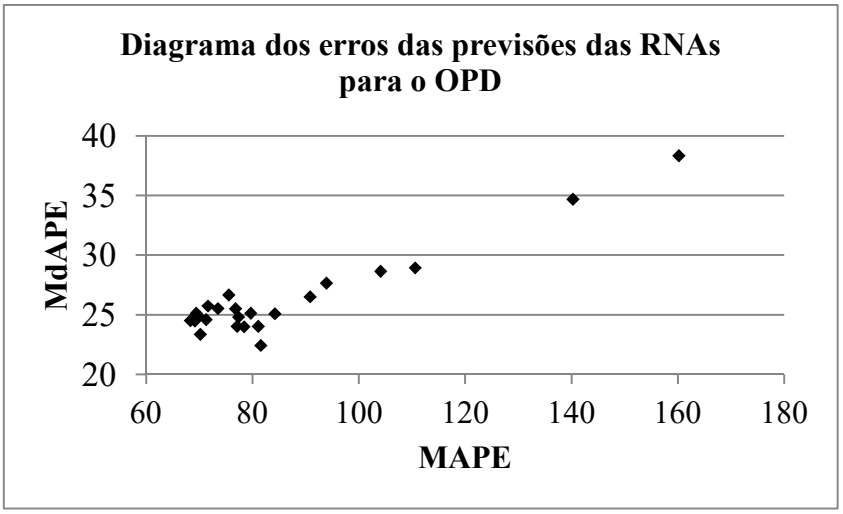

Figura 5.16 – Diagrama de dispersão entre as medidas de erro do desempenho das RNAs

O comportamento dos pontos discordantes não pode ser explicado apenas pela combinação dos fatores, já que elas não guardam similaridades entre si como um todo (Tabela 4.7).

Tabela 4.7 – Combinação dos valores dos parâmetros dos pontos discordantes

HL	UL	P2	LR	SC	ET	W1	W2	SM	Ordem
2	1	LevMarq	0,01	0,1	0,0	No	Yes	Rand	10
2	1	LevMarq	0,10	0,0	0,0	No	No	CrossVal	15
2	1	BFGS	0,10	0,0	0,0	Yes	Yes	CrossVal	4
1	12	BFGS	0,01	0,1	-0,1	Yes	Yes	Rand	3
2	12	LevMarq	0,01	0,0	0,0	Yes	Yes	Rand	20

Já os pontos com os melhores MAPE e MdAPE compartilham as características HL, UL, P2, SC, ET e W2 (Tabela 4.8). A título de ilustração, a Figura 5.17 contém porções da previsão dos experimentos 6 e 24.

Tabela 4.8 - Combinação dos valores dos parâmetros dos pontos com os mais baixos erros para a PBCA da Fase 1 (apenas os fatores)

HL	UL	P2	LR	SC	ET	W1	W2	SM	Ordem	Melhor
2	12	BFGS	0,01	0,1	0,0	Yes	No	CrossVal	6	MAPE
2	12	BFGS	0,10	0,1	0,0	No	No	Rand	24	MdAPE

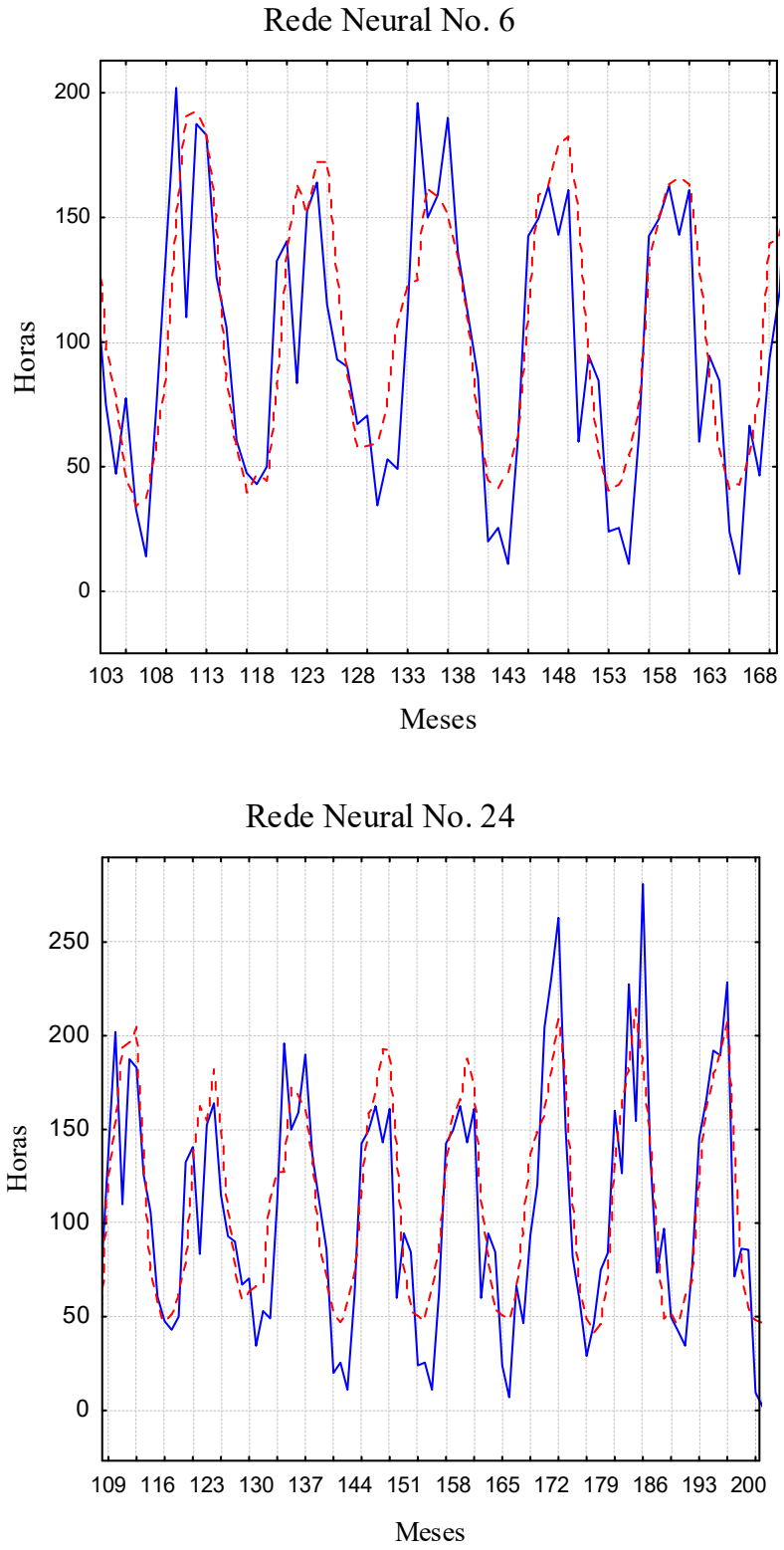

Figura 5.17– Previsões dada pelos experimentos de números 6 (sup.) e 24 (inf.). Em ambos os painéis, a linha contínua azul representa os dados experimentais e a linha tracejada vermelha representa os resultados dos modelos.

Na PBCA da Fase 2, os resíduos considerando todos os fatores e as diversas combinações dois a dois dos mesmos, os melhores resultados foram obtidos para os experimentos de No.s. 52 (MAPE) e 31 (MdAPE) – v. Tabela 4.9.

Tabela 4.9 - Combinação dos valores dos parâmetros dos pontos com os mais baixos erros para a PBCA da Fase 2 (todos os fatores mais as interações selecionadas)

HL*SC	HL*W1	HL*W2	SC*W1	SC*W2	W1*W2	Ordem	Melhor
1	1	1	1	1	0	31	MAPE
0	0	1	0	1	1	52	MdAPE

A Tabela 4.15 da seção 5.5 traz os resultados da aplicação da metodologia PBCA aos fatores (Fase 1) e aos quatro pares de interação de segunda ordem pré-escolhidos, conforme explicitado no Capítulo 4 (Fase 2). Apesar da delimitação do estudo dos pares de interações de ordem 2, para esses quatro pares, comparando-se os valores de MVPA para os casos de MAPE e de MdAPE resultantes da aplicação da PBCA em ambas as fases, nota-se que ambos diminuíram de valor. Isto indica a presença de interações significativas (COUTO, 2012).

A Figura 5.18 contém a série completa até dezembro de 2017 e a Figura 5.21 apresenta apenas os dados de janeiro de 2016 até dezembro de 2017 juntamente com as melhores previsões obtidas neste trabalho (experimentos de Nos. 6 e 24).

As Figuras Figura 5. 19 e Figura 5. 20 contêm os resíduos dos modelos de RNAs No. 6 e No. 24. Diagramas de densidade espectral em função do período não apresentaram nenhum pico proeminente, o que significa que os modelos ajustam a série temporal sem deixar vestígios de sazonalidade. Os histogramas são unimodais e simétricos em torno de zero, o que significa que os ajustes não apresentam tendenciosidade.

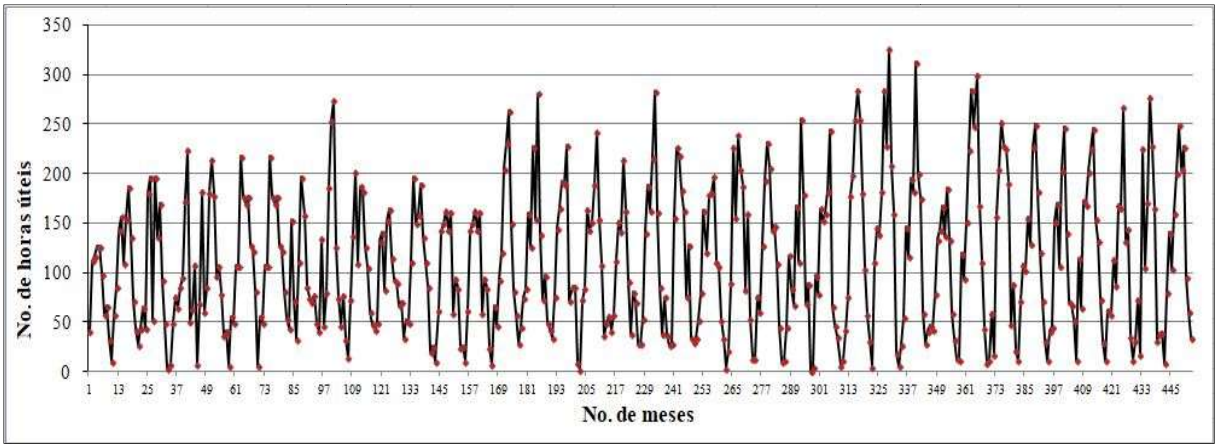

Figura 5.18 – Visão geral da série temporal do OPD.

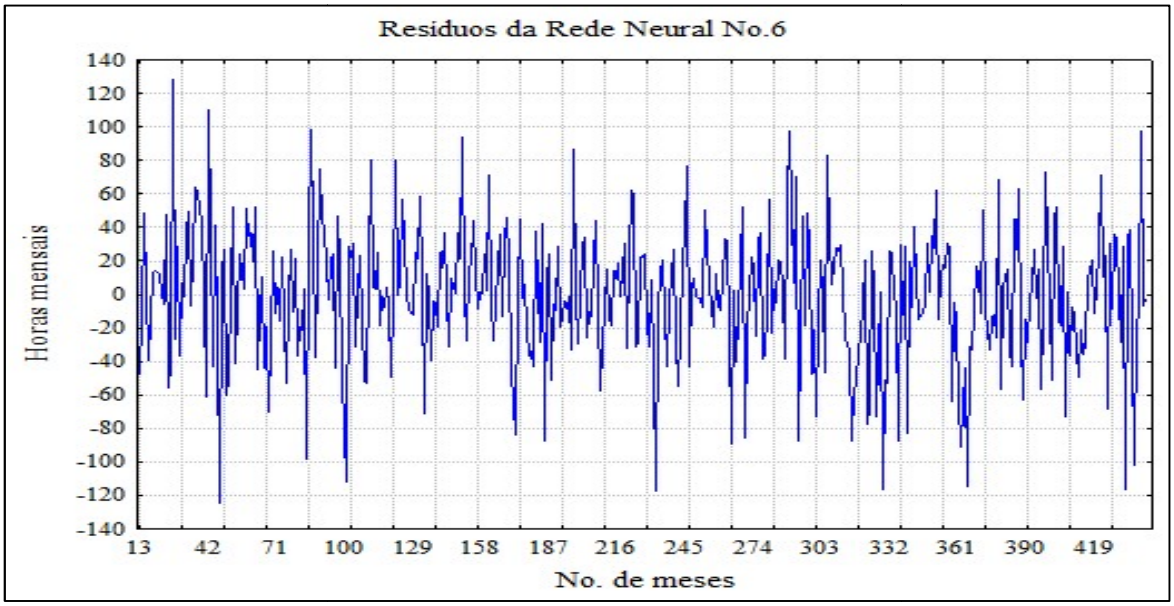

Figura 5. 19 – Visão geral dos resíduos em horas mensais da Rede Neural No. 6 com relação aos dados experimentais para o OPD.

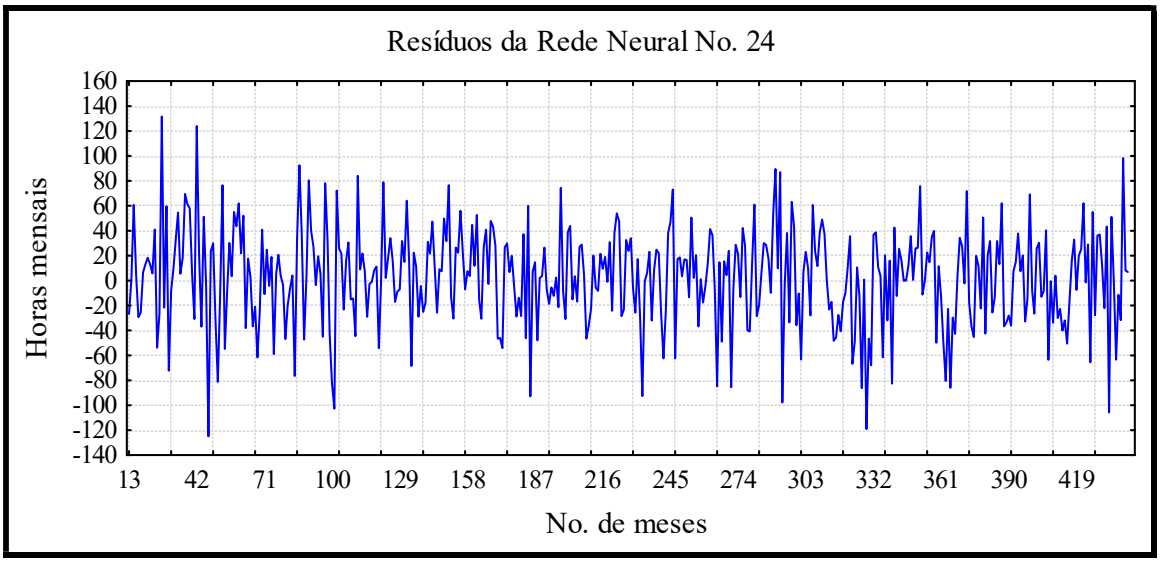

Figura 5. 20 – Visão geral dos resíduos em horas mensais da Rede Neural No. 24 com relação aos dados experimentais para o OPD.

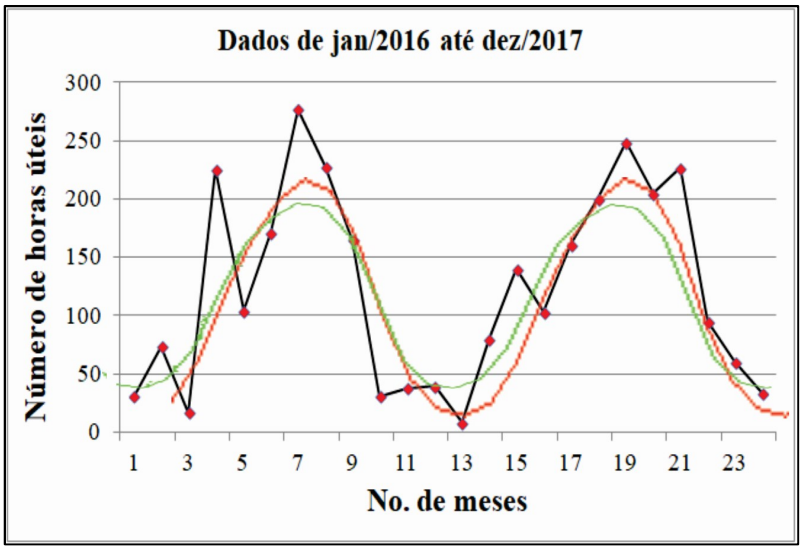

Figura 5.21 – Comparação entre os dados reais e as melhores previsões deste trabalho para os meses de janeiro de 2016 a dezembro de 2017**.** A curva em vermelho é a previsão do experimento No. 24, a curva em verde é a previsão do experimento No. 6.

5.4 . Aplicação II: distribuição de carga elétrica

A análise dos dados antes da aplicação da metodologia proposta nesta tese seguiu o protocolo apresentado na seção 5.1.

5.4.1. Os dados

Trata-se da evolução temporal da distribuição de carga elétrica fornecida pela subsidiária brasileira da empresa Duke Energy, durante o ano de 2008. As medidas foram realizadas a cada hora, de 1º. De janeiro a 31 de dezembro interruptamente; são 8.784 medidas no total, cuja amostra está caracterizada na Tabela 4.10 e nas Figura 5.18, Figura 5.23 e Figura 5.24. Os dados gentilmente cedidos pela Agência Nacional de Energia Elétrica−ANEEL. Não há informação sobre o método de coleta dos dados ou da precisão das medidas. Esta é uma série não estacionária; é volátil, sazonal, com média e tendência variáveis, em patamares.

Tabela 4.10 – Estatísticas básicas da série temporal Duke8

Parâmetro	Valor
Média	2.461,73
Mediana	2.388,05
Desvio Padrão	538,49
№ de medidas válidas	8.784
Valor mínimo	1.300,90
Valor máximo	3.747,90
1o. percentil	2.014,50
3o. percentil	2.934,45

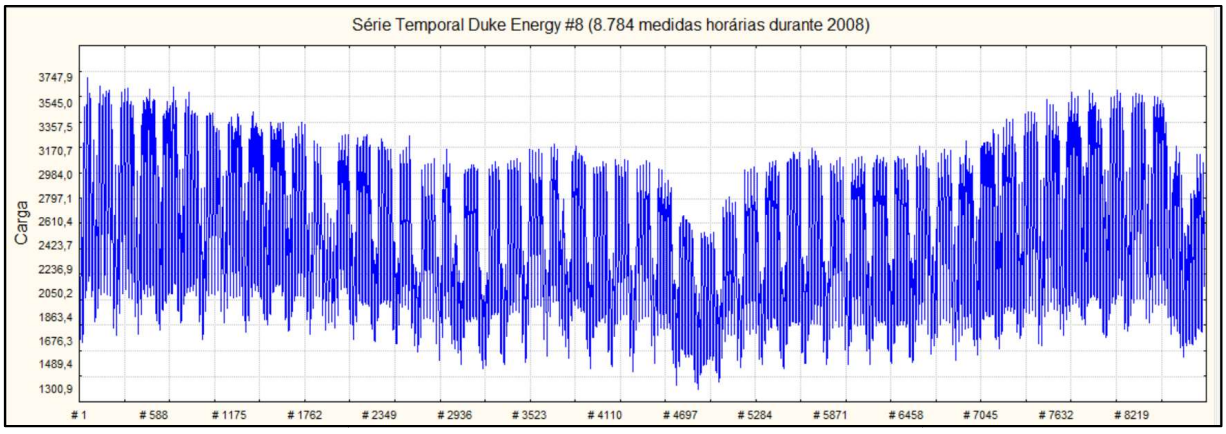

Figura 5.22 – Diagrama compactado da série temporal de distribuição de carga elétrica Duke8 em MWatts.

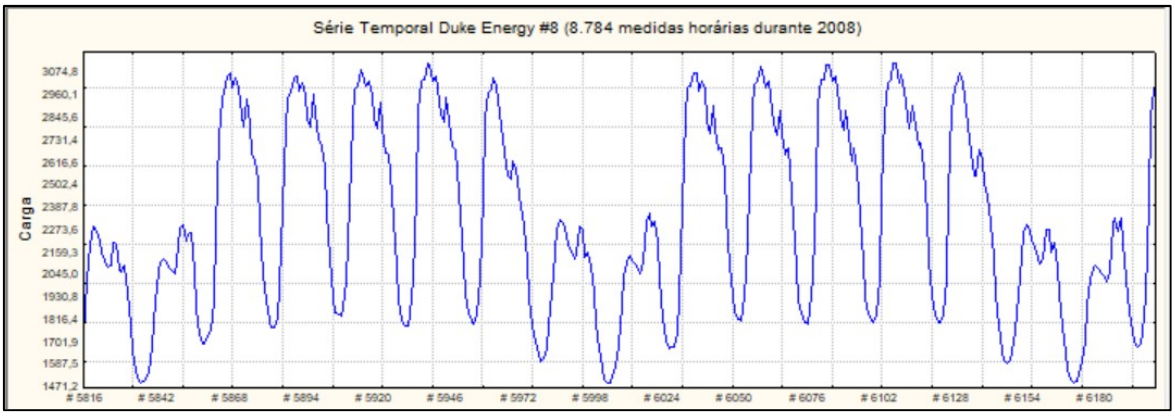

Figura 5.23 – Visão de porção da série em maior detalhe. Pode-se discernir os períodos semanais, inclusive dos fins de semana. A carga é dada em MW.

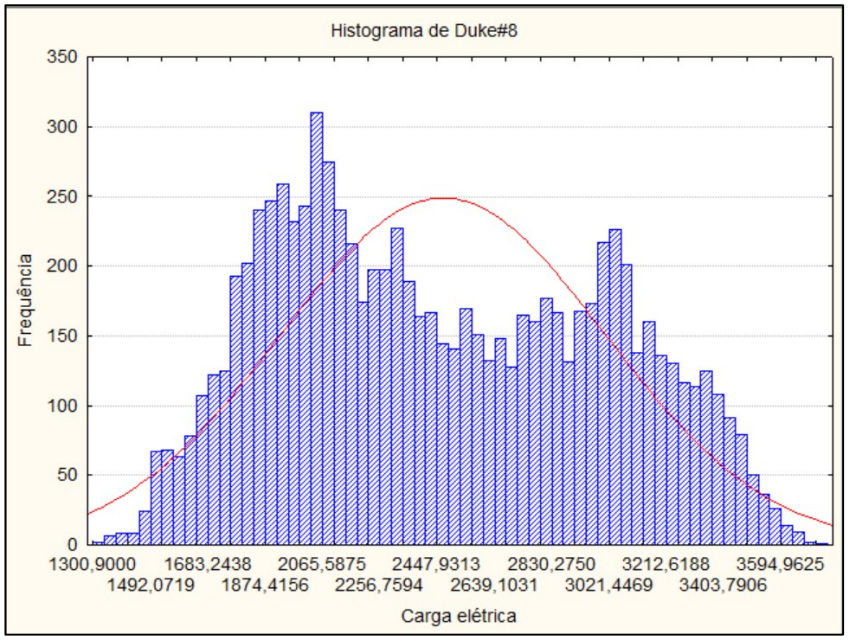

Figura 5.24 – Histograma dos dados da série Duke8 em Mw evidenciando sua natureza bimodal.

5.4.2. Análise espectral

A análise espectral (Figuras Figura 5.25, Figura 5.26 e Figura 5.27) revela a existência dos seguintes períodos notáveis: 12 horas, 24 horas, 84 e 168 horas. O uso de energia elétrica depende do tipo de vida e das atividades das comunidades às quais uma distribuidora serve, incluindo aí as outras distribuidoras e clientes com os quais comercializa a carga elétrica. O pico de densidade espectral em 84 horas indica a presença de uma componente periódica de 3,5 dias (meia semana), que não é imediatamente identificável ou associada a algum

fenômeno social, necessitando um estudo mais abrangente, com maior número de variáveis, na busca de uma ou mais causas. Isto, no entanto, foge ao escopo deste trabalho

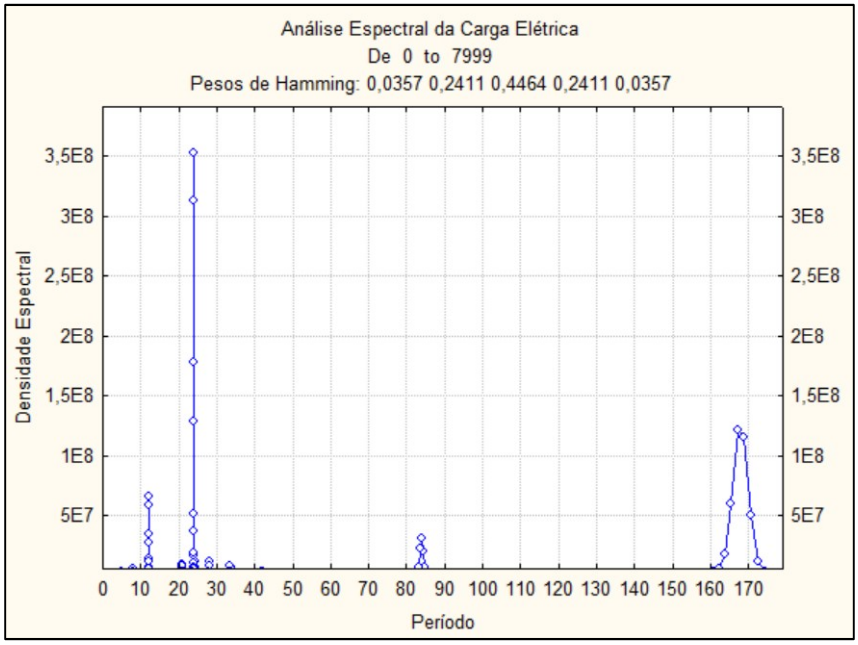

Figura 5.25 – Detalhe do diagrama de densidade espectral da série Duke8, evidenciando os picos principais.

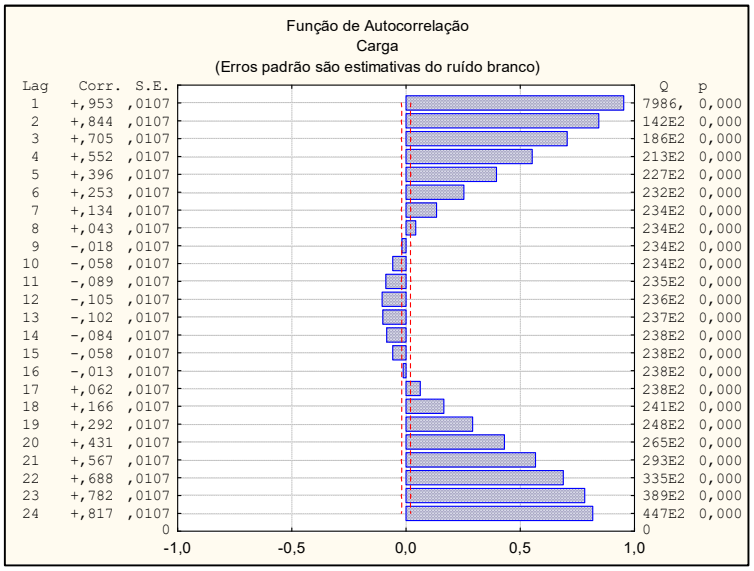

Figura 5.26 – Diagrama em barras da Função de Autocorrelação da série Duke8

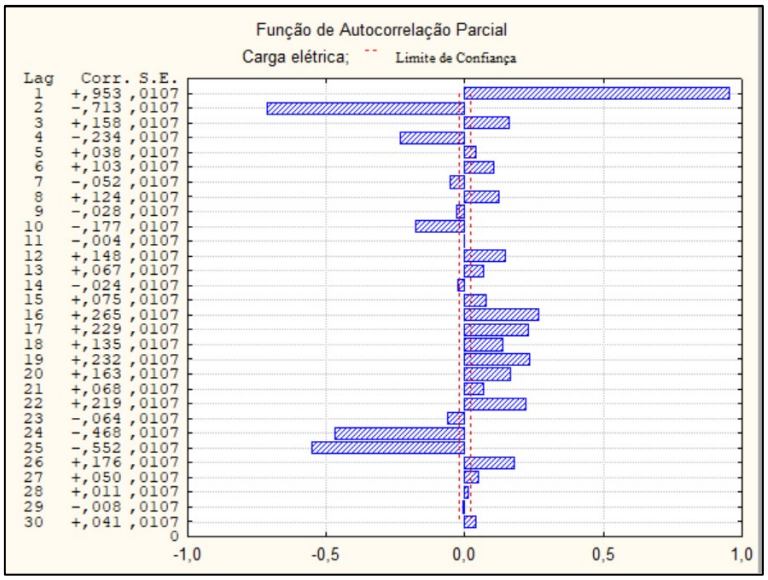

Figura 5.27 – Diagrama em barras da Função de Autocorrelação Parcial da série Duke8, que evidencia a presença do período de 24 horas e fornece indicação da presença de componentes de autorregressão e médias móveis, úteis para ajustes de modelos SARIMA.

Dadas as características de variabilidade e abrangência temporal da série, não foi realizado estudo de tendência ou sazonalidade.

5.4.3. Adequação do caso

Com base nos diagramas de Densidade Espectral e Periodograma, foram realizados experimentos exploratórios com *lag* = 12, 24, 84 e 168 e os parâmetros fixos do

a fim de avaliar a influência desta variável sobre a qualidade do treinamento das RNAs. Os, resultados propiciaram a escolha de SP = 168 (v. Tabela 4.11). O mosaico da Figura 5.28 corrobora a escolha do período de 168 horas como aquele a ser usado neste trabalho.

Tabela 4.11 – Desempenho das RNAs em função do tamanho da amostra para treinamento

Lag	**Período**	**Coeficiente de regressão**	**Erro**
12	Meio dia	0,175	0,032
24	Dia	0,146	0,026
84	3,5 dias	0,104	0,019
168	7 dias = 1 semana	0,987	0,018

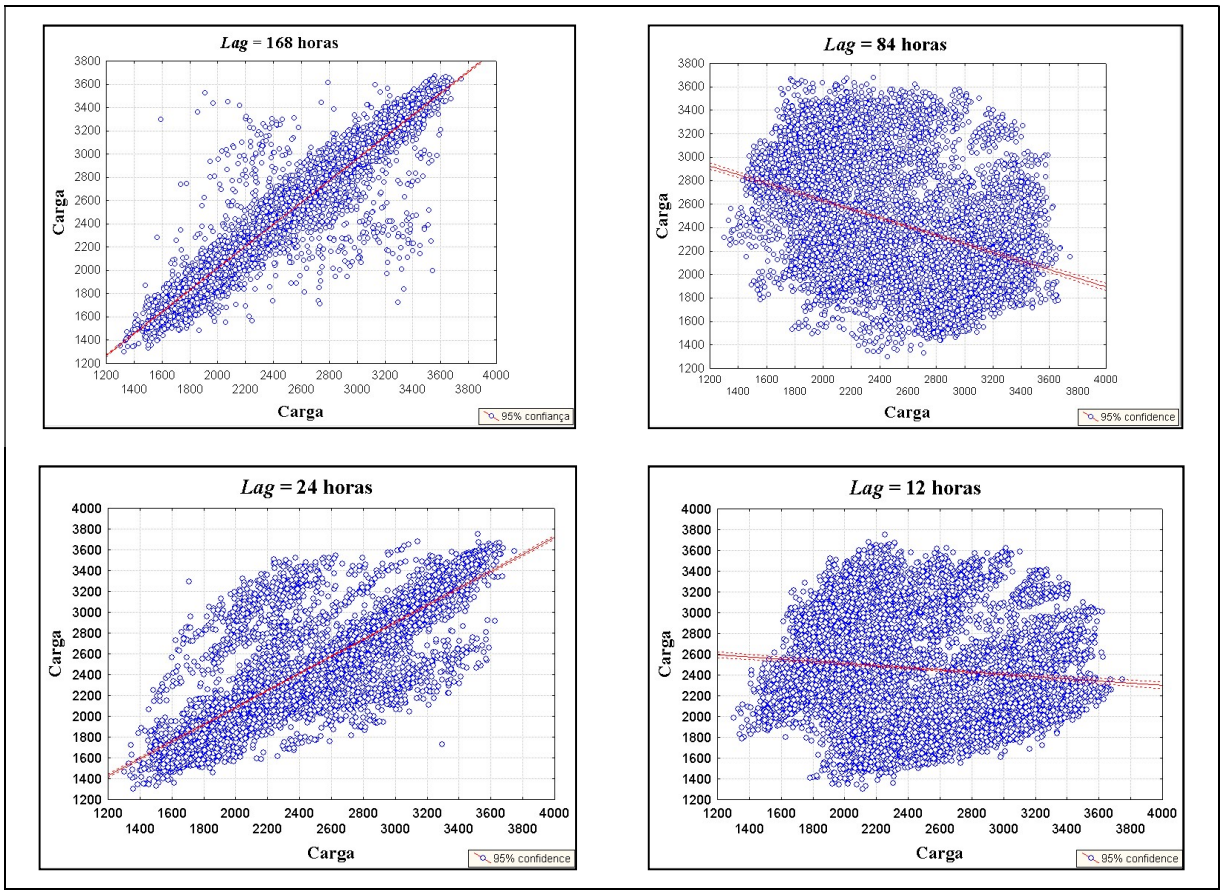

Figura 5.28 – Diagramas de carga *versus* carga (em MW) com diferenciação indicada nos títulos individualmente. As linhas em vermelho representam correlação de cada distribuição. Embora haja picos em 12 e 84 horas na PACF, a distribuição circular dos pontos comprovam que estas não são significativas, afinal.

5.4.4. Ajuste de modelo SARIMA

Procedeu-se à modelagem SARIMA com base em simulação de Monte Carlo sazonal com medida de erro MAPE. O melhor modelo foi SARIMA(2,0,2)(1,1,1)$_{12}$ com MAPE = 0,99% e parâmetros de ajuste apresentados na Tabela 4.12. A componente sazonal (1,1,1) está em acordo com o comportamento da ACF e PACF.

Tabela 4.12 – Resultados do modelo SARIMA obtidos por simulação de Monte Carlo

Variável	Coeficiente	Erro Padrão
AR(1)	1,9000	~ 0
AR(2)	-0,9053	~0
MA(1)	0,8579	0,0108
MA(2)	0,0995	0,0108
Sazonal AR(1)	0,1393	0,0053
Sazonal MA(1)	0,5994	0,0053

Alguns resultados deste modelo encontram-se na Figura 5.29, juntamente com os dados históricos, para fins de comparação. Considera-se que este modelo serve ao propósito de *benchmarking*.

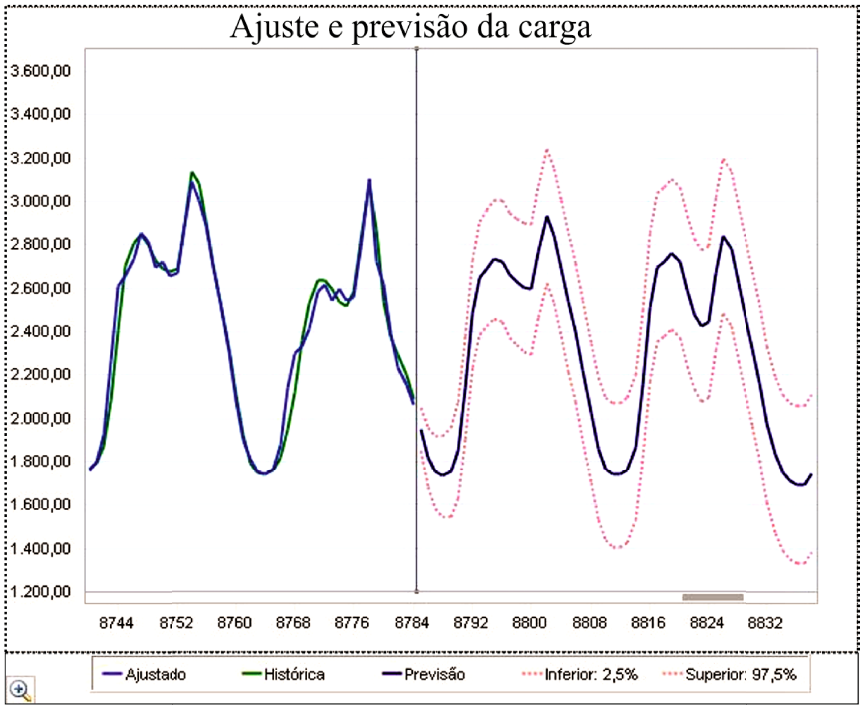

Figura 5.29 – Melhor modelo SARIMA para a série Duke8: comparação com a série original e a previsão (carga em unidades de MW *versus* No. de horas).

5.4.5. Aplicação da metodologia PBCA

Devido à sazonalidade de 168 horas, escolheu-se o número de medidas usadas para previsão também de 168 (SP). A Tabela 4.13 contém as medidas do erro de cada RNA empregando os fatores fixos e a combinação respectiva dos variáveis. A ordem dos experimentos é aleatória. A Figura 5.30 apresenta a distribuição de MAPE *versus* MdAPE. Os menores valores de MPAE e MdAPE foram apresentados pelo mesmo experimento, o de No. 17. A título de ilustração, a Figura 5.31 contém porções da previsão do experimento No. 17.

Tabela 4.13 - Resultados das RNAs para uso com a PBCA

Ordem	MAPE [%]	MdAPE [%]
9	3,4	2,5
12	4,5	3,6
21	4,3	3,2
7	2,6	1,8
13	3,3	2,4
18	2,7	1,9
2	4,4	3,1
11	3,5	2,7
1	3,3	2,3
5	7,1	5,7
23	11,1	9,6
14	8,4	6,8
6	3,0	2,2
24	7,3	5,9
10	8,1	6,6
15	2,4	1,6
17	2,1	1,5
8	8,4	6,8
16	8,6	7,1
22	2,4	1,6
19	2,8	2,0
4	6,9	5,4
3	2,4	1,7
20	3,1	2,2

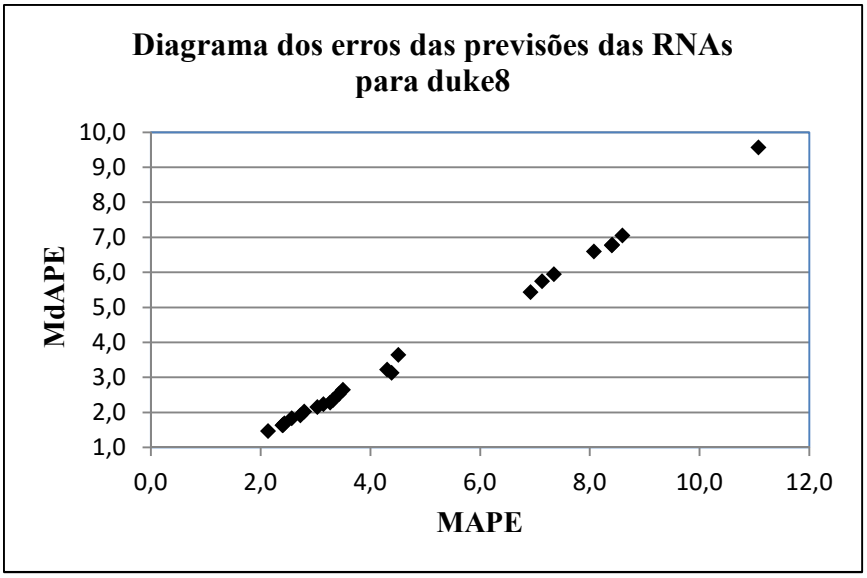

Figura 5.30 - Diagrama de dispersão entre as medidas de erro (%) do desempenho das RNAs.

Tabela 4.14 - Combinação dos valores dos parâmetros que caracterizam o experimento de No. 17.

HL	UL	P2	LR	SC	ET	W1	W2	SM	MAPE	MdAPE
1	2	BFGS	0,01	0,0	0,0	Yes	No	Rand	2,1 %	1,5 %

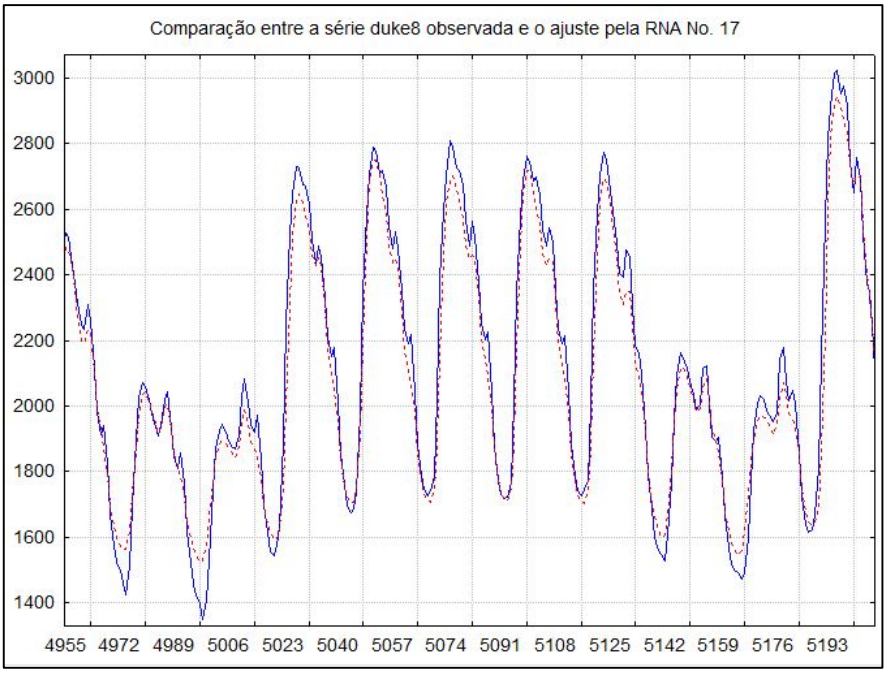

Figura 5.31 – Diagrama de carga em MW *versus* No. de horas contendo um trecho da série Duke8 (em azul) com sobreposição do resultado da RNA de No. 17 (em vermelho).

O ajuste da série compreendendo o final do ano de 2008 e o começo de 2009 encontra-se na Figura 5.32 abaixo:

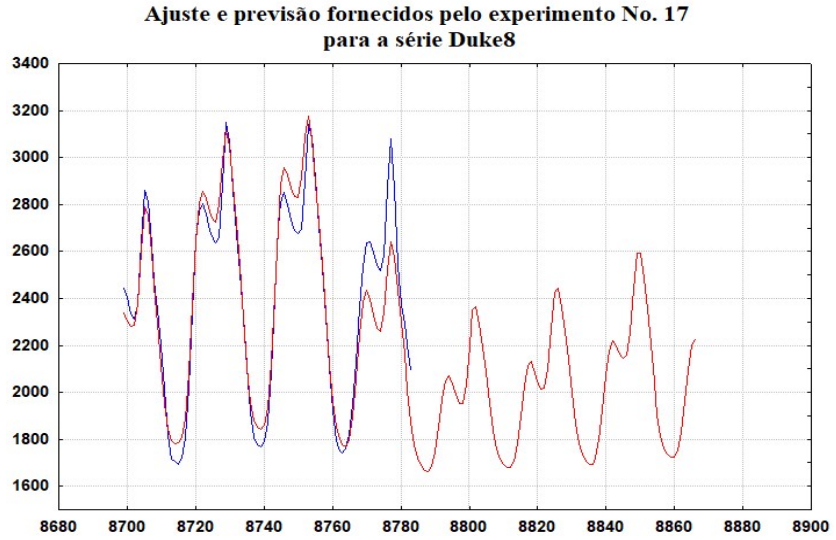

Figura 5.32 – Previsão da carga (MW) para as primeiras horas de 2009 dada pela RNA No. 17.

A Figura 5. 33 contém os resíduos do modelo de RNAs No. 17. O diagrama de densidade espectral em função do período não apresentou nenhum pico proeminente, o que significa que o modelo ajusta a série temporal sem deixar vestígios de sazonalidade. O histograma é unimodal e simétrico em torno de zero, o que significa que o ajuste não apresenta tendenciosidade notável.

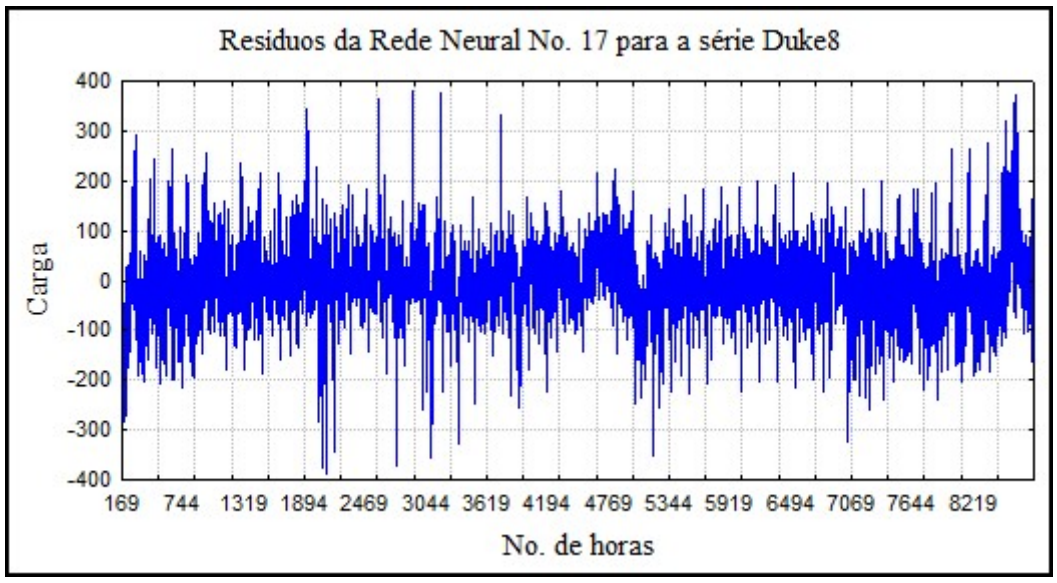

Figura 5. 33 – Resíduos do ajuste da Rede Neural No. 17 aos dados da série de carga elétrica Duke8 (MW).

A Tabela 4.15 da seção 5.5 traz os resultados da aplicação da metodologia PBCA ao fatores (Fase 1) e aos quatro pares de interação de segunda ordem pré-escolhidos, conforme explicitado no Capítulo 4 (Fase 2). Apesar da delimitação do estudo dos pares de interações de ordem 2, para esses quatro pares, comparando-se os valores de MVPA para os casos de MAPE e de MdAPE resultantes da aplicação da PBCA em ambas as fases, nota-se que ambos diminuíram de valor. Isto indica a presença de interações significativas (COUTO, 2012).

5.5 . Análise dos resultados

A análise exploratória de ambas as séries permitiu uma visão ao mesmo tempo global e também mais detalhada de seu comportamento e propriedades estatísticas. Parte das escolhas de parâmetros e de decisões sobre como melhor usar essas característica para se chegar à melhor aplicação da PBCA baseou-se nessas preliminares. Exemplo disto são os gráficos das PACFs e das densidades espectrais, que evidenciaram os principais períodos de cada série.

A série OPD, que a princípio apresentou-se como de simples comportamento com sazonalidade anual, revelou-se a mais difícil de se modelar. Já a série Duke8, interessante por suas sazonalidades e médias mutantes, sua variação diária, semanal e anual, terminou por ser a mais "domesticável".

O estudo aprofundado de cada série, embora tentador, foge ao escopo deste trabalho e portanto, não se estendeu mais.

O recurso de providenciar *benchmarks* através de modelagem por Monte Carlo e pela "tradicional" SARIMA foi oportuno e, juntamente com a modelagem de cada RNA da parte de DOE por PB com rebatimento, serviu de baliza para os resultados da PBCA. Estes resultados foram apresentados nas Figuras Figura 5.17, Figura 5.21 e Figura 5.31.

As previsões dos experimentos para o ano de 2017 para a série OPD foram concordantes (Figura 5.21), mas as previsões da série Duke8 não puderam ser comparadas a dados reais pela falta dos mesmos e apenas a comparação com o ano de 1988 foi possível (v. Figura 5.32).

Os resultados das PBCAs das Fases 1 e das Fases 2 das séries temporais OPD e Duke8 estão elencados na Tabela 4.15. No caso da série Duke8, nota-se o decréscimo no valor da MVPA tanto para MAPE como para MdAPE. Já o caso da série OPD não apresenta decréscimo evidente. Dada a falta de erros nas medidas de Duke8 e a adoção de um erro geral de 0,5 hora

para OPD, além da falta de um preceito para a estimativa e/ou propagação de erros na metodologia PBCA, não se pode quantificar esse decréscimo (ou sua falta) em termos estatisticamente significativos.

Tabela 4.15 – Medidas da qualidade dos melhores resultados das RNAs dentro da metodologia PBCA.

PBCA da Fase 1: apenas fatores					
Série	Estatística	No. do resíduo	Linha da Matriz 2	Fatores	MVPA
OPD	MAPE	211	110010110	HL, UL,SC,W1, W2	0,962
OPD	MdAPE	509	101111111	HL, P2, LR, SC, ET, W1, W2, SM	0,966
Duke8	MAPE	211	110010110	HL, UL,SC,W1, W2	0,955
Duke8	MdAPE	211	110010110	HL, UL,SC,W1, W2	0,956
PBCA da Fase 2: fatores e interações					
Série	Estatística	No. resíduo	Linha da Matriz 2	Interações	MVPA
OPD	MAPE	52	001011	HL*W2,SC*W2, W1*W2	0,966
OPD	MdAPE	31	111110	HL*SC, HL*W1, HL*W2, SC*W1, SC*W2	0,974
Duke8	MAPE	46	011101	HL*W1, HL*W2, SC*W1, W1*W2	0,873
Duke8	MdAPE	46	011101	Id.	0,869

No tocante à PBCA da Fase 1, é interessante notar que ambas as séries temporais indicam basicamente o mesmo conjunto de fatores da arquitetura de RNAs como sendo os que melhor as caracterizam. Este resultado, dentro do universo dos experimentos, independe do tipo de estatística empregado para medir a qualidade do ajuste, MAPE ou MdAPE.

No tocante à PBCA da Fase 2, apenas Duke8 apresenta resultados coincidentes entre si, independentemente da estatística considerada. Já a série OPD apresenta apenas duas interações em comum entre ambas as estatísticas.

Um olhar mais atento às colunas dos Fatores e Interações da Tabela 4.15 mostra outro resultado interessante: as melhores RNAs para os problemas aqui estudados podem ser caracterizadas pelos parâmetros HL, UL, SC, W1 e W2, e o experimentador não precisa levar em conta a interação entre HL e SC. Assim sendo, pelo menos um parâmetro de cada grupo classificatório do Quadro 4. 3 é relevante à modelagem. Este resultado pode ser entendido da seguinte forma: HL e UL definem a estrutura básica da RNA, SC determina a parada da modelagem e evita a superaprendizagem, W1 eW2 regulam o trânsito da informação entre os

neurônios, e os parâmetros restantes são ruído de fundo e podem assumir valores-padrão fixos.

A série temporal do OPD parece constituir um problema intrincado ("*Wicked Problem",* termo cunhado por Rittel e Webber em 1973, em políticas sociais, e revisitado por Crowley e Head em 2017). Ao longo de 37 anos, vários técnicos colheram os dados sem metodologia científica, baseados nos diários das observações feitas por pesquisadores, estudantes de pós-graduação e técnicos, sem seguir protocolo específico de descrição do aproveitamento do tempo (avaliação esta também dependente do tipo de projeto executado), com sono e/ou cansados − isto demonstra a forte influência subjetiva do fator humano e da falta de rigor científico. A partir de 2006, há registros de uma estação meteorológica no sítio, e isto fornece mais variáveis incontroláveis ao problema; a falta de um banco de previsões climáticas com o qual manter um registro para fins comparativos, um levantamento cruzado dos diversos tipos de instrumentação e melhorias de infraestrutura, hardware e software com data de entrada em funcionamento e dos tipos de projetos desenvolvidos no OPD também contribui para a parca caracterização do problema. O "mau comportamento" desta série, no tocante às tentativas de modelagem contribui positivamente para as conclusões apresentadas no Capítulo 6.

Na verdade, o próprio problema de modelagem da arquitetura de uma RNA previsora aplicada a séries temporais univariadas não lineares é um *wicked problem*. Embora o pesquisador lance mão de vários recursos estatísticos e formule hipóteses bem fundamentadas teoricamente acerca da modelagem de um fenômeno, seja ele mercadológico, natural ou produtivo, a impossibilidade de introduzir as possíveis interações entre as variáveis em algum ponto da modelagem da RNA com os recursos computacionais aqui disponíveis deixa inacabada a completa aplicação da metodologia.

A PBCA tem aplicação típica em processos de modelagem que permitem a consideração das interações que porventura sejam identificadas e/ou previamente conhecidas. A identificação do modelo matemático que descreve um determinado processo e que leva o experimentador ao uso ótimo dos recursos disponíveis tampouco pode ser aplicada a uma RNA devido à forma como é construída pelos pacotes computacionais usados e como a mesma se comporta.

Relembrando o Objetivo desta tese (seção 1.2), o qual é implementar a metodologia PBCA a RNAs e a este tipo de séries temporais, se a diminuição do valor de MVPA para MDAPE e MdAPE da série Duke8 for considerado, então, sim, a metodologia é aplicável com bom resultado. Se, por outro lado, a falta de uso direto das interações entre as variáveis constituir-se em um revés, então a conclusão possível é que a PBCA não se presta a este tipo de

problema. Em contraponto, os resultados da PBCA para a série OPD demonstraram que interações de segunda ordem entre os fatores não são necessárias, caso este em que o uso apenas dos fatores já proporcionam RNAs boas modeladoras – o que também foi previsto por Couto (2012) na proposta da metodologia.

Existiria algum uso para o conhecimento da existência das interações neste caso? *Plackett_Burman* é um DOE para redução do número de variáveis de um problema com base no efeito, na significância das mesmas e na sensibilidade do fenômeno às mudanças impostas a essas variáveis. Esta informação poderia ajudar o pesquisador a tomar decisões ainda na fase de testes exploratórios, sobre metodologias e/ou técnicas mais adequadas para a melhor modelagem do problemas de interesse.

5.6 . Considerações finais

Neste capítulo foram caracterizados dois casos de séries temporais que foram escolhidas devido às suas diferentes naturezas: distribuição de horas úteis em um observatório astronômico e distribuição de carga elétrica. Foram apresentados os resultados tanto de estratégias exploratórias como da aplicação da metodologia PBCA às mesmas. Apesar das poucas características em comum das duas séries, os resultados apontaram para o mesmo conjunto de variáveis significativas em ¾ dos casos. Para a série elétrica, a diminuição do parâmetro MVPA ao se considerar a existência de interações entre algumas variáveis comprova a eficiência da metodologia, mesmo sem a possibilidade de uso desta informação na especificação da arquitetura das RNAs.

Demonstrou-se, aqui, que a total aplicação da PBCA não é possível neste tipo de problema, qual seja, a definição de parâmetros arquitetônicos de RNAs, do tipo MLP, previsoras de séries temporais univariadas não lineares.

No Capítulo 6 são apresentadas algumas possibilidades atuais e promissoras de aplicação da PBCA em toda sua extensão em outros campos do conhecimento. A questão das RNAs previsoras mereceu comentários à parte.

6. Conclusões e perspectivas

6.1 .Considerações gerais

Nos Capítulos anteriores, foram apresentados: a motivação do trabalho, a metodologia, as inovações da proposta, os benefícios tanto para a indústria, mercado e outros, como para a academia. As etapas da condução da pesquisa, os casos escolhidos para estudo e aplicação da metodologia, os resultados e sua análise foram detalhados de modo a poderem ser reproduzidos. Alguns comentários a respeito dos resultados já foram adiantados no final do Capítulo 5. Aqui dá-se o fechamento das conclusões e a apresentação de possíveis trabalhos futuros relacionados à tese (seção 6.2).

As principais conclusões desta pesquisa são:

- A PBCA pode ser aplicada com sucesso na identificação dos parâmetros arquitetônicos mais significativos de Redes Neurais Artificiais do tipo *Multilayer Perceptron*.
- Embora a PBCA indique a existência de interações de segunda ordem entre alguns parâmetros e que, portanto, devam ser consideradas no estudo, essa informação não pode ser devidamente utilizada com o *setup* de *hardware* e de *software* desta pesquisa. Consequentemente, todo o potencial da metodologia PBCA não pode ser explorado na definição das Redes Neurais Artificiais do tipo *Multilayer Perceptron* dentro das condições experimentais desta tese.

Com relação ao caso da série temporal das horas noturnas úteis do Observatório do Pico dos Dias, de um modo geral, as previsões para o ano de 2017 foram similares aos dados experimentais no que diz respeito à sazonalidade. Neste caso, a Comissão de Programas do Observatório pode seguir com os procedimentos e critérios de distribuição de tempo de telescópio usados até o momento.

Um resultado interessante é que, de acordo com o ajuste de tendência (*detrending)*, há um aumento no número de horas úteis de cerca de 5 minutos ao mês, o que implica em quase 1 hora ao ano. Consequentemente, a cada 10 anos, há um aumento de 10 horas, ou seja, aproximadamente 1 noite em cada um dos cinco telescópios. Dependendo dos tipos de projetos executados, essa “noite” adicional pode permitir a publicação de pelo menos 5

artigos em periódicos arbitrados, baseados nas observações desse tempo extra, mesmo diluído ao longo dos meses.

Um estudo como este pode, também, auxiliar na tomada de decisões a respeito de se investir no desenvolvimento de nova instrumentação para o OPD a curto e médio prazos, a respeito de novas políticas sobre os melhores tipos de projetos a serem contemplados com tempo de telescópio segundo critérios de custo-benefício em ciência.

Com relação ao caso da distribuição de carga elétrica da Duke Energy, a PBCA indica a existência de algumas interações entre os parâmetros arquitetônicos das Redes Neurais MLP empregadas, através da diminuição do coeficiente de correlação entre as séries de resíduos. Isto é exatamente onde reside a força da metodologia PBCA, segundo Couto (2012): com a realização de apenas um Delineamento de Experimentos, todos os parâmetros e suas interações de segunda ordem significativas são identificados. Teria sido útil e complementar a este estudo a comparação com o restante da série, do ano seguinte, mas estes dados não estiveram disponíveis. Fica, então, fortemente sugerida a aplicação da PBCA aos casos de séries de potência e distribuição de carga elétrica, desde que outros tipos de Inteligência Artificial e/ou técnicas estatísticas de análise que permitam o aproveitamento dos resultados da PBCA sejam empregadas conjuntamente. Uma possibilidade promissora é associar a PBCA à gestão de carteiras de contratos (ARCE, 2014) sob a luz da Teoria de Portfólios proposta por Markowitz (1952).

Esta última sugestão se aplica, também, ao caso do OPD, para o qual outras técnicas de análise e previsão poderiam sair-se melhores que a PBCA. A Gestão de Portfólios aplicada à instauração do modo fila de observações, aliada a uma inteligência artificial que empregasse a PBCA como parte de seu acervo de recursos, poderia garantir o retorno científico das operações do Observatório, ao mesmo tempo em que diminuiria os riscos de desperdício de tempo e recursos financeiros, humanos e de infraestrutura.

Um subproduto desta tese é a constatação de que os mecanismos de busca de palavras-chave nas bases nacionais e internacionais de publicações podem fornecer resultados com viés. Por exemplo, as classificações de áreas do conhecimento usadas pelas bases são por vezes restritas, provavelmente pela necessidade prática de lidar com uma extensa gama de assuntos. Já as palavras-chave fornecidas pelos autores compõem conjuntos mais maleáveis e por vezes subjetivos, por serem constituídas de palavras que mais se coadunam com os critérios de busca consciente- e/ou inconscientemente utilizados por pessoal acadêmico ou não. Este trabalho sugere que, cada vez mais, não apenas os programadores das máquinas de busca, mas

também os usuários adquiram experiência em mineração de dados e técnicas de análise de grandes volumes de informação.

6.2 . Considerações finais e trabalhos futuros

Este trabalho contribui para o acervo de recursos estatísticos para a redução do número de parâmetros em modelagem de fenômenos naturais, mercadológicos, financeiros, empresariais e em Pesquisa e Desenvolvimento (P&D). Também é útil na identificação da existência (ou não) de interações entre esses parâmetros, o que auxilia o pesquisador ou técnico a não cometer erros do Tipo II, quais sejam, deixar de lado interações por não considerá-las importantes dentro de seu problema de interesse quando, na verdade, elas o são, ou por sequer ter noção de sua existência.

Propõe-se a seguir, então, algumas linhas de pesquisa e ações que seriam interessantes como desdobramentos e/ou continuação desta tese:

- Confeccionar o código completo da PBCA em R ou Python, por exemplo, e utilizar um cluster como o do Laboratório Nacional de Astrofísica/MCTIC, constituído por um servidor Dell Poweredge R910 com 140 processadores de 1,87 GHz e 124 GB de memória operando com sistema Fedora 19.
- Verificar se outras séries temporais uni- e multivariadas não lineares sendo estudadas através da PBCA com RNAs do tipo MLP teriam como significativos os mesmos parâmetros e interações de segunda ordem. Isto poderia vir a caracterizar as séries e/ou as RNAs nesse tipo de problema.
- Explorar o potencial da PBCA selecionando problemas com interações conhecidas previamente e/ou esperadas.
- Explorar o potencial da PBCA em problemas de modelagem que exigem mineração de dados (*Data Mining*) e que requerem o uso de inteligências artificiais capazes de aprendizagem profunda (*Deep Learning*), especialmente nas áreas de processos químicos e biotecnologia, os quais se caracterizam pelo elevado número de dados, de interações entre as variáveis e custo computacional.

Muitos desses problemas certamente apresentam alto nível de complexidade, mas, por isso mesmo, merecerão ser estudados com esta metodologia. Isto devido ao fato de que, segundo os resultados desta tese, a PBCA é capaz de identificar as variáveis mais significativas de um problema de séries temporais não lineares e, ao mesmo tempo, acusar de forma inequívoca a não consideração de interações de segunda ordem. Ademais, a realização de apenas um Delineamento de Experimentos faz dela um poderosa e econômica ferramenta na análise de diversos fenômenos mercadológicos, financeiros e naturais.

APÊNDICE A – Sobre a pesquisa

A.1. Caracterização da pesquisa

A.1.1. Classificação

Trata-se de pesquisa de natureza aplicada, com objetivo normativo, cuja abordagem é quantitativa e o método é modelagem e simulação (MIGUEL, 2012).

A.1.2. Modelo

O modelo proposto por Mitroff *et al.* (1974) para este método é o fio condutor desta pesquisa. O modelo é dividido em quatro etapas: conceitualização, modelagem, solução do modelo e aplicação, e as mesmas estão ilustradas na .

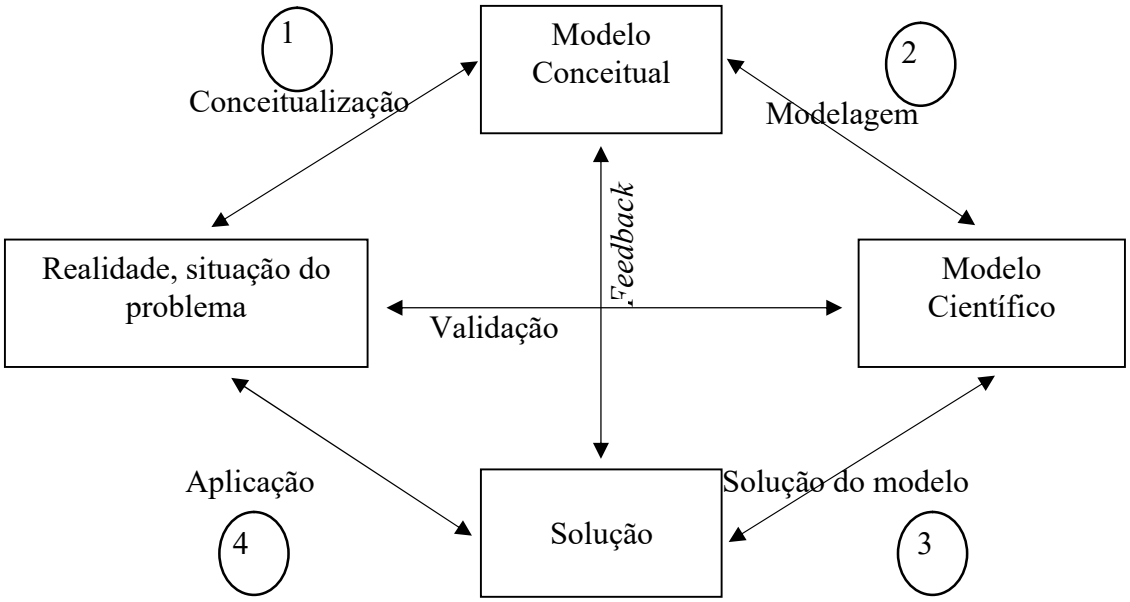

Figura A.1 - Etapas do método de pesquisa a ser seguido: Modelagem e Simulação.
Fonte: Amorim (2018), numa adaptação de Mitroff *et al*. (1974)

É na etapa de conceitualização que se definem as variáveis do modelo conceitual, o escopo e demais condições de contorno iniciais intrínsecas ao problema. A etapa de modelagem é

quando se constrói o modelo matemático-científico que define as relações causais entre as variáveis, este modelo traduz algebricamente o modelo conceitual. Uma vez de posse do modelo algébrico, por assim dizer, pode-se proceder à sua solução e otimização, convergindo para o modelo final, pronto para ser aplicado, se assim for desejado, a um problema real, o que ocorre na etapa seguinte. Nesta fase de aplicação, por fim, soluciona-se o problema inicial e finaliza-se o ciclo; note-se, porém, que o ciclo também pode ser reiniciado caso ajustes sejam necessários, a critério do experimentador.

Bertrand e Fransoo (2002) afirmam que a definição das etapas a serem realizadas depende do tipo da pesquisa pretendida. Amorim (2018) esclarece que:

> Nas pesquisas quantitativas axiomáticas descritivas, o pesquisador cria apenas o modelo conceitual e o modelo científico. Já nas pesquisas quantitativas axiomáticas normativas, o pesquisador parte do modelo conceitual para obter o científico e o soluciona; o modelo conceitual pode ou não ser realimentado para que sejam feitas as devidas alterações ou contribuições. Nota-se que nas pesquisas de classe axiomática não existe a preocupação de aplicar a solução proposta no modelo real.

Uma pesquisa é empírica quando há o retorno ao fenômeno e/ou processo real. Se, ademais, ela for quantitativa e descritiva, é necessário proceder-se à validação através do confronto dos os resultados do modelo matemático com o comportamento e medidas reais. Nos casos semelhantes a este projeto, que se caracteriza por ser do tipo quantitativo empírico normativo, todo o ciclo proposto originalmente para o método modelagem e simulação. Na estão ilustradas as etapas a serem conduzidas no estudo de acordo com a classificação de cada pesquisa (AMORIM, 2017).

Assim sendo, este projeto, o qual, no quesito de modelo, usará modelagem e simulação, também se caracteriza como sendo empírico normativo.

A.2. Justificativa da escolha

A aplicação de um método inédito (PBCA) na solução de um problema prático comum em Engenharia de Produção (identificação de modelo ótimo sem desperdícios) precisa ser validada (modelos matemáticos sintéticos e aplicação a caso real já estudado) antes de estabelecer-se como recurso de valor.

Escolheu-se modelar a arquitetura interna de RNAs devido ao seu crescente emprego na solução de problemas nas mais diversas áreas do conhecimento (v. Seção 2.3), especialmente naquelas de alta complexidade e consequente dificuldade (ou até impossibilidade) de representação paramétrica (TIROZZI, BRUNELLO *et al.*, 2006; HSIEH, 2004). A habilidade

e vantagem da PBCA em identificar variáveis e suas interações de segunda ordem que descrevam o modelo técnico-científico de um processo qualquer com economia de recursos (COUTO, 2012) tornam-na interessante nesses casos.

Uma vez validada a metodologia, não há porque não aplicá-la imediatamente a casos que potencialmente podem se beneficiar dessa abordagem. Selecionaram-se dois problemas tático-estratégicos que envolvem a previsão de comportamento temporal em série a curto e médio prazos. O primeiro é o auxílio à organização noite a noite e minuto a minuto de observações astronômicas com telescópio em terra, onde decisões dependem das restrições impostas pela ciência pretendida e das condições atmosféricas e reinantes ao longo da noite. Uma RNA dedicada, baseada em dados históricos de medidas de desempenho instrumental e de condições de céu, é o primeiro passo em direção à automatização dessa organização. O segundo caso é o auxílio à identificação de possíveis cenários futuros de melhor uso de sítio astronômico e consequente recomendação de estratégia de sobrevivência com base em dados históricos de medidas de horas úteis.

O projeto de pesquisa aqui proposto e acima classificado, então, requer a execução de todas as quatro etapas do modelo proposto por Mitroff *et al*. (1974).

A.3. Procedimento metodológico adotado

Parte dos aspectos e cuidados que envolvem uma experimentação já foi introduzida no Capítulo 2. Vale reiterar aqui a vantagem da metodologia proposta sobre outros métodos de seleção de variáveis em modelos computacionais que envolvem dezenas ou centenas delas e as relações entre as mesmas são complexas e não lineares (v. SCHONLAU e WELCH, 2006).

Neste trabalho, têm-se preocupações e cuidados como, por exemplo: escala de esforço, precisão e poder, amostragem e aspectos similares às análises quase econômicas e de como planejar a robustez do método.

Cox e Reid (2000) enfatizam que o planejamento deve estimar apropriadamente o esforço a ser exigido pela busca de uma solução para o problema de interesse, seja estimando-se tempos de processamento, custos, tamanho e forma de coleta da amostra, etc. O estudo dos contrastes calculados no DOE leva a estimativas aproximadas da variância alcançada, e o cálculo do poder de um teste de hipótese sobre a nulidade deles fornecem essencialmente a mesma informação sobre o andamento − neste caso – da definição da topologia da RNAs simulada

com o objetivo de parar o processo. A escolha da amostra e da metodologia envolve o balanço energético entre os custos e as perdas oriundas da imprecisão das conclusões; se for possível expressar esses parâmetros quantitativamente, então uma solução ótima (robusta, aqui) pode ser encontrada; tratar-se-ia, então, de uma pesquisa orientada à decisão.

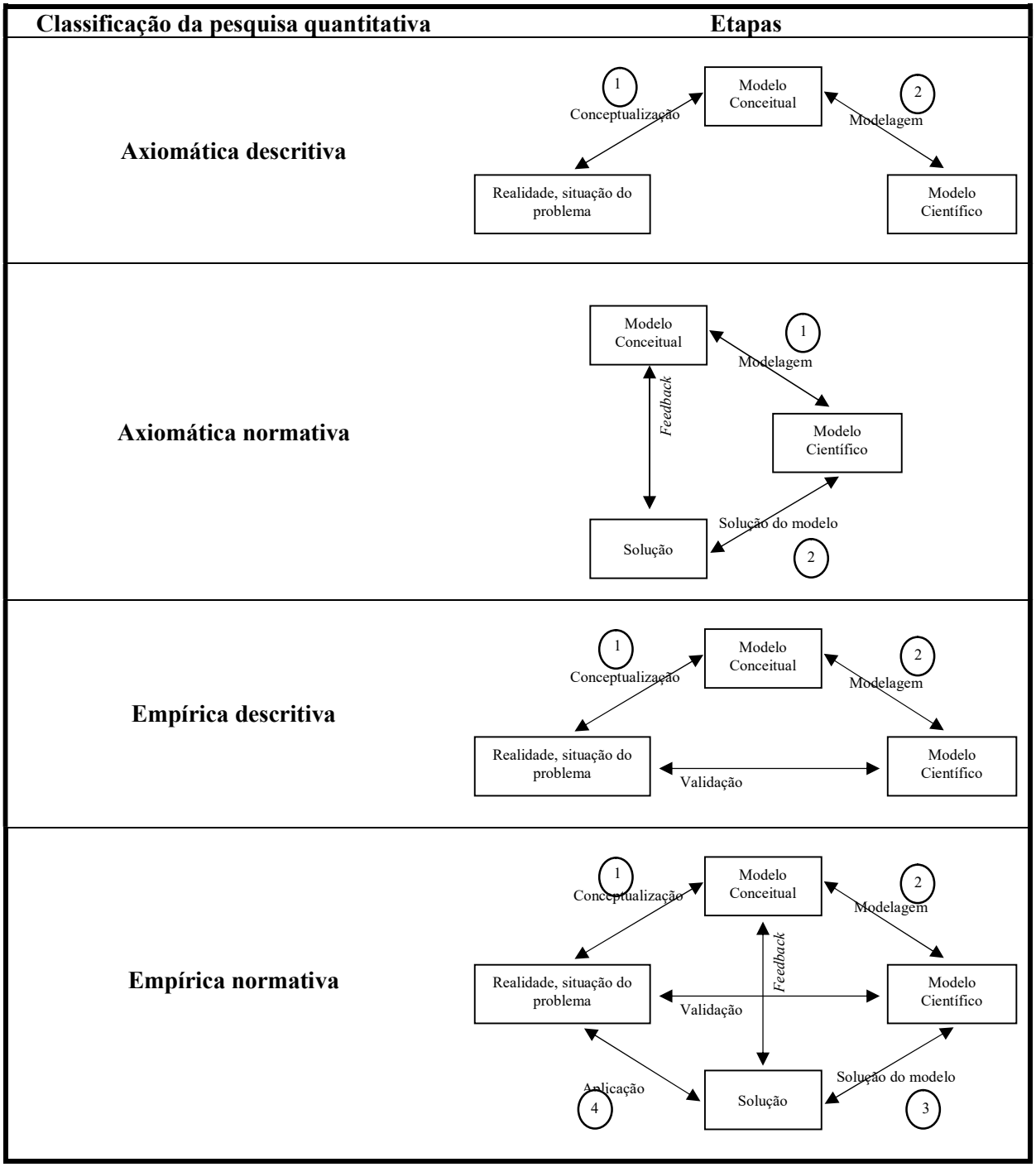

Figura A.2 - Classificação da pesquisa científica quantitativa com as etapas do método modelagem e simulação
Fonte: Amorim (2018)

A robustez da metodologia merece atenção especial. Este é um conceito ligado aos experimentos fatoriais de Taguchi, proposto originalmente visando a melhoria da produção e a satisfação do cliente através da consideração cuidadosa dos fatores incontroláveis e dos custos com falhas (PHADKE, 2013). Aqui, a preocupação é com a magnitude da variabilidade dos resultados frente à existência de erros (os aleatórios e os prováveis caóticos) levando-se em conta o grande número de variáveis de entrada de parâmetros das RNAs.

APÊNDICE B - Pseudocódigo do passo a passo dos cálculos

Autoria: Mariângela de Oliveira-Abans

Procedimento
* Preparar o arranjo experimental PB com rebatimento no Scilab® *\
* Usar Scilab, Statistica, Excel e/ou Minitab, por exemplo
* Saber de antemão (i) o número de fatores para o DOE (*k)*, (ii) se prefere arranjo geométrico *\
* ou não, e (iii) o número de interações de ordem dois que pode usar nos cálculos *\
* Exemplo desta tese: k = 9 e número de experimentos = 24 *\
Executar *o programa em Scilab, respondendo às perguntas e tomando as decisões adequadas*
Se *escolher arranjo não geométrico* Então
    Executar *o Scilab até o fim*
    Executar a Subrotina A * criar os arranjos *\
    Executar a Subrotina B * guardar os arquivos *\
Senão * o arranjo é geométrico *\
    Parar *o Scilab*
    Executar a Subrotina B
Fim do Se
Se arranjo geométrico
    Executar a Subrotina A
    Fazer *análise de regressão e*
    Fazer *ANOVA com a Matrix1, todos os termos e todas as interações de ordem dois*
    Contar *o número de blocos de interações*
    Executar *novamente o Scilab informando o número de blocos de interações acima*
    Executar a Subrotina B
Fim do Se
Procedimento * para quaisquer casos: realização dos experimentos *\
Criar *planilha Excel com os valores máximos e mínimos reais dos fatores*
Criar *coluna com o número de cada experimento, linha a linha do arranjo fatorial*
Criar *colunas vazias para MAPE e MdAPE*
Executar *omódulo de RNAs do Statistica em modo personalizado, linha por linha*
Calcular *MAPE e MdAPE dos resíduos, linha por linha*
Gerar *gráficos das previsões*

```
Guardar os valores ajustados, as previsões, o modelo, a análise de sensibilidade, a regressão
Gerar macro identificando-a pelo número do experimento
Incorporar a macro, gráfico e planilhas num só workbook do Statistica
Denominar este workbook com o número do experimento
Identificar os experimentos com os MAPE e MdAPE mais baixos
Comparar os resultados: há fatores comuns ou mais significativos?
Anotar estes resultados para análises futuras
Procedimento \* PBCA da Fase 1 *\
\* montagem da Matrix1(MAPE) e Matrix1(MdAPE) para PBCA dos fatores *\
Criar planilha com a macro em VBA
Criar aba com o arranjo fatorial codificado em -1s e +1s e uma coluna MAPE ou MdAPE
Denominar essa aba Matriz1
Criar aba com o nome de Residuos
Procedimento \* montagem da Matrix2 para uso com fatores *\
Criar aba com a matriz para os resíduos dos fatores (0s e 1s)
Denominar essa aba Matriz2
Executar a macro para MAPE e depois para MdAPE
Guardar a planilha identificada por k, número de resíduos dos fatores, MAPE ou MdAPE
Identificar , na aba denominada Final, o sinal de menor MVPA, para MAPE ou MdAPE
Anotar o VPPA e MVPA correspondentes
\* Exemplo desta tese: há 511 linhas de combinações dos 9 fatores *\
Identificar a combinação de fatores significativos desse sinal, para MAPE ou MdAPE
Anotar quais são esses fatores para, também, o caso de ser necessário escolher as interações
\* Exemplo desta tese: dos 5 fatores significativos, escolheu-se 4 \*
Procedimento \* PBCA da Fase 2 *\
\* montagem da Matrix1(MAPE) e Matrix1(MdAPE) para PBCA das interações. *\
\* Notar que é a mesma sequência de cálculos para os fatores sozinhos *\
Criar planilha com a macro em VBA
Criar aba com o arranjo fatorial codificado em -1s e +1s e uma coluna MAPE ou MdAPE
Denominar essa aba Matriz1
Criar aba com o nome de Residuos
Procedimento \* montagem da Matrix2 para uso com interações *\
Criar aba com a matriz para os resíduos dos fatores (0s e 1s)
Denominar essa aba Matriz2
Criar tantas colunas quanto o número de fatores e preenchê-las apenas com 1s
Inserir as linhas de 0s e 1s para as interações depois da última coluna de 1s
```

67: * Exemplo desta tese: os 4 fatores escolhidos geraram 6 pares de interações *\
68: * Neste caso, foi necessário executar o Scilab novamente com estes números *\
69: Executar *a macro para MAPE e depois para MdAPE*
70: Guardar *a planilha identificada por k, número de resíduos dos pares ou blocos de interação,*
71: *MAPE ou MdAPE*
72: Identificar *, na aba denominada Final, o sinal de menor MVPA, para MAPE ou MdAPE*
73: Anotar *o VPPA e MVPA correspondentes*
74: * Exemplo desta tese: há 63 linhas de combinações dos 6 interações *\
75: Identificar *a combinação de fatores significativos desse sinal, para MAPE ou MdAPE*
76: Anotar *quais são essas combinações para análise futura*
77: * Exemplo desta tese: das 6 interações, 5 foram significativas *
78: Comparar os 4 resultados: MAPE e MdAPE para os fatores e para os resíduos
79: Decidir *pelos fatores mais significativos e quais podem ser considerados ruído (default)*
80: Decidir *pelo melhor uso da informação sobre as interações de segunda ordem*
81: Comparar *as melhores previsões com os dados reais/experimentais e os de benchmarking, se houver*
82: Fim *dos cálculos (DOE, PBCA dos fatores e PBCA das interações)*
83: Subrotina A
84: * Preparar os arranjos fatoriais *\
85: Usar *k e o número de experimentos dado pelo Scilab*
86: Criar *arranjo fatorial de Plackett-Burman com rebatimento*
87: Guardar o *arranjo em formato Excel habilitado para macros e em formato Statistica*
88: Voltar
89: Subrotina B
90: Se *arquivos tiverem sido guardados em formato .csv* Então
91: Importar *arquivos em formato texto para planilha Excel*
92: Senão
93: Abrir *diretamente no Excel*
94: Fim do Se
95: Guardar *o arquivo relativo aos resíduos dos fatores como Matrix1 (0s e 1s)*
96: Guardar *o arquivo relativo aos resíduos das interações como Matrix2 (0s e 1s)*
97: Voltar

APÊNDICE C - Código em Scilab® para preparo do arranjo PB com rebatimento e matrizes para as regressões e ANOVA

```
// This is file the_one_forcef.sce: an attempt to join all scilab
// codes into one
// Aug. 15, 2017
// Mariângela de Oliveira-Abans
//
// ---------------------------------------------------------------------
// Dec. 14, 2016
// Mariângela de Oliveira-Abans
//
// Number of residues series
//
clc;clear;errclear
//
// Number of parameters to be tested
//
k = input("Please enter number of parameters = ");
//
// Determine the appropriate Plackett-Burman factorial design (PB)
//
N = (k + 1)
//
// Test whether N is a multiple of 4
//
t = 1111
while t ~= 0
   t = pmodulo (N,4);
   if t ~= 0          // Not multiple
      N = (N+1);
   end
end
mprintf('N before foldover = %i', N)
//
// Fold PB over to have the total number of lines/experiments
//
QtRUNS = (N * 2)
// Qtnew = 0
mprintf('   and   \nNumber of PBCA experiments = %i',QtRUNS)
//
// Verify whether the design is geometrical or not:
// Verify whether QtRUNS is a multiple of 2 (geometrical)
// or not (non-geometrical)
//
dum = factor (QtRUNS)    // factorization
le = length (dum)
disp('dum, le = '); disp (dum, le)
for i = 1:le
   if dum(i) ~= 2 then
      geom = %F
      mprintf('\n Your design is not geometrical,')
      mprintf(' \n    so you will have pairs of second-order interactions.')
      forc = input("\n Force a geometrical design? (1=yes or 0=no)")
      disp('change design? = '); disp(forc)
      if forc == 1 then
         if QtRUNS < 16 then
```

122

```
            QtRUNS = 16
            mprintf('\n Qtnew = 16 %i',QtRUNS)
         elseif QtRUNS < 32 then
            QtRUNS = 32
            mprintf('\n Qtnew is 32 %i',QtRUNS)
         elseif QtRUNS < 64 then
            QtRUNS = 64
            mprintf('\n Qtnew = 64 %i',QtRUNS)
         elseif QtRUNS < 128 then
            QtRUNS = 128
            mprintf('\n Qtnew = 128 %i',QtRUNS)
         else
            disp("lost Qtnew")
         end
         mprintf('\n QtRUNS now is really = %i',QtRUNS)
         geom = %T
      else
         disp('QtRUNS unchanged'); disp(QtRUNS)
      end
      //       break
      if geom == %F then
         NPAIRS = factorial(k) / (2 * factorial (k -2))
         printf('\n NPAIRS = %d \n',NPAIRS)
         //       Calculating the number of freedom degrees
         w = QtRUNS - k - 2
         mprintf('\n There are %i degrees of freedom.',w)
         // Caution: maybe cannot study all interactions
         poss = w - k
         mprintf(' You can only study %i second-order interactions!', w)
         mprintf('\n If you are using MSExcel, you can only study %i
interactions due to memory size',poss, '\n')
         //       break  // first non-2 prime
         break
      else
         end
      else
         if i == le then
            geom = %T
            // all elements = 2
            mprintf('\n Your design is geometrical,')
            mprintf('\n   so you will have blocks of second-order interactions.')
         end
      end
   end
   // pause
   //
   // Stop this code.
   //
   // Choose appropriate column-generating line in Plackett & Burman's
   original
   // paper (1946) or use a comercial software like Statistica(R) to generate
   // the DOE matrix.
   //
   // Make all experiments, run all scenarios and collect data (x) for
   // each combination of variables/levels, storing the responses (y).
   //
   exper = %T
   //
   // dum = factor (QtRUNS)    // factorization
   // le = length (dum)
   //
   dumm = input('\n Have you made all experiments? (1 = yes or 0 = no)  ')
   if dumm == 0
   exper == %F, quit, end
   //
   // You have the experimental or synthetic responses (y)
   // and you have run the factorial ANOVA analysis.
   // So you have identified the blocks of second-order interactions if the
   // design is geometrical.
```

```
//
// Now it is time to computationally run the signal analysis.
//
// Verify whether the design is geometrical or not:
// Verify whether N is a multiple of 2 (geometrical) or not (non-
geometrical)

if geom
   QtBLK = input('Please enter number of blocks from factorial ANOVA=
')
end
//
// Calculate number of factors' residues series to be generated
computationally
//
NRESIFAC = (2^k - 1)
//
// If design is geometrical, calculate the number of block residues series
// If design is non-geometrical, calculate the number of pair residues
// series for w degrees of freedom.
//
if geom then  // geometrical design: QtRUNS is a multiple of four and a
power of two
   NRESIBLK = (2^QtBLK - 1)
else        //   non-geometrical design: N is a multiple of four,
   NRESIPAIR = (2^w - 1)   //   but not a power of two
end
//
// Write output in .csv format
//
// filename = fullfile("C:\Users\Sony\Desktop\", "fig5_out.csv");
// csvWrite(k, filename);    // just for testing purposes
//
// Write in Excel format .xml
//
// filename = fullfile("C:\Users\Sony\Desktop\", "fig5_out.xml");
// writeXmlExcel(filename,k)
// quit
// pause
// ========================================

// Dec. 21, 2016
// Mariângela de Oliveira-Abans
//
// Building of factors' configuration matrix RESIFAC(m,n)
//
printf('\n NRESIFAC = %d \n',NRESIFAC)
RESIFAC01 = zeros(NRESIFAC,k)
//
// Building column by column
//
// column one starts with 1
//
for j = 1:k
   du = (j -1)
   //   printf('\n du = %d \n',du)
   count = (2^du)
   nihil = (count - 1)
   //   printf('\n nihil = %d \n',nihil)
   if j ==1 then                    // column one starts with 1
      for i = 1:NRESIFAC
         if modulo(i,2) == 0 then
            RESIFAC01(i,j) = 0
         else
            RESIFAC01(i,j) = 1
         end
      end
   end
   //
   // j for columns other than the first
```

```
    //
    nex = count
    last = count + nihil
    while last <= NRESIFAC
        for blo = nex:last
            RESIFAC01(blo,j) = 1
        end
        nex = nex + 2^j
        last = last + 2^j
    end
end
//
// Write output in .csv format
//
filename = fullfile("C:\Users\Sony\Desktop\", "fig6_resifac01.csv");
csvWrite(RESIFAC01, filename);    // just for testing purposes
//
// Write in Excel format .xml
//
// filename = fullfile("C:\Users\Sony\Desktop\", "fig6_out.xml");
// writeXmlExcel(filename,RESIFAC)
// quit
// pause
// ============================================
if geom then

    // Dec. 21, 2016
    // Mariângela de Oliveira-Abans
    //
    // Building of blocks' configuration matrix RESIBLK(m.n)
    //           for a geometrical design
    //
    printf('\n QtRUNS = %d \n',QtRUNS)
    printf('\n NRESIBLK = %d \n',NRESIBLK)
    // if forc == 1 then
    //    RESIBLK01 = zeros(NRESIBLK,
    RESIBLK01 = zeros(NRESIBLK,QtBLK)
    //
    // Building column by column
    //
    for j = 1:QtBLK
        du = (j -1)
        //   printf('\n du = %d \n',du)
        count = (2^du)
        nihil = (count - 1)

        if nihil == NRESIBLK then
            printf('\n nihil = NRESIBLK so I stop loop %d \n',nihil)
            break
        end

        //   printf('\n nihil = %d \n',nihil)
        if j ==1 then                    // column one starts with 1
            for i = 1:NRESIBLK
                if modulo(i,2) == 0 then
                    RESIBLK01(i,j) = 0
                else
                    RESIBLK01(i,j) = 1
                end
            end
        end
        //
        // j for columns other than the first
        //
        nex = count
        last = count + nihil
        while last <= NRESIBLK
            for blo = nex:last
                RESIBLK01(blo,j) = 1
            end
```

```
        nex = nex + 2^j
        last = last + 2^j
      end
    end
    // Write output in .csv format
    //
    filename = fullfile("C:\Users\Sony\Desktop\", "fig7_resiblk01.csv");
    csvWrite(RESIBLK01, filename);    // just for testing purposes
    //
    // Write in Excel format .xml
    //
    // filename = fullfile("C:\Users\Sony\Desktop\", "fig7_out.xml");
    // writeXmlExcel(filename,RESIBLK01)
    // quit
    //
    mprintf('Residues series of blocks for geometric design ready')
    break
    // ========================================
else

    // Dec. 23, 2016
    // Mariângela de Oliveira-Abans
    //
    // Building of second-order interactions' configuration matrix
    // RESIPAIR(m.n) for a non-geometrical design
    printf('\n QtRUNS = %d \n',QtRUNS)
    printf('\n w = %d \n',w)
    printf('\n NRESIPAR = %d \n',NRESIPAIR)
    //
    // Building column by column
    //
    if NPAIRS > w then
       mprintf('\n Sorry: you must comply to having %i degrees of
freedom.\n',w)
       newj = input("Please, enter number of second-order interactions less
than or equal to the number of degrees of freedom")
       NPAIRS = newj
       NRESIPAIR = (2^NPAIRS - 1)
       printf('\n New NRESIPAIR = %i',NRESIPAIR)
    end
    RESIPAIR01 = zeros(NRESIPAIR,NPAIRS)

    for j = 1:NPAIRS
       printf('\n j = %d \n',j)
       du = (j -1)
       //    printf('\n du = %d \n',du)
       count = (2^du)
       //    printf('\n count = %d \n',count)
       nihil = (count - 1)
       printf('\n nihil = %d \n',nihil)

       if nihil == NRESIPAIR then
          printf('\n nihil = NRESIPAIR, so I stop loop %d \n',nihil)
          printf('\n Please wait.\n')
          break
       end

       if j ==1 then                    // column one starts with 1
          for i = 1:NRESIPAIR
             if modulo(i,2) == 0 then
                RESIPAIR01(i,j) = 0
             else
                RESIPAIR01(i,j) = 1
             end
          end
       else
          //       if j ~= 1
          //
          // j for columns other than the first
```

126

```
//
//      printf('\n j = %d \n',j)
nex = count
//          printf('\n nex = %d \n',nex)
last = count + nihil
// Caution: if count + nihil for j=12 is larger than
// 2047 available lines, the matrix will have zeroes from here on!
// Then, e.g., the residuals from j=12 to j=55 will not be
// considered in the factorial ANOVA! **************
//

while last <= NRESIPAIR
   for blo = nex:last
      //            printf('\n blo = %d \n',blo)
      RESIPAIR01(blo,j) = 1
      // printf('\n RESIPAIR01 = %d \n',RESIPAIR01(blo,j))
   end

   nex = nex + 2^j
   //             printf('\n new nex = %d \n',nex)
   last = last + 2^j
   //             printf('\n new last = %d \n',last)
         end
      end
      //    mprintf('saiu do if j = 1')
   end

   //
   // Write output in .csv format
   //
   filename = fullfile("C:\Users\Sony\Desktop\", "fig8_out.csv");
   csvWrite(RESIPAIR01, filename);    // just for testing purposes
   //
   // Write in Excel format .xml
   //
   // filename = fullfile("C:\Users\Sony\Desktop\", "fig8_out.xml");
   // writeXmlExcel(filename,RESIPAIR01)
   // quit
   //
   mprintf('\nResidues series of second-order interaction pairs')
   mprintf(' \nfor non-geometric design are ready')
   // pause
   // ========================================
end
```

APÊNDICE D - Macro em VBA para o cálculo de VPPA e MVPA em MSExcel®

Algoritmo: Mariângela de Oliveira-Abans

Código VBA: Júlio Didier Maciel

```
Sub Macro()

    'inicializa Y
'    Sheets("Matrix1").Select
'    Dim Yvalue As Range
'    Range("F1").Select
'    Range(Selection, Selection.End(xlDown)).Select
'    Set Yvalue = Selection

    'conta numero de colunas na matriz 1
    Sheets("Matrix1").Select
    range("A1").Select
    Selection.End(xlToRight).Select
    Dim numCol As Integer
    numCol = ActiveCell.Column - 1

    'conta numero de linhas na matriz 2
    Sheets("Matrix2").Select
    range("A1").Select
    Selection.End(xlDown).Select
    Dim numLin As Integer
    numLin = ActiveCell.Row

    For j = 1 To numLin
        'monta planilha TEMP
        Sheets("Matrix2").Select
        range("A1").Select
        If ActiveCell.Offset(0, numCol).Value <> "" Then
            Do
                ActiveCell.Offset(1, 0).Select
            Loop Until ActiveCell.Offset(0, numCol).Value = ""
        End If

        For i = 0 To numCol - 1
            Sheets("Matrix2").Select
            If (ActiveCell.Offset(0, i).Value = 1) Then
                'checa se planilha existe
                Dim flag
                flag = 0
                For k = 1 To Worksheets.Count
                    If Worksheets(k).Name = "TEMP" Then
                        flag = 1
                        Exit For
                    End If
                Next k
                If flag = 0 Then
                    Sheets.Add After:=ActiveSheet
                    ActiveSheet.Name = "TEMP"
                End If

                Sheets("Matrix1").Select
                Columns(i + 1).Select
```

```
      Selection.Copy
      Sheets("TEMP").Select
      range("A1").Select

      If (ActiveCell.Offset(0, 0).Value = "") Then
         ActiveSheet.Paste
      Else
         If (ActiveCell.Offset(0, 1).Value <> "") Then
            Selection.End(xlToRight).Select
            ActiveCell.Offset(0, 1).Select
            ActiveSheet.Paste
         Else
            ActiveCell.Offset(0, 1).Select
            ActiveSheet.Paste
         End If
      End If
   End If

Next i
Application.CutCopyMode = False

'seta range de X para a regressao
Sheets("TEMP").Select
range("A1").Select
If (ActiveCell.Offset(0, 1).Value <> "") Then
   If (ActiveCell.Offset(0, 1).Value <> "Y") Then
      range(Selection, Selection.End(xlToRight)).Select
   End If
End If
range(Selection, Selection.End(xlDown)).Select
Dim rangeX As range
Set rangeX = Selection

'copia coluna Y
Sheets("Matrix1").Select
range("A1").Select
ActiveCell.Offset(0, numCol).Select
range(Selection, Selection.End(xlDown)).Select
Selection.Copy
      Sheets("TEMP").Select
      range("A1").Select
      If (ActiveCell.Offset(0, 0).Value = "") Then
         ActiveSheet.Paste
      Else
         If (ActiveCell.Offset(0, 1).Value <> "") Then
            Selection.End(xlToRight).Select
         End If
         ActiveCell.Offset(0, 1).Select
         ActiveSheet.Paste
      End If
      'seta range de Y para a regressao - JA ESTÁ SELECIONADO (VER O
SELECTION)
      Dim rangeY As range
      Set rangeY = Selection

      Sheets("TEMP").Select
      range("A1").Select

      'executa regressao
      Application.Run "ATPVBAEN.XLAM!Regress", rangeY, _
         rangeX, False, True, , "", True, False, False _
         , False, , False

      'copiando residuos
      range("C80").Select
      Selection.End(xlUp).Select
      range(Selection, Selection.End(xlUp)).Select
      Selection.Copy

      Sheets("Residuos").Select
      range("A1").Select

      If (ActiveCell.Offset(0, 0).Value = "") Then
         ActiveSheet.Paste
      Else
         If (ActiveCell.Offset(0, 1).Value <> "") Then
            Selection.End(xlToRight).Select
         End If
```

```
    ActiveCell.Offset(0, 1).Select
    ActiveSheet.Paste
  End If
  Selection.End(xlUp).Select
  ActiveCell.Offset(0, 0).Value = "Residuo " & j


  'finalizando
  Application.DisplayAlerts = False
  Worksheets(3).Activate
  ActiveSheet.Name = "Residuo " & j
  ActiveSheet.Move After:=Sheets(5)
  'adicionado depois q pedido para deletar
  ActiveSheet.Delete
  Worksheets(3).Delete
  Application.DisplayAlerts = True

  'marcando linha como feita
  Sheets("Matrix2").Select
  range("A1").Select
  ActiveCell.Offset(0, numCol).Select
  If (ActiveCell.Offset(0, 0).Value = "") Then
    ActiveCell.Offset(0, 0).Value = 1
  Else
    Do
      ActiveCell.Offset(1, 0).Select
    Loop Until ActiveCell.Offset(0, 0).Value = ""
    ActiveCell.Offset(0, 0).Value = 1
  End If

'Loop Until ActiveCell.Offset(0, 0).Value = ""
Next j

'deletando colunas de 1 auxiliar
Sheets("Matrix2").Select
range("A1").Select
Selection.End(xlToRight).Select
ActiveCell.EntireColumn.Delete

  'mostrar matriz 1
  Sheets("Matrix1").Select
  range("A1").Select

  'etapa2

End Sub
-----------------------------------------------------------------------------------------------
Sub etapa2()

  Sheets("Residuos").Select
  range("A1").Select

  'transpor titulos e montar indices
  If (ActiveCell.Offset(0, 1).Value <> "") Then
    range(Selection, Selection.End(xlToRight)).Select
    Selection.Copy
  End If
  range("A1").Select
  If (ActiveCell.Offset(0, 0).Value <> "") Then
    Selection.End(xlToRight).Select
  End If
  ActiveCell.Offset(0, 5).Select
  ActiveCell.Offset(0, 0).Value = "VALMAX"
  ActiveCell.Offset(0, 1).Value = "VALMIN"
  ActiveCell.Offset(0, 2).Value = "VPPA"
  ActiveCell.Offset(1, -1).Select
  Selection.PasteSpecial Paste:=xlPasteAll, Operation:=xlNone, SkipBlanks:= _
          False, Transpose:=True
  Application.CutCopyMode = False

  'seta primeira celula para ser preenchida
  ActiveCell.Offset(0, 1).Select
  Dim celulaBasica As range
  Set celulaBasica = Selection

  'conta numero de residuos
  Dim numRes As Integer
  numRes = 0
```

130

```
    range("A1").Select
    Do
        If (ActiveCell.Offset(0, 0).Value <> "") Then
            numRes = numRes + 1
        End If
        ActiveCell.Offset(0, 1).Select
    Loop Until ActiveCell.Offset(0, 0).Value = ""

    'preenche dados
    celulaBasica.Select
    For i = 1 To numRes
        Dim rangeAtual As range
        Columns(i).Select
        Set rangeAtual = Selection
        celulaBasica.Select
        Do
            If (ActiveCell.Offset(0, 0).Value <> "") Then
                ActiveCell.Offset(1, 0).Select
            End If
        Loop Until ActiveCell.Offset(0, 0).Value = ""
        ActiveCell.Value = Application.WorksheetFunction.Max _
                    (rangeAtual)
        ActiveCell.Offset(0, 1).Value = Application.WorksheetFunction.Min _
                    (rangeAtual)
        ActiveCell.Offset(0, 2).Value = ActiveCell.Offset(0, 0).Value -
ActiveCell.Offset(0, 1).Value

    Next i

    'executa correlacao
    Sheets("Residuos").Select
    range("A1").Select
    range(Selection, Selection.End(xlToRight)).Select
    range(Selection, Selection.End(xlDown)).Select
    Dim correlRange As range
    Set correlRange = Selection
     Application.Run "ATPVBAEN.XLAM!Mcorrel", correlRange, _
        "", "C", True
    ActiveSheet.Name = "CORREL"

    'copia e corta menu criado - passa para aba CORREL
'    Sheets("Residuos").Select
'    celulaBasica.Offset(-1, -1).Select
'    ActiveCell.Offset(0, 0).Value = "Tabela"
'    Range(Selection, Selection.End(xlToRight)).Select
'    Range(Selection, Selection.End(xlDown)).Select
'    Selection.Cut
'    Sheets("CORREL").Select
'    Range("A1").Select
'    Selection.End(xlToRight).Select
'    Selection.End(xlToRight).Select
'    ActiveCell.Offset(0, 5).Select
'    ActiveSheet.Paste

   'cria menu em CORREL
    range("B1").Select
    Selection.End(xlToRight).Select
    ActiveCell.Offset(0, 5).Select
    ActiveCell.Offset(0, 0).Value = "VALMAX"
    ActiveCell.Offset(0, 1).Value = "VALMIN"
    ActiveCell.Offset(0, 2).Value = "VPPACC"
    ActiveCell.Offset(0, 3).Value = "VALMED"
    ActiveCell.Offset(0, 4).Value = "MVPA"
    ActiveCell.Offset(0, 5).Value = "RESI"
    Dim celulaBasicaMenu As range
    Set celulaBasicaMenu = ActiveCell.Offset(2, 0)

    'conta numero de iteracoes
    range("B1").Select
    Selection.End(xlToRight).Select
    Dim numIteracoes As Integer
    numIteracoes = ActiveCell.Column - 2

    'preenche dados no menu
    For i = 1 To numIteracoes
        range("B3").Select
        Dim cont As Integer
        ActiveCell.Offset(i - 1, 0).Select
```

131

```
    Dim maior As Double
    Dim menor As Double
    maior = ActiveCell.Offset(0, 0).Value
    menor = ActiveCell.Offset(0, 0).Value
    Do
       ActiveCell.Offset(0, 1).Select
       If ActiveCell.Offset(0, 0).Value <> 1 Then
          If ActiveCell.Offset(0, 0).Value > maior Then
             maior = ActiveCell.Offset(0, 0).Value
          End If
          If ActiveCell.Offset(0, 0).Value < menor Then
             menor = ActiveCell.Offset(0, 0).Value
          End If
       End If
     Loop Until ActiveCell.Offset(0, 0).Value = 1
     'seleciona celula para preencher
     celulaBasicaMenu.Select
     Do
       If ActiveCell.Offset(0, 0).Value <> "" Then
          ActiveCell.Offset(1, 0).Select
       End If
     Loop Until ActiveCell.Offset(0, 0).Value = ""
     ActiveCell.Offset(0, 0).Value = maior
     ActiveCell.Offset(0, 1).Value = menor
     ActiveCell.Offset(0, 2).Value = ActiveCell.Offset(0, 0).Value -
ActiveCell.Offset(0, 1).Value
   Next i

  'preenche coluna RESI
  range("B1").Select
  If ActiveCell.Offset(0, 1).Value <> "" Then
     range(Selection, Selection.End(xlToRight)).Select
  End If
  Selection.Copy
  celulaBasicaMenu.Offset(-1, 5).Select
  Selection.PasteSpecial Paste:=xlPasteAll, Operation:=xlNone, SkipBlanks:= _
     False, Transpose:=True

  'preenche VALMED
  Dim rangeMedia As range
  For i = 0 To numIteracoes - 1
     range("B3").Select
     ActiveCell.Offset(0 + i, 0).Select
     Set rangeMedia = ActiveCell
     Do
        ActiveCell.Offset(0, 1).Select
        If ActiveCell.Offset(0, 0).Value <> 1 Then
           Set rangeMedia = Union(rangeMedia, ActiveCell.Offset(0, 0))
        End If
     Loop Until ActiveCell.Offset(0, 0).Value = 1
     rangeMedia.Select
     'seleciona celula para preencher
     celulaBasicaMenu.Offset(0, 3).Select
      Do
        If ActiveCell.Offset(0, 0).Value <> "" Then
           ActiveCell.Offset(1, 0).Select
        End If
      Loop Until ActiveCell.Offset(0, 0).Value = ""
      ActiveCell.Offset(0, 0).Value = Application.WorksheetFunction.Average _
                (rangeMedia)
  Next i

  'preenche MVPA
  celulaBasicaMenu.Offset(0, 4).Select
  For i = 1 To numIteracoes
     ActiveCell.Offset(0, 0).Value = ActiveCell.Offset(0, -2).Value / 2 +
ActiveCell.Offset(0, -1).Value
     ActiveCell.Offset(1, 0).Select
  Next i

  'criando final

  'copiando RESIDUOS
  Sheets.Add After:=ActiveSheet
  ActiveSheet.Name = "FINAL"
  Sheets("CORREL").Select
  celulaBasicaMenu.Offset(-2, 5).Select
  range(Selection, Selection.End(xlDown)).Select
```

132

```
Selection.Copy
Sheets("FINAL").Select
range("A1").Select
ActiveSheet.Paste

'copiando VPPA
Sheets("Residuos").Select
celulaBasica.Offset(-1, 2).Select
Columns(ActiveCell.Column).Select
Selection.Copy
Sheets("FINAL").Select
range("B1").Select
ActiveSheet.Paste

'copiando MVPA
Sheets("CORREL").Select
celulaBasicaMenu.Offset(-2, 4).Select
Columns(ActiveCell.Column).Select
Selection.Copy
Sheets("FINAL").Select
range("C1").Select
ActiveSheet.Paste
With ActiveWorkbook.Sheets("FINAL").Tab
    .Color = 5287936
    .TintAndShade = 0
End With

'ordenar do menor para o maior
Sheets("FINAL").Select
range("A1").Select
Selection.End(xlDown).Select
Selection.End(xlToRight).Select
Dim ultimaCelulaFinal As range
Set ultimaCelulaFinal = ActiveCell
Dim primeiraCelulaFinal As range
Set primeiraCelulaFinal = range("A2")
Columns("B:B").Select
ActiveWorkbook.Worksheets("FINAL").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets("FINAL").Sort.SortFields.Add
Key:=range("B1"), _
        SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets("FINAL").Sort
        .SetRange range(primeiraCelulaFinal, ultimaCelulaFinal)
        .Header = xlNo
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

    'selecionar 25%
    range("A1").Select
    range(Selection, Selection.End(xlToRight)).Select
    Selection.Copy
    range("F1").Select
    ActiveSheet.Paste
    range("A3").Select
    range(Selection, Selection.End(xlToRight)).Select
    range(Selection, Selection.End(xlDown)).Select
    Selection.Copy
    range("F2").Select
    ActiveSheet.Paste
    range("F2").Select
    Selection.End(xlDown).Select
    Dim umQuarto As Integer
    umQuarto = Round((ActiveCell.Row - 1) / 4)
    range("F2").Select
    For i = 1 To umQuarto
        ActiveCell.Offset(1, 0).Select
    Next i
    range(Selection, Selection.End(xlToRight)).Select
    range(Selection, Selection.End(xlDown)).Select
    Selection.ClearContents

    'ordenar por MVPA
    range("F2").Select
    range(Selection, Selection.End(xlDown)).Select
```

133

```
    range(Selection, Selection.End(xlToRight)).Select
    Dim rangeMVPA As range
    Set rangeMVPA = Selection
    Columns("H:H").Select
    ActiveWorkbook.Worksheets("FINAL").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets("FINAL").Sort.SortFields.Add
Key:=range("H1"), _
        SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets("FINAL").Sort
        .SetRange rangeMVPA
        .Header = xlGuess
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

    'finalizando
    range("H1").Select
    range(Selection, Selection.End(xlDown)).Select
    With Selection.Interior
        .Pattern = xlSolid
        .PatternColorIndex = xlAutomatic
        .Color = 49407
        .TintAndShade = 0
        .PatternTintAndShade = 0
    End With
    range("A1").Select



End Sub
```

134

REFERÊNCIAS[5]

ACOCK, A. C. Working with missing values. **Journal of Marriage and Family**, v. 67, n. 4, p. 1012–1028, 2005.

AHMED, S. N. **Physics and Engineering of Radiation Detection**. 2nd ed. Oxford, UK: Elsevier, 2015.

DE ALMEIDA, G. A massa de ar e a sua determinação analítica. **Caderno Brasileiro de Ensino de Física**, v. 31, n. 1, p. 156–166, 2014. Disponível em: <https://periodicos.ufsc.br/index.php/fisica/article/view/2175-7941.2014v31n1p156>. .

AMORIM, G. DA F. DE. **Uma abordagem para Design For Six Sigma (DFSS) baseada no odelo de quatro fases do Desdobramento da Função Qualidade**, 2018. Universidade Federal de Itajubá.

ANGUS, J. E. **Criteria for Choosing the Best Neural Network Part I**. 1991.

ARCE, P. E. B. **Aplicação da Teoria do Portfólio para Otimização de Carteiras de Contratos de Energia Elétrica e Gestão de Risco**, 2014. Universidade de São Paulo.

EL ATI-HELLAL, M.; HELLAL, F.; HEDHILI, A. Application of *Plackett_Burman* and Doehlert designs for optimization of selenium analysis in plasma with electrothermal atomic absorption spectrometry. **Clinical Biochemistry**, 2014.

EL ATY, A. A.; WEHAIDY, H. R.; MOSTAFA, F. A. Optimization of inulinase production from low cost substrates using Plackett–Burman and Taguchi methods. **Carbohydrate Polymers**, v. 102, p. 261–268, 2014.

BALESTRASSI, P. P.; POPOVA, E.; PAIVA, A. P.; MARANGON LIMA, J. W. Design of experiments on neural network's training for nonlinear time series forecasting. **Neurocomputing**, v. 72, n. 4–6, p. 1160–1178, 2009.

BANERJEE, A.; SARKAR, P.; BANERJEE, S. Application of statistical design of experiments for optimization of As(V) biosorption by immobilized bacterial biomass. **Ecological Engineering**, 2016.

BARRETO, H.; HOWLAND, F. M. **INTRODUCTORY ECONOMETRICS: Using Monte Carlo Simulation with Microsoft Excel**. 2nd. ed. Cambridge, UK: Cambridge University Press, 2006.

BARRETO, J. M. Indrodução as Redes Neurais Artificiais. , 2002.

BERES, D. L.; HAWKINS, D. M. Plackett–Burman technique for sensitivity analysis of many-parametered models. **Ecological Modelling**, v. 141, p. 171–183, 2001. Disponível em: <www.elsevier.com/locate/ecolmodel>. .

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p. 241–264, 2002. MCB UNIVERSITY PRESS. Disponível em: <http://www.emeraldinsight.com/10.1108/01443570210414338>. .

[5] Baseadas na NBR-6023 de agosto de 2002, da Associação Brasileira de Normas Técnicas (ABNT). Abreviatura dos títulos dos periódicos em conformidade com o MEDLINE.

BONTEMPI, G.; BEN TAIEB, S.; LE BORGNE, Y. A. Machine learning strategies for time series forecasting. **Lecture Notes in Business Information Processing**, 2013.

BOX, G. E. P.; BEHNKEN, D. W. American Society for Quality Some New Three Level Designs for the Study of Quantitative Variables Society for Quality Stable URL : http://www.jstor.org/stable/1266454 Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use. **Technometrics**, v. 2, n. 4, p. 455–475, 1960.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. **Statistics for experimenters: an introduction to design, data analysis, and model building**. Wiley, 1978.

BOX, G. E. P.; WILSON, K. B. On the Experimental Attainment of Optimum Conditions. **Journal of the Royal Statistical Society, Series B (Methodological)**, v. 13, n. 1, p. 1–45, 1951. Disponível em: <http://www.jstor.org/stable/2983966>. .

BOX, G.; JENKINS, G. Time series analysis: forecasting and control. **Holden-Day INC, California**, 1976.

CAETANO, T. C.; DIAS, W. S.; CAMPOS, R. P.; HICKEL, G. Determinação da extinção atmosférica e brilho do cé u no OPD. Workshop OPD, SOAR e Gemini: Passado, Presente e Futuro . **Anais...** . p.1, 2010. Campos do Jordão, SP: Laboratório Nacional de Astrofísica. Disponível em: <http://www.lna.br/workshop2010/Proc-OSG/posters/ThiagoCostaCaetano.pdf>. Acesso em: 5/3/2017.

CAI, K.; ZHAO, H.; XIANG, Z.; et al. Enzymatic hydrolysis followed by gas chromatography-mass spectroscopy for determination of glycosides in tobacco and method optimization by response surface methodology. **Analytical Methods**, v. 6, n. 17, p. 7006, 2014. Disponível em: <http://xlink.rsc.org/?DOI=C4AY01056F>. Acesso em: 30/9/2016.

CHAPIN III, F. S., RUPP, T. S., STARFIELD, A. M., DEWILDE, L., ZAVALETA, E. S., FRESCO, N., ET AL. Planning for resilience: modeling change in human–fire interactions in the Alaskan boreal forest. **Frontiers in Ecology and the Environment**, v. 1, n. 5, p. 255–261, 2003.

CHATFIELD, C. **The Analysis of Time Series - An Introduction**. 5th ed. Boca Raton, USA: Chapman & Hall/CRC, 1996.

CHIANG, W.-C.; URBAN, T. L.; BALDRIDGE, G. W. A Neural Network Approach to Mutual Fund Net Asset Value Forecasting. **Omega, Int. J. Mgmt Sci.**, v. 24, n. 2, p. 205–215, 1996.

CNPQ/MCTIC. Tabela de Áreas do Conhecimento do CNPq. Disponível em: <http://cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf>. .

COLEMAN, D. E.; MONTGOMERY, D. C. A Systematic Approach to Planning for a Designed Industrial Experiment. **Technometrics**, v. 35, n. 1, p. 1–12, 1993.

CONFALONIERI, R.; BELLOCCHI, G.; BREGAGLIO, S.; DONATELLI, M.; ACUTIS, M. Comparison of sensitivity analysis techniques: A case study with the rice model WARM. **Ecological Modelling**, v. 221, n. 16, p. 1897–1906, 2010.

CORRADI, O.; OCHSENFELD, H.; MADSEN, H.; PINSON, P. Controlling electricity consumption by forecasting its response to varying prices. **IEEE Transactions on Power Systems**, v. 28, n. 1, p. 421–429, 2013.

COUTO, A. F. DO. **Análise de correlação no reconhecimento de interações em arranjos Plackett-Burman**, 2012. UNIFEI.

COX, D. R.; REID, N. **The Theory of the Design of Experiments**. New York, USA:

Chapman & Hall/CRC, 2000.

CRESSIE, N. A. C. **Statistics for spatial data**. 2nd. ed. Wiley-Interscience, 2015.

CROWLEY, K.; HEAD, B. W. The enduring challenge of 'wicked problems': revisiting Rittel and Webber. **Policy Sciences**, v. 50, n. 4, p. 539–547, 2017. Springer US.

CRUMEY, A. Human contrast threshold and astronomical visibility. **MNRAS**, v. 442, p. 2600–2619, 2014.

DIAMOND, N. T. Some Properties of a Foldover Design. **Australian & New Zealand Journal of Statistics**, v. 37, n. 3, p. 345–352, 1995.

DOMINICI, T. Poluição Luminosa : Monitoramento da qualidade do céu no Observatório do Pico dos Dias utilizando o SQM-L. Disponível em: <http://poluicaoluminosa.blogspot.com.br/2013/02/monitoramento-da-qualidade-do-ceu-no.html>. Acesso em: 5/3/2017.

DOMINICI, T. P.; RANGEL, M. F. Utilizando conceitos de patrimônio como uma estratégia de proteção do direito à luz das estrelas. **Museologia e Patrimônio**, v. 10, n. 1, p. 32–64, 2017. Disponível em: <http://revistamuseologiaepatrimonio.mast.br/index.php/ppgpmus/article/view/529/541>. .

DONDERS, A. R. T.; VAN DER HEIJDEN, G. J. M. G.; STIJNEN, T.; MOONS, K. G. M. Special Series: Missing Data - Review: A gentle introduction to imputation of missing values. **Journal of Clinical Epidemiology**, v. 59, p. 1087–1091, 2006. Disponível em: <http://www.uvm.edu/~rsingle/JournalClub/papers/Donders+2006-JClinEpi_imputation.pdf>. Acesso em: 24/5/2016.

DOUCOURE, B.; AGBOSSOU, K.; CARDENAS, A. Time series prediction using artificial wavelet neural network and multi-resolution analysis: Application to wind speed data. **Renewable Energy**, v. 92, p. 202–211, 2016. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.renene.2016.02.003>. .

DOWD, M. A bio-physical coastal ecosystem model for assessing environmental effects of marine bivalve aquaculture. **Ecological Modelling**, v. 183, n. 2–3, p. 323–346, 2005.

DRAVINS, D.; LINDEGREN, L.; MEZEY, E.; YOUNG, A. T. Atmospheric Intensity Scintillation of Stars. I. Statistical Distributions and Temporal Properties. **Publications of the Astronomical Society of the Pacific**, v. 109, p. 173–207, 1997. Disponível em: <http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1997PASP..109..173D&data_type=PDF_HIGH&whole_paper=YES&type=PRINTER&filetype=.pdf>. .

EVAGORA, A, M.; MURRAY, N. J.; HOLLAND, A. D.; ENDICOTT, J. Novel method for identifying the cause of inherent ageing in Electron Multiplying Charge Coupled Devices. In: SISSA (Org.); The 9th International Conference on Position Sensitive Detectors. **Anais...** . p.11, 2012. Aberystwyth , U.K.: IOP Publishing Ltd. Disponível em: <http://iopscience.iop.org/article/10.1088/1748-0221/7/01/C01023/pdf>. Acesso em: 25/1/2017.

FAULKNER, B. R.; LYON, W. G.; KHAN, F. A.; CHATTOPADHYAY, S. Modeling leaching of viruses by the Monte Carlo method. **Water Research**, v. 37, n. 19, p. 4719–4729, 2003.

FISHER, R. A. S. **The Design of Experiments**. 9th. ed. New York, USA: Hafner Press,a Division of Macmillan Publishing Co., Inc., 1974.

FISHER, R. A. S. **The Design of Experiments**. 9th ed. New York, USA: Hafner Press, a subdivision of Macmillan Publishing Co., Inc., 1974.

FISHER, R. A.; WISHART, J. The arrangement of field experimants and the statistical reduction of the results. **Technical Communication n°. 10**, v. 10, n. 10, p. 24, 1930.

FRIES, A.; HUNTER, W. G. Minimum Aberration 2k-p Designs. **Technometrics**, v. 22, n. 4, p. 601–608, 1980. Disponível em: <http://www.jstor.org/stable/1268198>. Acesso em: 24/6/2016.

GAILLARD, P.; GOUDE, Y.; NEDELLEC, R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. **International Journal of Forecasting**, v. 32, n. 3, p. 1038–1050, 2016. Elsevier B.V.

GARGAGLIONI, S. Poluição luminosa e a necessidade de uma legislação. **ComCiëncia**, v. 112, 2009. Disponível em: <http://comciencia.scielo.br/pdf/cci/n112/a08n112.pdf>. Acesso em: 5/3/2017.

GARGAGLIONI, S. R. **Análise Legal Dos Impactos Provocados Pela Poluição**, 2007. UNIFEI. Disponível em: <http://saturno.unifei.edu.br/bim/0032988.pdf>. Acesso em: 5/3/2017.

GARGAGLIONI, S. R. Aprendendo mais sobre a Poluição Luminosa. Disponível em: <http://www.lna.br/lp/definicao.html#A_Polui%E7%E3o_Luminosa_nos_arredores_do>. Acesso em: 5/3/2017.

GEVREY, M.; DIMOPOULOS, I.; LEK, S. Two-way interaction of input variables in the sensitivity analysis of neural network models. **Ecological Modelling**, v. 195, n. 1–2, p. 43–50, 2006. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380005005752>. Acesso em: 24/9/2016.

GIESBRECHT, F. G.; GUMPERTZ, M. L. **Planning, Construction and Statistical Analysis of Comparative Experiments**. Hoboken, New Jersey: Wiley-Interscience, John Wiley & Sons, Inc., 2004.

HASSAN, M.; ESSAM, T.; YASSIN, A. S.; SALAMA, A. Optimization of rhamnolipid production by biodegrading bacterial isolates using *Plackett_Burman* design. **International Journal of Biological Macromolecules**, 2016.

HAYKIN, S. **Neural Networks: a Comprehensive Foundation**. 2nd. ed. New Jersey, USA: Prentice Hall International, Inc., 1999.

HIGGINS, J. P. T.; GREEN, S. (EDITORS). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Disponível em: <http://handbook.cochrane.org/front_page.htm>. Acesso em: 6/3/2017.

HINKELMANN, K.; KEMPTHORNE, O. **Design and analysis of experiments: Vol.3 Special designs and applications**. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012.

HOLTON, J. M.; NIELSEN, C.; FRANKEL, K. A. The point-spread function of fiber-coupled area detectors. **Journal of Synchrotron Radiation**, v. 19, n. 6, p. 1006–1011, 2012. International Union of Crystallography.

HORTON, N. J.; KLEINMAN, K. P. Much Ado About Nothing : A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. **The American Statistician**, v. 61, n. 1, p. 79–90, 2007. Disponível em: <http://www-jstor-org.ez38.periodicos.capes.gov.br/stable/pdf/27643843.pdf?_=1464095896304>. .

HOWELL, D. C. Treatment of Missing Data--Part 1. Disponível em: <https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html>. Acesso em: 6/3/2017a.

HOWELL, D. C. Treatment of Missing Data--Part 2. Disponível em: <https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing-Part-Two.html>. Acesso em: 6/3/2017b.

HSIEH, W. W. Nonlinear multivariate and time series analysis by neural network methods. **Reviews of Geophysics**, v. 42, n. 2002, p. 10–1029, 2004. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.4298&rep=rep1&type=pdf%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.4298>. .

INMAN, R. H.; PEDRO, H. T. C.; COIMBRA, C. F. M. Solar forecasting methods for renewable energy integration. **Progress in Energy and Combustion Science**, v. 39, n. 6, p. 535–576, 2013. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.pecs.2013.06.002>. .

JAYABAL, S. **Development of coir-polyester composites by e-glass hybridization and assessment of drilling behavior**, 2010. Anna. Disponível em: <http://shodhganga.inflibnet.ac.in/handle/10603/11432>. .

JIJU, A. **Design of Experiments for Engineers & Scientists**. Burlington, MA: Butterworth Heinemann, 2012.

JONES, D. Norman Pogson and the Definition of Stellar Magnitude. **Astronomical Society of the Pacific Leaflets**, v. 10, n. 469, p. 145–152, 1967. Disponível em: <http://articles.adsabs.harvard.edu//full/1968ASPL...10..145J/0000150.000.html>. .

JURAN, J. M. **Quality Control Handbook**. 3rd ed. McGraw Hill Higher Education, 1974.

JURAN, J. M.; DE FEO, J. **Juran's quality handbook : the complete guide to performance excellence**. 6th ed. New York, USA: McGraw Hill, 2010.

KHOTANZAD, A.; ZHOU, E.; ELRAGAL, H. A neuro-fuzzy approach to short-term load forecasting in a price-sensitive environment. **IEEE Transactions on Power Systems**, v. 17, n. 4, p. 1273–1282, 2002.

KLEIJNEN, J. P. .; SANCHEZ, S. M.; LUCAS, T. W.; CIOPPA, T. M. A user's guide to the brave new world of designing simulation experiments. **INFORMS Journal on Computing**, v. 17, n. 3, p. 263–289, 2005. Disponível em: <https://harvest.nps.edu/papers/UserGuideSimExpts.pdf>. .

LATGÉ, C.; SIMON, N. II . Experimental study: Randomized experiment. Disponível em: <www.poliklinika-harni.hr%2FuserFiles%2Fupload%2Fdocuments%2FSTATISTIKA%2FII%2520Experimental%2520study.docx&usg=AFQjCNEtNhqDDG0NNO3sJu2YvprwXmQVlg>. Acesso em: 19/2/2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015. Disponível em: <http://dx.doi.org/10.1038/nature14539>. .

LEINERT, C.; BOWYER, S.; HAIKALA, L. K.; et al. The 1997 reference of diffuse night sky brightness. **Astronomy and Astrophysics Supplement Series**, v. 127, n. 1, p. 1–99, 1998.

LENTH, R. V. Quick and Easy Analysis of Unreplicated Factorials. **Technometrics**, v. 31, n. 4, p. 469–473, 1989. Disponível em:

<http://www.tandfonline.com/doi/abs/10.1080/00401706.1989.10488595>. Acesso em: 26/9/2016.

LI, G.; YANG, H. A Prediction Method for Underwater Acoustic Chaotic Signal Based on RBF Neural Network. **Journal of Software**, v. 9, n. 6, p. 1581–1587, 2014. Disponível em: <http://ojs.academypublisher.com/index.php/jsw/article/view/11357>. .

LI, H.; MEE, R. W. Better Foldover Plans for Resolution III 2k−p Designs. **Technometrics**, v. 44, n. 3, p. 278–283, 2002.

LI, W. K. **Diagnostic Checks in Time Series**. Chapman & Hall/CRC, 2004.

LI, W.; LIN, D. K. J. Optimal Foldover Plans for Two-Level Fractional Factorial Designs. **Technometrics**, v. 45, n. 2, p. 142–149, 2003. Taylor & Francis. Disponível em: <http://www.tandfonline.com/doi/abs/10.1198/004017003188618779>. Acesso em: 26/9/2016.

LIN, C. D.; MILLER, A.; SITTER, R. R. Folded over non-orthogonal designs. **Journal of Statistical Planning and Inference**, v. 138, n. 10, p. 3107–3124, 2008.

LINDBERG, T. **An application of DOE in the evaluation of optimization functions in a statistical software**, 2010. Universitet Umea. Disponível em: <http://www.diva-portal.se/smash/get/diva2:393126/FULLTEXT01.pdf>. Acesso em: 2/10/2016.

LIU, J. Adaptive forgetting factor OS-ELM and bootstrap for time series prediction. **International Journal of Modeling, Simulation, and Scientific Computing**, v. 8, n. 4, p. 1750029, 2017. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S1793962317500295>. .

MAGALLANES, J. F.; OLIVIERI, A. C. The effect of factor interactions in *Plackett_Burman* experimental designs. Comparison of Bayesian-Gibbs analysis and genetic algorithms. **Chemometrics and Intelligent Laboratory Systems**, 2010.

MAKRIDAKIS, S. G.; WHEELWRIGHT, S. C.; HYNDMAN, R. **Forecasting – methods and applications**. 3rd ed. New Jersey, USA: John Wiley & Sons, Inc., 1998.

MAKRIDAKIS, S.; HIBON, M. The M3-Competition: results, conclusions and implications. **International Journal of Forecasting**, v. 16, n. 4, p. 451–476, 2000.

MALLMITH, D. DE M. **PRÉ-SELEÇÃO DE SÍTIOS ASTRONÔMICOS POR IMAGENS DE SATÉLITES METEOROLÓGICOS**, 2004. UFRGS. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/5097/000509942.pdf?sequence=1>. Acesso em: 5/3/2017.

MARKOV, A. A. Spreading the law of large numbers by quantities that depend on each other. **Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete**, v. 15, n. 2, p. 135–156, 1906.

MARKOWITZ, H. Portfolio Selection. **The Journal of Finance**, v. 7, n. 1, p. 77–91, 1952. Disponível em: <http://links.jstor.org/sici?sici=0022-1082%28195203%297%3A1%3C77%3APS%3E2.0.CO%3B2-1>. Acesso em: 17/6/2018.

MARUJO, D.; SANTOS, M. V.; SOUZA, A. C. Z. DE; LOPES, B. I. L. Avaliação de Segurança de Tensão Considerando uma Técnica Híbrida de Previsão de Carga. XXIII SNPTEE - Seminário Nacional de Produção e Transmissão de Energia Elétrica. **Anais...** , 2015. Foz do Iguaçu, PR.

MATEOS, D. M.; RIVEAUD, L.; LAMBERTI, P. W. Detecting dynamical changes in time series by using the Jensen Shannon Divergence. , p. 1–14, 2017. Disponível em:

<http://arxiv.org/abs/1702.08276>. .

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943. Disponível em: <http://download.springer.com/static/pdf/691/art%253A10.1007%252FBF02478259.pdf?originUrl=http%253A%252F%252Flink.springer.com%252Farticle%252F10.1007%252FBF02478259&token2=exp=1488835856~acl=%252Fstatic%252Fpdf%252F691%252Fart%2525253A10.1007%2525252FBF02>. Acesso em: 6/3/2017.

MCRAE, G. J.; TILDEN, J. W.; SEINFELD, J. H. Global sensitivity analysis—a computational implementation of the Fourier Amplitude Sensitivity Test (FAST). **Computers & Chemical Engineering**, v. 6, n. 1, p. 15–25, 1982. Pergamon.

MIGUEL, P. C. (COORD. . **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 2a. ed. Rio de Janeiro: Elsevier: ABEPRO, 2012.

MILLER, A.; SITTER, R. R. Using the Folded-Over 12-Run Plackett—Burman Design to Consider Interactions. **Technometrics**, v. 43, n. 1, p. 44–55, 2001. Taylor & Francis. Disponível em: <http://www.tandfonline.com/doi/abs/10.1198/00401700152404318>. Acesso em: 26/9/2016.

MILLER, A.; SITTER, R. R. Using Folded-Over Nonorthogonal Designs. **Technometrics**, v. 47, n. 4, p. 502–513, 2005. Taylor & Francis. Disponível em: <http://www.tandfonline.com/doi/abs/10.1198/004017005000000210>. Acesso em: 26/9/2016.

MITROFF, I. I.; BETZ, F.; PONDY, L. R.; SAGASTI, F. On managing science in the systems age: two schemas for the study of science as a whole systems phenomenon. **Interfaces**, v. 4, n. 3, p. 46–58, 1974.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 7th ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2009.

MONTGOMERY, D. C. **Design and analysis of experiments**. 8th ed. John Wiley & Sons, Inc, 2013.

MONTGOMERY, D. C.; JENNINGS, C.; KULAHCI, M. **Introduction to time series analysis and forecasting**. Hoboken, New Jersey: John Wiley & Sons, Inc., 2008.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 5th ed. John Wiley & Sons, Inc., 2011.

MORSTYN, T.; FARRELL, N.; DARBT, S. J.; MCCULLOCH, M. D. Using peer-to-peer energy-trading platforma to incentivize prosumers to form federated power plants. **Nature Energy**, v. 3, p. 94–101, 2018.

MOTAMEDI, A.; ZAREIPOUR, H.; ROSEHART, W. D. Electricity Price and Demand Forecasting in Smart Grids. **IEEE Transactions on Smart Grid**, v. 3, n. 2, p. 664–674, 2012. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6204245&queryText=Electricity+Price+and+Demand+Forecasting+in+Smart+Grids>. .

MYERS, R. H.; MONTGOMERY, D. C.; ANDERSON-COOK, C. M. **Response surface methodology: process and product optimization using designed experiments**. 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2009.

NEWHAM, L. T. H. **Catchment Scale Modelling of Water Quality and Quantity**, 2002.

The Australian National University.

NIELSEN, M. A. Neural Networks and Deep Learning. Disponível em: <http://neuralnetworksanddeeplearning.com/index.html>. Acesso em: 21/3/2017.

NISBET, R.; ELDER, J.; MINER, G. **Handbook of Statistical Analysis & Data Minig Applications**. 1st. ed. Canada: Academic Press/Elsevier, 2009.

NOAO. Calculation of Time on Keck | ast.noao.edu. Disponível em: <http://ast.noao.edu/system/tsip/more-info/time-calc-keck>. Acesso em: 18/1/2017.

DE OLIVEIRA, K.; SARAIVA, M. DE F. **Astronomia e Astrofísica**. São Paulo, Brasil: Livraria da Física, 2013.

DE OLIVEIRA, L. G. **Fundamentos da metodologia de superfície de resposta e suas aplicações em manufatura avançada: uma análise crítica**, 2018. UNIFEI.

PERIAGO, M. J.; RINCÓN, F.; JACOB, K.; GARCÍA-ALONSO, J.; ROS, G. Detection of Key Factors in the Extraction and Quantification of Lycopene from Tomato and Tomato Products. **Journal of Agricultural and Food Chemistry**, v. 55, n. 22, p. 8825–8829, 2007. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/jf0705623>. Acesso em: 26/9/2016.

PHADKE, B. M. S. Introduction To Robust Design (Taguchi Method). .

PIPINO, L. L.; LEE, Y. W.; YANG, R. W. Data Quality Assessment. **Communication of the ACM**, v. 45, n. 4, p. 211–218, 2002. Disponível em: <http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf>. .

PLACKETT, R. L.; BURMAN, J. P. Biometrika Trust. **Biometrika**, v. 33, n. 4, p. 305–325, 1946a. Disponível em: <http://www.jstor.org/stable/2332195>. .

PLACKETT, R. L.; BURMAN, J. P. The Design of Optimum Multifactorial Experiments. **Source: Biometrika**, v. 33, n. 4, p. 305–325, 1946b. Disponível em: <http://www.jstor.org>. .

PUSCHNIG, J.; POSCH, T.; UTTENTHALER, S. Night sky photometry and spectroscopy performed at the Vienna University Observatory. **Journal of Quantitative Spectroscopy and Radiative Transfer**, v. 139, p. 64–75, 2014. Disponível em: <http://ac-els-cdn-com.ez38.periodicos.capes.gov.br/S002240731300352X/1-s2.0-S002240731300352X-main.pdf?_tid=59891a20-e4b1-11e6-b362-00000aacb35d&acdnat=1485536284_3fa0ef0a991f7103997c1d9db0860660>. .

RIBEIRO JÚNIOR, H. J.; MOTA, R. L. M.; LEME, R. C.; SANTOS, P. E. S. Ensaios *Plackett_Burman* para identificação de elementos de custo tarifário de energia elétrica. XXXIII Encontro Nacional de Engenhari de Produção. **Anais...** . p.1–16, 2013. Salvador, BA: ABEPRO.

RITTEL, H. W. J.; WEBBER, M. M. Dilemmas in a general theory of planning. **Policy Sciences**, v. 4, n. 2, p. 155–169, 1973.

ROMERO, J.; LÓPEZ, P.; RUBIO, C.; BATLLE, R.; NERÍN, C. Strategies for single-drop microextraction optimisation and validation. Application to the detection of potential antimicrobial agents. **Journal of Chromatography A**, v. 1166, n. 1–2, p. 24–29, 2007.

SALTELLI, A. (ANDREA); CHAN, K. (KAREN); SCOTT, E. M. **Sensitivity analysis**. Wiley, 2009.

SALVADOR, S.; KOREVAAR, M. A N.; HEEMSKERK, J. W. T.; et al. Improved EMCCD gamma camera performance by SiPM pre-localization. **Physics in Medicine and Biology**, v. 57, p. 7709–7724, 2012. Disponível em:

<http://iopscience.iop.org.ez38.periodicos.capes.gov.br/article/10.1088/0031-9155/57/22/7709/pdf>. .

SCHEURER, D. L. **A spatially-explicit framework for investigating patchiness effects in aquatic ecosystems**, 2006. University of Maryland.

SCHONLAU, M.; WELCH, W. J. **Screening - Methods for Experimentation in Industry, Drug Discovery, and Genetics**. New York, USA: Springer Science+Business Media, Inc., 2006.

SHEN, L.; MORRIS, M. D. Augmented Plackett–Burman designs with replication and improved bias properties. **Journal of Statistical Planning and Inference**, 2016.

SIVARAO, S.; ANAND, T. J. S.; AMMAR, A. DOE Based Statistical Approaches in Modeling of Laser Processing – Review & Suggestion. **International Journal of Engineering and Technology**, v. 10, n. 04, p. 1–8, 2010.

SMITH, S. K. **Digital Signal Processing - A Practical Guide for Engineers and Scientists**. Newness for Elsevier, 2003.

SOUZA, J. R.; PESSIN, G.; OSÓRIO, F. S.; WOLF, D. F. **Vision-based autonomous navigation using supervised learning techniques**. EANN/AIAI ed. IFIP International Federation for Information Processing, 2011.

SOUZA, M. F. Z. DE. **Modelagem e simulação integrada ao processamento e diagnóstico do desempenho de parâmetros elétricos no contexto de redes inteligentes**, 2017. UNIFEI.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research**, v. 15, p. 1929–1958, 2014.

ŠTĚPNIČKA, M.; CORTEZ, P.; DONATE, J. P.; ŠTĚPNIČKOVÁ, L. Forecasting seasonal time series with computational intelligence: On recent methods and the potential of their combinations. **Expert Systems with Applications**, v. 40, n. 6, p. 1981–1992, 2013.

STERNE, J. A. C.; WHITE, I. R.; CARLIN, J. B.; et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. **British Medical Journal**, v. 338, p. b2393, 2009. Disponível em: <http://www.bmj.com/content/338/bmj.b2393.full.print>. .

STRONG, D.; LEE, Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103–110, 1997. Disponível em: <http://dl.acm.org/citation.cfm?id=253804>. Acesso em: 31/3/2012.

TANGIRALA, A. K. **Principles of System Identification: Theory and Practice**. Boca Raton, USA: CRC Press, Taylor & Francis Group, 2015.

TAYLOR, B. J.; SMITH, J. T. **Methods and Procedures for the Verification and Validation of Artificial Neural Networks**. Fairmont, WV: Springer Science+Business Media, Inc., 2006.

TIROZZI, B.; PUCA, S.; PITTALIS, S.; et al. **Neural Networks and Sea Time Series**. Boston: Birkhäuser/Springer Science+Business Media Inc., 2006.

TIROZZI, B.; PUCA, S.; PITTALIS, S.; et al. **Neural Networks and Sea Time Series Reconstruction and Extreme-Event Analysis**. Boston: Birkhäuser, 2006.

VAHIDNIA, M. H.; ALESHEIKH, A. A.; ALIMOHAMMADI, A.; HOSSEINALI, F. A GIS-based neuro-fuzzy procedure for integrating knowledge and data in landslide

susceptibility mapping. **Computers and Geosciences**, v. 36, n. 9, p. 1101–1114, 2010. Elsevier. Disponível em: <http://dx.doi.org/10.1016/j.cageo.2010.04.004>. .

WU, W.; HALL, C. A. S.; SCATENA, F. N. Modelling the impact of recent land-cover changes on the stream flows in northeastern Puerto Rico. **Hydrological Processes**, v. 21, n. 21, p. 2944–2956, 2007. John Wiley & Sons, Ltd. Disponível em: <http://doi.wiley.com/10.1002/hyp.6515>. Acesso em: 26/9/2016.

XU, C.; GERTNER, G. Extending a global sensitivity analysis technique to models with correlated parameters. **Computational Statistics and Data Analysis**, v. 51, n. 12, p. 5579–5590, 2007.

XU, C.; GERTNER, G. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). **Computational Statistics and Data Analysis**, v. 55, n. 1, p. 184–198, 2011.

XU, C.; GERTNER, G. Z. A general first-order global sensitivity analysis method. **Reliability Engineering & System Safety**, v. 93, n. 7, p. 1060–1071, 2008a.

XU, C.; GERTNER, G. Z. Uncertainty and sensitivity analysis for models with correlated parameters. **Reliability Engineering & System Safety**, v. 93, n. 10, p. 1563–1573, 2008b.

YE, K. Q.; LI, W. Some properties of blocked and unblocked folovers of 2k-p designs. , v. 13, p. 403–408, 2003.

YEUNG, D. S.; CLOETE, I.; SHI, D.; NG, W. W. Y. **Sensitivity Analysis for Neural Networks**. Berlin: Springer Science+Business Media, Inc., 2010.

ZHANG, G. P. An investigation of neural networks for linear time-series forecasting. **Computers & Operations Research**, v. 28, n. 12, p. 1183–1202, 2001. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0305054800000332>. Acesso em: 13/3/2017.

ZHANG, L.; UME, I. C.; GAMALSKI, J.; GALUSCHKI, K. P. Detection of flip chip solder joint cracks using correlation coefficient and auto-comparison analyses of laser ultrasound signals. **IEEE Transactions on Components and Packaging Technologies**, v. 29, n. 1, p. 13–19, 2006.

ZHAO, A.; CHEN, F.; NING, C.; et al. Use of real-time cellular analysis and *Plackett_Burman* design to develop the serum-free media for PC-3 prostate cancer cells. **PLoS ONE**, v. 12, n. 9, p. 1–16, 2017.

ANEXO 1 - Produção acadêmica

Artigos publicados

(1) Amorim, Gabriela de Fonseca de ; Balestrassi, Pedro Paulo ; Paiva, Anderson Paulo de ; OLIVEIRA-ABANS, MARIÂNGELA DE . Detecção de mudança de nível em séries temporais não lineares usando Descritores de Hjorth. Production, v. 25, p. 812-825, 2015.

(2) Monticeli, A. R. ; Pereira, L. D. ; DE OLIVEIRA-ABANS, M. ; Josa, J. L. . Análise do uso de um sistema informatizado na gestão do conhecimento em um órgão da administração direta federal. Revista Gestão do Conhecimento (Curitiba. Impresso), v. 9, p. 1-19, 2015.

(3) Amorim, Gabriela Da Fonseca De ; Balestrassi, P. P. ; Sawney, R. ; DE OLIVEIRA-ABANS, M. ; da Silva, D. L. F. . Six Sigma learning evaluation model using Bloom's Taxonomy. International Journal Of Lean Six Sigma, 9(1), p.156-174, 2018.

(4) Campos, P. H. S. ; Belinato, G. ; Incerti, T. ; DE OLIVEIRA-ABANS, M. ; Ferreira, J. R. ; de Paiva, A. P. ; Balestrassi, P. P. . Multivariate Mean Square Error for the multiobjective optimization of AISI 52100 hardened steel turning with wiper ceramic inserts tool: a comparative study. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 39(10), p. 4121-4036, 2017.

Artigo em preparação

(1) "*Plackett_Burman* Correlation Analysis: a new method for multivariate models' sensitivity analysis", Anderson Fagundes do Couto, Mariângela de Oliveira-Abans, Gabriela da Fonseca Amorim, Anderson Paulo de Paiva, Pedro Paulo Balestrassi.

Trabalhos completos publicados em anais de congressos

(1) DE OLIVEIRA-ABANS, M.; Pereira, A. P. ; Vieira, L. F. S. ; Carvalho, M. E. F. ; Pires, R. M. . Um exemplo didático de Design For Six Sigma explorando o pêndulo de Newton. In: XXI SIMPEP - Simpósio de Engenharia de Produção, 2014, Bauru, SP. Anais do XXI Simpósio de Engenharia de Produção, 2014.

(2) Amorim, G. F. ; Pereira, T. F. ; DE OLIVEIRA-ABANS, M. ; Balestrassi, P. P. ; Montevechi, J. A. B. . Pesquisa operacional: modelagem matemática do planejamento de culturas em uma fazenda familiar. In: XX SIMPEP - Simpósio de Engenharia de Produção, 2013, Bauru, SP. SIMPEP - Simpósio de Engenharia de Produção, 2013.

Participação em bancas de Trabalhos de Conclusão de Curso de graduação (TCC) junto ao IEPG

(1) Balestrassi, P. P.; Amorim, G. F.; M. DE OLIVEIRA-ABANS. Participação em banca de Pedro Granato Pissinato.Desenvolvimento de uma *template* para DFSS considerando as quatro casas da qualidade de um QFD. 2015. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Federal de Itajubá.

(2) Balestrassi, P. P.; Peruchi, R.; DE OLIVEIRA-ABANS, M.. Participação em banca de Guilherme de Assis Mendes.Monitoramento de Eventos Raros Através da Carta de Controle G. 2013. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Federal de Itajubá.

(3) Balestrassi, P. P.; Peruchi, R.; DE OLIVEIRA-ABANS, M.. Participação em banca de Natália Silva Braga.Otimização de métodos de alisamento exponencial na previsão de séries temporais. 2013. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Federal de Itajubá.

(4) Balestrassi, P. P.; Peruchi, R.; DE OLIVEIRA-ABANS, M.. Participação em banca de Rafael Rachid de Almeida. Transformações de Johnson em estudos de capabilidade. 2013. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Federal de Itajubá.

Co-orientação de TCC junto ao IEPG

Pedro Granato Pissinato. Desenvolvimento de uma template para DFSS considerando as quatro casas da qualidade de um QFD. 2015. Trabalho de Conclusão de Curso. (Graduação em Engenharia de Produção) - Universidade Federal de Itajubá. Orientador: MARIÂNGELA DE OLIVEIRA ABANS.

Periódico para o qual atuou como revisora técnica em Engenharia de Produção

Production (ABEPRO, Brasil)