

**UNIVERSIDADE FEDERAL DO PARÁ  
CAMPUS UNIVERSITÁRIO DE SALINÓPOLIS  
FACULDADE DE MATEMÁTICA**

**TRABALHO DE CONCLUSÃO DE CURSO**

**BRENO DE CASTRO BARBOSA**

**APLICAÇÕES DE TEMAS EM PROBABILIDADE E SÉRIES  
TEMPORAIS NO ÂMBITO DE INICIAÇÃO CIENTÍFICA DE UM  
ALUNO DE GRADUAÇÃO**

**Campus Universitário de Salinópolis  
Salinópolis-Pará-Brasil  
2024**

**UNIVERSIDADE FEDERAL DO PARÁ  
CAMPUS UNIVERSITÁRIO DE SALINÓPOLIS  
FACULDADE DE MATEMÁTICA**

**BRENO DE CASTRO BARBOSA**

**APLICAÇÕES DE TEMAS EM PROBABILIDADE E SÉRIES  
TEMPORAIS NO ÂMBITO DE INICIAÇÃO CIENTÍFICA DE UM  
ALUNO DE GRADUAÇÃO**

Trabalho de Conclusão de Curso submetido à Banca Examinadora da Faculdade de Matemática do Campus Universitário de Salinópolis como quesito a obtenção do Grau de Licenciado em Matemática.

**Campus Universitário de Salinópolis  
Salinópolis-Pará-Brasil  
2024**


**UNIVERSIDADE FEDERAL DO PARÁ  
CAMPUS UNIVERSITÁRIO DE SALINÓPOLIS  
FACULDADE DE MATEMÁTICA**

**APLICAÇÕES DE TEMAS EM PROBABILIDADE E SÉRIES  
TEMPORAIS NO ÂMBITO DE INICIAÇÃO CIENTÍFICA DE UM  
ALUNO DE GRADUAÇÃO**

**AUTOR:** Breno de Castro Barbosa

TRABALHO DE CONCLUSÃO DE CURSO SUBMETIDO À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DA FACULDADE DE MATEMÁTICA DO CAMPUS UNIVERSITÁRIO DE SALINÓPOLIS E JULGADA ADEQUADA PARA OBTENÇÃO DO GRAU DE LICENCIADO EM MATEMÁTICA

**APROVADA EM:** 29/02/2024

Documento assinado digitalmente  
 **RAMZ LUIZ FRAIHA LOPES**  
Data: 22/05/2024 21:56:27-0300  
Verifique em <https://validar.iti.gov.br>

---


Prof. Dr. Ramz Luiz Fraiha Lopes

**ORIENTADOR**  
Documento assinado digitalmente  
 **AMILCAR MONTALBAN SAYAGO**  
Data: 22/05/2024 17:02:51-0300  
Verifique em <https://validar.iti.gov.br>

---

Prof. Dr. Amilcar Montalban Sayago

**MEMBRO – UFPA Salinópolis**

Documento assinado digitalmente  
 **VINÍCIUS DUARTE LIMA**  
Data: 21/05/2024 17:26:57-0300  
Verifique em <https://validar.iti.gov.br>

---

Prof. Dr. Vinícius Duarte Lima

**MEMBRO – UFPA/ICEN/Belém**

## Agradecimentos

Primeiramente gostaria de agradecer a Deus. Pois em minha concepção sem sua permissão nada disso seria possível. Segundo aos meus pais: Zannerri e Rosimeire, por me proporcionarem toda estrutura e dedicação, para que eu mantivesse o foco principal aos meus estudos e ao meu irmão: Bruno, por também contribuir de diversas formas no decorrer de toda graduação.

Agradeço a Universidade Federal do Pará (UFPA) em especial ao campus Salinópolis por me proporcionar iniciar a vida acadêmica na cidade onde vivo e à Pró-Reitoria de Pesquisa e Pós-Graduação (PROPESP) pela oportunidade de participação nos programas PARD e PIBIC enquanto bolsista. Também agradeço a Faculdade de Matemática (FAMAT) por todo apoio prestado no decorrer de toda a graduação, e a todos professores pelo excelente profissionalismo.

Ao meu orientador Prof. Ramz Fraiha, pela acessibilidade e pela excelente condução em todo o trabalho de pesquisa, agradeço pela confiança depositada em mim.

Também agradeço ao meu colega de graduação: Cid Tadeo por toda troca de conhecimento, e também à Emilly Rocha por todo apoio prestado.

# Sumário

Agradecimentos .....	3
Sumário.....	4
Resumo .....	6
Abstract.....	7
Lista de Acrônimos.....	8
Lista de Figuras .....	9
Lista de Tabelas .....	10
Capítulo 1 - Introdução.....	11
Capítulo 2 – Fundamentos Básicos em Probabilidade .....	12
2.1- Considerações Iniciais .....	12
2.2- Axiomas da probabilidade .....	12
2.3- Probabilidade condicional.....	13
2.4- Eventos independentes.....	13
2.5- Variável Aleatória.....	14
2.6- Distribuição de probabilidade .....	14
2.7- Função distribuição de probabilidade para VAs.....	15
2.8- Variável aleatória contínua .....	15
2.9- Distribuições conjuntas.....	15
2.9.1- Caso discreto.....	16
2.9.2- Caso contínuo .....	17
2.10- Variáveis aleatórias independentes .....	17
2.11- Esperança matemática.....	18
2.11.1- Alguns teoremas sobre esperança matemática.....	18
2.12- Variância e desvio padrão .....	19
2.13- Variáveis aleatórias padronizadas.....	20
2.14- Variância para distribuições conjuntas e covariância .....	21
2.15- Coeficiente de correlação.....	22
2.16- Considerações Finais .....	22
Capítulo 3 – Fundamentos Básicos em Séries Temporais.....	23
3.1- Considerações iniciais.....	23
3.2- Previsão de séries temporais .....	23
3.3- Modelos matemáticos dinâmicos estocásticos e determinísticos.....	24
3.3.1- Modelos e operadores importantes .....	24
3.3.1.1- Operadores deslocamento ao passado e deslocamento ao futuro.....	24

3.3.3.2- Modelos de filtro linear.....	25
3.3.3.3- Modelos autorregressivos .....	26
3.3.3.4- Modelos de médias móveis.....	27
3.3.3.5- Modelos autorregressivos de médias móveis.....	28
3.3.3.6- Modelos não estacionários.....	28
3.4- Séries temporais.....	29
3.5- Processos estocásticos estacionários.....	30
3.5.1- Média e variância de um processo estacionário.....	30
3.5.2- Coeficientes de autocovariância e autocorrelação .....	31
3.6- Estacionariedade fraca .....	31
3.7- Funções de autocovariância e autocorrelação.....	32
3.8- Considerações finais .....	33
Capítulo 4 – Aplicações e seus resultados.....	34
4.1- Considerações Iniciais .....	34
4.2- Modelo ARIMA aplicado à propagação de ondas eletromagnéticas – Parte 1.....	34
4.2.1- Introdução .....	34
4.2.2- Modelo híbrido ARIMA-RNA .....	34
4.2.3- Campanhas de medição.....	35
4.2.4- Tratamento dos dados .....	35
4.2.5- Resultados e conclusões.....	36
4.3- Modelo ARIMA aplicado à propagação de ondas eletromagnéticas – Parte 2.....	40
4.3.1- Introdução .....	40
4.3.2- Resultados e conclusões.....	40
4.4- Cálculo do Coeficiente de Correlação entre duas VAs.....	42
4.4.1- Introdução .....	42
4.4.2- Metodologia e tratamento dos dados .....	43
4.4.3- Resultados e conclusões.....	44
4.5- Considerações Finais .....	48
Capítulo 5 – Conclusões.....	49
Referências Bibliográficas.....	51

# Resumo

Este trabalho apresenta duas aplicações de temas como probabilidade, estatística e séries temporais a problemas reais nas quais o autor participou enquanto pesquisador em companhia de seu orientador. Uma delas é uma aplicação direta da tese de doutorado de seu orientador e outro uma segunda situação problema, também em um tema que orbita a área de probabilidade e estatística. Todas as aplicações mostradas resultaram em resumos/artigos publicados e se desenvolveram entre os anos de 2020 e 2023. Procura-se também destacar que atividades de pesquisa como as desenvolvidas pelo autor enriquecem a formação acadêmica de um aluno de graduação. Isto é feito ao recapitular os fundamentos básicos de probabilidade e séries temporais que são as bases das aplicações apresentadas neste trabalho. Além disso, toda a fundamentação básica apresentada busca deixar o trabalho acessível a um leitor eventualmente não iniciado nos temas deste estudo.

**PALAVRAS-CHAVE:** Probabilidade; Séries-Temporais; Modelos ARIMA; Correlação.

# **Abstract**

This work presents two applications of topics such as probability, statistics and time series to real problems in which the author participated as a researcher along with his supervisor. One of these is a direct application of his supervisor's doctorate thesis and the other one is another problem, also in a topic that orbits the great area of probability and statistics. All applications shown in this work resulted in published short works/articles and were carried out between 2020 and 2023. The author aims to highlight, as well, the importance of research activities in a graduation course to the enrichment of the academic education of a student. This is done by recapitulating basic fundamentals of probability and time series theory, which are the foundations of the applications presented in this work. Besides, all the fundamentals presented here aim to turn this study accessible to an undergraduate student that is not initiated in the topic of this work.

**KEY-WORDS:** Probability; Time series theory; ARIMA Model; Correlation.

## Lista de Acrônimos

**ACF** – Função de Autocorrelação

**ANATEL** – Agência Nacional de Telecomunicações

**ARIMA** – Auto-Regressivo Integrado de Médias Móveis

**ARIMA-RNA** – ARIMA-Rede Neural Artificial

**ITU** – *International Communication Union*

**MQ** – Mínimos Quadrados

**FACP** – Função de Autocorrelação Parcial

**ERMS** – Erro Médio Quadrático

**RNI** – Radiação Não-Ionizante

**UFPA** – Universidade Federal do Pará

## Lista de Figuras

<b>Figura 3.1-</b> Representação de uma série temporal como saída de um filtro linear ...	32
<b>Figura 3.2-</b> Matriz de autocorrelação e função de autocorrelação correspondente de um processo estacionário .....	40
<b>Figura 3.3-</b> Metade positiva da função de autocorrelação .....	40
<b>Figura 4.1-</b> Cenário de medições em salinópolis .....	42
<b>Figura 4.2-</b> Fluxograma da metodologia da modelagem híbrida .....	43
<b>Figura 4.3-</b> Diagrama da RNA .....	44
<b>Figura 4.4-</b> Comparações entre ajuste ARIMA-RNA e modelos da literatura .....	44
<b>Figura 4.5-</b> Conjunto de treino e ajuste do modelo híbrido .....	45
<b>Figura 4.6-</b> Conjunto 2 e estimativa do modelo híbrido .....	45
<b>Figura 4.7-</b> Conjunto 3 e estimativa do modelo híbrido .....	46
<b>Figura 4.8-</b> Comparações com o conjunto de treino (modelo ARIMA versus ARIMA-RNA) .....	47
<b>Figura 4.9-</b> Comparações com o conjunto 2 (modelo ARIMA versus ARIMA-RNA) .....	48
<b>Figura 4.10-</b> Comparações com o conjunto 3 (modelo ARIMA versus ARIMA-RNA) .....	48
<b>Figura 4.11-</b> Resultados notas e renda per capita de 7 municípios .....	51
<b>Figura 4.12-</b> Resultados notas e renda per capita de 7 municípios .....	52

## Lista de Tabelas

<b>Tabela 2.1-</b> Ilustração de eventos de uma variável aleatória .....	21
<b>Tabela 2.2-</b> Tabela de probabilidade conjunta .....	23
<b>Tabela 4.1-</b> Valores de ERMS entre as estimativas híbridas e os dados medidos .....	46
<b>Tabela 4.2-</b> Erros RMS do modelo linear e do modelo híbrido .....	48
<b>Tabela 4.3-</b> Escala interpretada .....	49
<b>Tabela 4.4-</b> Dados de correlação e parâmetro $p$ para 7 municípios .....	51
<b>Tabela 4.5-</b> Dados de correlação e parâmetro $p$ para 21 municípios .....	53

## Capítulo 1 - Introdução

O percurso acadêmico dentro de um curso de graduação é marcado pela quantidade significativamente maior de aulas (e outras atividades de ensino) em comparação com atividades de extensão e pesquisa, contudo, legislações e demais normativas, atendendo a demandas reais da sociedade brasileira, vem paulatinamente transformando essa realidade, ao exigir a integração maior de atividades de pesquisa e de extensão na formação de um aluno de graduação.

O presente trabalho tem por objetivo apresentar duas aplicações de pesquisa realizadas pelo autor no âmbito de seu curso de graduação e, em segundo lugar, ligar estas aplicações com conteúdos que são abordados normalmente na formação de um licenciado em Matemática da UFPA Salinópolis.

Inicialmente serão apresentados os fundamentos basilares (que podem ser facilmente estudados a nível de graduação) das aplicações mostradas nos capítulos finais. Estas aplicações são: (i) um modelo híbrido para representar a propagação de ondas eletromagnéticas na frequência de telefonia celular aplicado em um cenário da cidade de Salinópolis e; (ii) a avaliação da existência de correlação entre notas de português e matemática da “Provinha Brasil” e a renda per capita de municípios do estado do Pará. Ambos os problemas abordados neste trabalho possuem seu ferramental matemático de origem em assuntos de probabilidade, estatística e séries temporais. Temas esses que, em um processo de iniciação científica, tornam-se acessíveis a um aluno de graduação.

No capítulo 2 são apresentados fundamentos básicos em probabilidade. No capítulo 3 são apresentados fundamentos básicos em séries temporais. No capítulo 4 os problemas práticos atacados pelo autor (todos com publicações como resultado) são mostrados. Finalmente, o capítulo 5 mostra as conclusões deste trabalho de conclusão de curso e o elo entre os problemas práticos e a trajetória acadêmica de um aluno de graduação é retomado, fechando o raciocínio iniciado aqui, na introdução.

## Capítulo 2 – Fundamentos em Probabilidade

### 2.1- Considerações Iniciais

Estudar probabilidade é estudar as chances de determinado resultado acontecer em uma determinada situação. Dentro do estudo de probabilidade existem modelos matemáticos que são utilizados para modelar determinado fenômeno a partir das chances de determinado resultado ocorrer. Neste trabalho, uma das aplicações apresentadas é o estudo da correlação entre duas variáveis estudadas através de dados obtidos de eventos reais. A correlação é uma das métricas utilizadas no estudo de probabilidade para analisar a relação entre variáveis. Por outro lado, o que chamamos de estatística é a área de estudo que envolve as formas de tratamento de dados brutos para que estes sejam analisados posteriormente, por exemplo, através da probabilidade. Neste capítulo são apresentados fundamentos básicos em probabilidade e estatística que norteiam o ferramental matemático utilizado em uma das aplicações mostrada neste trabalho. A referência base para a fundamentação teórica exposta nesse capítulo é (Spiegel, 2009).

### 2.2- Axiomas da probabilidade

Seja um experimento aleatório, no qual todos os resultados são imprevisíveis ainda que feitos em igual circunstâncias. Por exemplo, ao lançar um dado (de seis faces) sabe-se que nesse lançamento será obtida uma das seis faces, porém, antes do experimento, a resposta final é desconhecida (Spiegel, 2009). O conceito de probabilidade está relacionado a situações desse tipo. Seja, então,  $S$  o espaço amostral de determinado experimento,  $S$  é formado pelo conjunto dos possíveis resultados do experimento. No lançamento de um dado esse conjunto é dado por  $S = \{1,2,3,4,5,6\}$ , denominamos qualquer subconjunto desse espaço amostral de evento aleatório  $A$ , isto é, um conjunto de resultados possíveis. A cada evento  $A$  associaremos um número real  $P(A)$ . Sendo  $P$  a função de probabilidade e  $P(A)$  a probabilidade do evento  $A$ .  $P(A)$  deve satisfazer os seguintes axiomas:

**Axioma 2.1** Para o evento  $A$  em  $S$ ,

$$P(A) \geq 0. \quad (2.1)$$

Isto quer dizer que não podemos ter probabilidade negativa.

**Axioma 2.2** Dado um espaço amostral  $S$  de todas as possibilidades, ao calcularmos a probabilidade desse espaço amostral o resultado será igual a 1, ou seja:

$$P(S) = 1 \quad (2.2)$$

**Axioma 2.3** Seja um conjunto de eventos disjuntos  $A_1, A_2, \dots$ . A probabilidade da união dos eventos  $A_i, i = 1, 2, \dots$  é dada pela soma das probabilidades de cada evento, isto é

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots \quad (2.3)$$

Analogamente, tem-se

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i). \quad (2.4)$$

### 2.3- Probabilidade condicional

Sejam  $A$  e  $B$  dois eventos tais que  $P(A) > 0$ . Denota-se por  $P(B|A)$  a probabilidade de  $B$  dado que  $A$  ocorreu. Como se sabe que  $A$  já ocorreu, este passa a ser o novo espaço amostral no lugar de  $S$ . Daí tem-se a definição

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (2.5)$$

### 2.4- Eventos independentes

A igualdade

$$P(A|B) = P(A) \quad (2.6)$$

significa que a probabilidade de  $A$  ocorrer visto que  $B$  ocorreu continua sendo a mesma, pois, a ocorrência de  $A$  independe de qualquer resultado de  $B$ . Analogamente, tem-se:

$$P(B|A) = P(B) \quad (2.7)$$

Afirmamos que  $A$  e  $B$  são eventos independentes quando a probabilidade de um ocorrer não depende da ocorrência do outro. Por exemplo, sejam dois eventos, um sendo o lançamento de um dado, o outro um lançamento de uma moeda, a probabilidade de obter a face quatro do dado independe da probabilidade do resultado cara no lançamento da moeda. Em linguagem matemática, tem-se:

$$P(A \cap B) = P(A)P(B) \quad (2.8)$$

A probabilidade da intersecção é igual ao produto da probabilidade de  $A$  e  $B$ . Assumindo que (2.7) é verdade dessa forma,  $A$  e  $B$  são independentes.

Em caso de três eventos  $A_1, A_2, A_3$  são independentes, se eles, dois a dois, forem independentes.

$$P(A_j \cap A_k) = P(A_j)P(A_k) \quad j \neq k \text{ com } j, k = 1, 2, 3 \quad (2.9)$$

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \quad (2.10)$$

Notar que nenhuma das equações (2.8) e (2.9) são suficientes individualmente para caracterizar a independência de três eventos. Para situações com mais de três eventos, é possível definir a independência de forma análoga (Spiegel, 2009).

## 2.5- Variável Aleatória

Uma variável aleatória (VA), ou variável estocástica representa cada ponto do espaço amostral por um número, isto é, trata-se de uma função que associa um número a um determinado evento no espaço amostral de um experimento. Pode-se chamá-la de função aleatória (função estocástica) e normalmente é representada por uma letra maiúscula,  $X$  ou  $Y$  geralmente.

Um exemplo clássico para uma melhor compreensão das VAs é o lançamento da moeda. Suponha que seja lançada uma moeda duas vezes, seu espaço amostral é  $S = \{HH, HT, TH, TT\}$ , tal que,  $H$  é cara e  $T$  é coroa. A variável  $X$  representará o número de caras enquanto  $Y$  indica o número de coroas, a cada evento será associado um número tanto para  $X$  quanto para  $Y$ , podemos observar na tabela 2.1.

Tabela 2.1: Ilustração de eventos de uma variável aleatória

Evento	$HH$	$HT$	$TH$	$TT$
$X$	2	1	1	0
$Y$	0	1	1	2

Diz-se que uma variável aleatória é discreta quando o conjunto de possíveis resultados é finito ou enumerável. Ao contrário, quando assume infinitos valores, ou finitos não enumeráveis, chamamos de variável aleatória não discreta.

## 2.6- Distribuição de probabilidade

Seja uma variável aleatória discreta  $X$ , com seus possíveis valores  $x_1, x_2, x_3, \dots$ , pode-se determinar para cada valor dessa variável uma probabilidade de ocorrência. Ao relaciona cada valor dessa variável, tem-se:

$$P(X = x_k) = f(x_k) \quad k = 1, 2 \dots \quad (2.11)$$

Dessa forma, pode-se definir o que se chama de função de probabilidade (distribuição de probabilidade),

$$P(X = x) = f(x) \quad (2.12)$$

Para que uma função  $f(x)$  possa ser uma função de probabilidade, é necessário que satisfaça as seguintes condições:

1.  $f(x) \geq 0$ ;
  2.  $\sum_x f(x) = 1$
- (2.13)

## 2.7- Função distribuição de probabilidade para VAs

A função distribuição de probabilidade, ou função cumulativa de probabilidade de uma variável aleatória  $X$  é definida por

$$F(x) = P(X \leq x) , \quad x \in \mathbb{R} . \quad (2.14)$$

A função  $F(x)$  tem as seguintes propriedades:

1.  $F(x)$  é não decrescente;
  2.  $\lim_{x \rightarrow -\infty} F(x) = 0$  e  $\lim_{x \rightarrow \infty} F(x) = 1$
  3.  $F(x)$  é contínua pela direita
- (2.15)

Para uma VA discreta, tem-se:

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u) . \quad (2.16)$$

## 2.8- Variável aleatória contínua

Uma VA não-discreta é entendida como uma VA contínua e sua distribuição de probabilidade é dada por

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du , \quad x \in \mathbb{R} \quad (2.17)$$

com  $f(x)$  satisfazendo

1.  $f(x) \geq 0$ ;
  2.  $\int_{-\infty}^{\infty} f(x) dx = 1$ .
- (2.18)

De (2.18) segue que não existe probabilidade de uma VA contínua em um valor específico, sempre é feito o cálculo de probabilidade em um determinado intervalo de valor, a saber  $[a, b]$ . Deste modo, pode-se dizer que

$$P(a < X < b) = \int_a^b f(x) dx . \quad (2.19)$$

## 2.9- Distribuições conjuntas

Até o momento foram abordados exemplos com apenas uma variável, porém, em certos experimentos pode-se estar interessado em observar mais de um resultado simultaneamente, é aí que entra a ideia de distribuição conjunta, esses exemplos podem naturalmente ser generalizados para duas ou mais variáveis sendo discretas ou contínuas (Spiegel, 2009).

### 2.9.1- Caso discreto

Sejam duas variáveis aleatórias discretas  $X$  e  $Y$ , no qual sua função de probabilidade conjunta é dada por:

$$P(X = x, Y = y) = f(x, y) \quad (2.20)$$

Essa função tem que satisfazer as seguintes propriedades:

1.  $f(x, y) \geq 0$ ;
  2.  $\sum_x \sum_y f(x, y) = 1$ .
- (2.21)

O somatório dos valores de  $x$  e  $y$  tem que ser igual a 1.

A função de probabilidade conjunta pode ser representada por uma tabela que denominamos de tabela de probabilidade conjunta. Sejam, portanto:

$$P(X = x_j) = f_1(x_j) = \sum_{k=1}^n f(x_j, y_k), \quad (2.22)$$

$$P(Y = y_k) = f_2(y_k) = \sum_{j=1}^m f(x_j, y_k). \quad (2.23)$$

A tabela 2.2 é chamada de tabela de probabilidade conjunta.

Tabela 2.2: Tabela de probabilidade conjunta

$X \backslash Y$	$y_1$	$y_2$	$\dots$	$y_n$	Totais ↓
$x_1$	$f(x_1, y_1)$	$f(x_1, y_2)$	$\dots$	$f(x_1, y_n)$	$f_1(x_1)$
$x_2$	$f(x_2, y_1)$	$f(x_2, y_2)$	$\dots$	$f(x_2, y_n)$	$f_1(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$x_m$	$f(x_m, y_1)$	$f(x_m, y_2)$	$\dots$	$f(x_m, y_n)$	$f_1(x_m)$
Totais →	$f_2(y_1)$	$f_2(y_2)$	$\dots$	$f_2(y_n)$	1

Fonte: Spiegel, 2009.

Ao observar a tabela pode-se perceber que o total de  $f_1(x_j)$  e  $f_2(y_k)$  é 1. As probabilidades obtidas nas margens da tabela referem-se às probabilidades marginais de  $X$  e  $Y$ , isto é,  $f_1(x)$  e  $f_2(y)$  e conclui-se que:

$$\sum_{j=1}^m \sum_{k=1}^n f(x_j, y_k) = 1 . \quad (2.24)$$

Com isso, tem-se que a função de distribuição de probabilidade conjunta de  $X$  e  $Y$  é definida por:

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v) \quad (2.25)$$

### 2.9.2- Caso contínuo

Diferente do caso discreto que faz o uso de somatório, no caso contínuo utiliza-se integrais. Com isso a função de probabilidade conjunta para a variáveis  $X$  e  $Y$  é definida por:

1.  $f(x, y) \geq 0$ ;
  2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$
- (2.26)

Ao analisar o gráfico de  $z = f(x, y)$ , tem-se uma superfície, que denominamos de superfície de probabilidade, no qual, o volume total limitado por esta superfície e o plano  $xy$  e sabe-se que ao calcular a integral desse plano o resultado tem que ser igual a 1. A probabilidade de  $X$  esta entra  $a$  e  $b$  e  $Y$  entre  $c$  e  $d$  e representado pelo gráfico abaixo:

$$P(a < X < b, c < Y < d) = \int_{x=a}^b \int_{y=c}^d f(x, y) dx dy . \quad (2.27)$$

### 2.10- Variáveis aleatórias independentes

Anteriormente foi definida a independência entre eventos, para variáveis aleatórias independentes a ideia é análoga. Tem-se que dois eventos são independentes se a probabilidade da intersecção for igual ao produto das probabilidades, no caso de variáveis aleatórias independentes tem-se que:

$$P(X = x, Y = y) = P(X = x)P(Y = y). \quad (2.28)$$

Diz-se que duas variáveis aleatórias são independentes se a função de probabilidade conjunta for igual ao produto das funções de probabilidade marginais.

Caso as variáveis  $X$  e  $Y$  sejam contínuas, temos que, os eventos  $X \leq x$  e  $Y \leq y$  são independentes para todo  $x$  e  $y$ .

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \quad (2.29)$$

## 2.11- Esperança matemática

A esperança matemática é bastante importante nos estudos de probabilidade e estatística, de uma forma mais sucinta pode ser denominada de esperança de uma variável aleatória. Então seja uma VA discreta  $X$  com valores possíveis  $x_1, \dots, x_n$  a esperança de  $X$  (Spiegel, 2009), é definida por:

$$E(X) = x_1P(X = x_1) + \dots + x_nP(X = x_n) = \sum_{j=1}^n x_jP(X = x_j). \quad (2.30)$$

De forma análoga, se  $P(X = x_j) = f(x_j)$

$$E(X) = x_1f(x_1) + \dots + x_nf(x_n) = \sum_{j=1}^n x_jf(x_j) = \sum xf(x). \quad (2.31)$$

Neste caso, a última soma é feita sobre todos os valores de  $x$ . Em 2.25 as probabilidades  $f(x_i)$  são todas iguais, com isso:

$$E(X) = \frac{x_1 + x_1 + \dots + x_n}{n} \quad (2.32)$$

é chamada de média aritmética.

Se uma VA discreta  $X$  assume um número infinito de valores  $x_1, x_2, \dots$ , então  $E(X) = \sum_{j=1}^{\infty} x_j f(x_j)$ , na condição de que as series converjam (Spiegel, 2009). No caso de uma VA contínua  $X$  com a função densidade  $f(x)$ , sua esperança é dada por

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (2.33)$$

na condição de que a integral converja absolutamente. A média de  $X$ , ou esperança de  $X$ , é representada por  $\mu_x$  ou somente,  $\mu$ , isso quando a variável aleatória partícula está implícita (Spiegel, 2009).

### 2.11.1- Alguns teoremas sobre esperança matemática

**Teorema 2.1:** Se  $c$  é uma constante qualquer, então

$$E(cX) = cE(X) \quad (2.34)$$

Prova:

$$\begin{aligned} E(cX) &= x_1cf(x_1) + \dots + x_ncf(x_n) \\ &= cx_1f(x_1) + \dots + cx_nf(x_n) \end{aligned}$$

$$\begin{aligned}
 &= c(x_1f(x_1) + \dots + x_nf(x_n)) \\
 &= cE(X)
 \end{aligned}$$

**Teorema 2.2:** Se  $X$  e  $Y$  são quaisquer variáveis aleatórias, então

$$E(X + Y) = E(X) + E(Y) \quad (2.35)$$

Prova:

$$\begin{aligned}
 E(X + Y) &= \sum_x \sum_y (x + y)f(x, y) \\
 &= \sum_x \sum_y xf(x, y) + \sum_x \sum_y yf(x, y) \\
 &= E(X) + E(Y)
 \end{aligned}$$

**Teorema 2.3:** Se  $X$  e  $Y$  são variáveis aleatórias independentes, então

$$E(XY) = E(X)E(Y) \quad (2.36)$$

Prova:

Pelo exposto na tabela 2.2,

$$\begin{aligned}
 E(XY) &= \sum_x \sum_y xyf(x, y) = \sum_x \sum_y xyf_1(x)f_2(y) \\
 &= \sum_x \left[ xf_1(x) \sum_y yf_2(y) \right] \\
 &= \sum_x [xf_1(x)E(y)] \\
 &= E(X)E(Y)
 \end{aligned}$$

## 2.12- Variância e desvio padrão

Outra métrica de grande importância na probabilidade e estatística é a variância, definida por:

$$Var(X) = E[(X - \mu)^2]. \quad (2.37)$$

A variância é sempre um número não negativo. A raiz quadrada da variância é chamada de desvio padrão (Spiegel, 2009), denotado por:

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{E[(X - \mu)^2]} . \quad (2.38)$$

Vale ressaltar que o desvio padrão de  $X$  é representado por  $\sigma$  ou por  $\sigma_X$ , e a variância é representada por  $\sigma^2$  (Spiegel, 2009).

Seja  $X$  uma VA discreta que assume  $x_1, x_2, \dots, x_n$  valores com sua função de probabilidade  $f(X)$ , a variância é dada por:

$$\sigma_X^2 = E[(X - \mu)^2] = \sum_{j=1}^n (x_j - \mu)^2 f(x_j) = \sum (x - \mu)^2 f(x) \quad (2.39)$$

Em (2.32) temos que as probabilidades são iguais, com isso:

$$\sigma^2 = [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2]/n \quad (2.40)$$

é a variância de um conjunto de  $n$  números  $x_1, \dots, x_n$ .

Supondo que  $X$  assumira um número infinito de valores  $x_1, x_2, \dots$ , logo  $\sigma_X^2 = \sum_{j=1}^{\infty} (x_j - \mu)^2 f(x_j)$ , contanto que as séries convirjam. Quando se tem uma VA contínua  $X$  com função densidade  $f(x)$ , a variância  $\sigma_X^2$  é dada por:

$$\sigma_X^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (2.41)$$

contanto que a integral convirja.

### 2.13- Variáveis aleatórias padronizadas

Seja uma variável aleatória  $X$  com média  $\mu$  e desvio padrão  $\sigma$  ( $\sigma > 0$ ), logo podemos definir uma variável aleatória padronizada associada dada por

$$X^* = \frac{X - \mu}{\sigma} \quad (2.42)$$

Uma propriedade importante de  $X^*$  é que ele tem média zero e variância 1, o que explica o termo “padronizado” (Spiegel, 2009), ou seja,

$$E(X^*) = 0, \quad \text{Var}(X^*) = 1 . \quad (2.43)$$

Chama-se em algumas vezes os valores de uma variável padronizada de *escores padronizados* e  $X$  é expresso em *unidades padronizadas*, i.e.,  $\sigma$  é declarado como unidade na mensuração de  $X - \mu$ .

Além das VAs padronizadas, as funções geradoras de momentos e funções características, são basilares nos estudos de probabilidade e estatística (Spiegel, 2009).

## 2.14- Variância para distribuições conjuntas e covariância

Sejam  $X$  e  $Y$  duas VAs contínuas no qual, sua função densidade conjunta  $f(x, y)$ , as esperanças ou médias são dadas por:

$$\mu_x = E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y) dx dy, \quad \mu_y = E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y) dx dy \quad (2.44)$$

e as variâncias são

$$\begin{aligned} \sigma_x^2 &= E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x, y) dx dy \\ \sigma_y^2 &= E[(Y - \mu_y)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_y)^2 f(x, y) dx dy . \end{aligned} \quad (2.45)$$

Outra métrica importante que surge de duas variáveis  $X$  e  $Y$  é a covariância determinada por:

$$\sigma_{xy} = Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] . \quad (2.46)$$

No que diz a respeito de função densidade conjunta  $f(x, y)$ . Temos:

$$\sigma_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y)f(x, y) dx dy \quad (2.47)$$

Quando se tem VAs discretas, (2.44) e (2.47) podem ser substituídas por:

$$\mu_x = \sum_x \sum_y xf(x, y) \quad \mu_y = \sum_x \sum_y yf(x, y) \quad (2.48)$$

$$\sigma_{xy} = \sum_x \sum_y (x - \mu_x)(y - \mu_y)f(x, y) \quad (2.49)$$

A seguir tem-se algumas propriedades importantes de covariância.

### Propriedade 2.1

$$\sigma_{xy} = E(XY) - E(X)E(Y) = E(XY) - \mu_x\mu_y \quad (2.50)$$

**Propriedade 2.2** Sejam  $X$  e  $Y$  VAs independentes, então

$$\sigma_{xy} = Cov(X, Y) = 0 \quad (2.51)$$

### Propriedade 2.3

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y) \quad (2.52)$$

ou

$$\sigma_{x\pm y}^2 = \sigma_x^2 + \sigma_y^2 \pm 2\sigma_{XY} \quad (2.53)$$

### Propriedade 3.7

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y \quad (2.54)$$

### 2.15- Coeficiente de correlação

Sendo  $X$  e  $Y$  duas VAs independentes, então  $Cov(X, Y) = \sigma_{XY}$ . Em contra-partida, se  $X$  e  $Y$  são totalmente dependentes, quando  $X = Y$ , logo  $Cov(X, Y) = \sigma_{XY} = \sigma_X \sigma_Y$ . Com isso, temos a *medida de independência* das variáveis  $X$  e  $Y$  (Spiegel, 2009). Dada por:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (2.55)$$

Denomina-se  $\rho$  de *coeficiente de correlação*. De 2.47 observa-se que  $-1 \leq \rho \leq 1$ . Na hipótese de  $\rho = 0$  covariância é zero, i.e., as variáveis  $X$  e  $Y$  são não correlacionadas. Nestes casos as variáveis podem ser ou não independentes (Spiegel, 2009)

### 2.16- Considerações Finais

Neste capítulo, alguns fundamentos básicos de probabilidade foram apresentados com o intuito de nortear minimamente um leitor não iniciado no assunto às bases matemáticas das aplicações apresentadas no Capítulo 4.

## Capítulo 3 – Fundamentos em Séries Temporais

### 3.1- Considerações iniciais

Segundo (Box & Jenkins, 2008), uma série temporal é uma sequência de observações feita ao longo do tempo. Dois exemplos desses conjuntos de dados são, a quantidade de acidentes que ocorrem durante a semana em uma cidade e o rendimento mensal de uma empresa. As séries temporais são bastante utilizadas em áreas como economia, engenharia, geofísica, meteorologia, dentre outras. As observações em séries temporais são tipicamente dependentes uma das outras, em particular as observações adjacentes, e a análise dessa dependência é de interesse prático. Para estudar dessas dependências são necessários modelos estacionários e dinâmicos para análise e utilização dessas séries em diferentes áreas de aplicação. Os métodos para construir, identificar, ajustar e verificar modelos para séries temporais e sistemas dinâmicos são adequados para modelos discretos, no qual, a observação ocorre em um intervalo de tempo igual (Box & Jenkins, 2008). Pode-se destacar cinco funções importantes de uma série temporal em áreas diversas:

- Previsões de valores futuros a partir de valores atuais e valores passados;
- Determinar uma função de transferência de um sistema sujeito a inércia;
- A utilização de variáveis de entrada de indicadores em modelos de funções de transferência para analisar o efeito eventos de intervenção incomuns no desempenho de uma série temporal;
- Examinar as inter-relações de diversas variáveis de importância de séries temporais compatíveis e detectar modelos dinâmicos multivariados adequados para representar essas relações conjuntas entre variáveis que variam no tempo;
- O projeto de um esquema de controle simples que possa compensar, tanto quanto possível, os desvios potenciais da saída do sistema em relação ao alvo desejado, ajustando os valores da série de entrada.

### 3.2- Previsão de séries temporais

O uso, no momento  $t$ , de observações disponíveis de uma série temporal para prever seu valor em algum momento futuro  $t + l$  pode fornecer uma base para (1) planejamento econômico e de negócios, (2) planejamento de produção, (3) estoque e controle de produção.

Supondo que as observações estão disponíveis em intervalos de tempo discretos e equidistantes como, por exemplo, em uma previsão de vendas  $Z_t$  no mês atual  $t$ , as vendas  $Z_{t-1}, Z_{t-2}, Z_{t-3}, \dots$  nos meses anteriores podem ser usadas para prever vendas para prazos de entrega  $l = 1, 2, 3, \dots, 12$  meses à frente. Denota-se por  $\hat{Z}_t(l)$  a previsão feita na origem  $t$  das vendas  $Z_{t+l}$  em algum momento futuro  $t + l$ . A função  $\hat{Z}_t(l)$ , que fornece as previsões na origem  $t$  para todos os tempos futuros, com base nas informações disponíveis dos valores atuais

e anteriores  $Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, \dots$  ao longo do tempo  $t$ , será chamada de função de previsão na origem  $t$ . O objetivo é obter uma função de previsão tal que o quadrado médio dos desvios entre os valores reais e previstos, dados por  $Z_{t+l} - \hat{Z}_t(l)$  seja o menor possível para cada valor de  $l$ .

Além de calcular as melhores previsões, é também necessário especificar a sua precisão, para que, possam ser calculados os riscos associados às decisões baseadas nas previsões. Métodos para esse tipo de previsão são desenvolvidos com base na suposição de que a série temporal  $Z_t$  segue um modelo estocástico de forma conhecida (Box & Jenkins 2008).

Uma classe útil de modelos de séries temporais que podem ser apropriados para representar o comportamento de uma série  $Z_t$  é a dos modelos autorregressivos integrados de médias móveis (ARIMA), um dos núcleos deste trabalho. Uma das aplicações mostrada no capítulo 4 é baseada em um modelo do tipo ARIMA.

### 3.3- Modelos matemáticos dinâmicos estocásticos e determinísticos

A ideia de usar modelo matemáticos para explicar comportamento de um evento físico está bem estabelecido, às vezes é possível derivar um modelo baseado em leis físicas, isto permite calcular o valor de determinada quantidade dependente do tempo, praticamente em qualquer instante (box & Jenkins 2008). Em certos problemas temos que considerar eventos que tem dependência do tempo e que não permitem escrever um modelo determinístico, no qual, permitiria um cálculo do comportamento futuro do evento. Entretanto, é possível derivar um modelo para calcular a probabilidade de um valor futuro e se estabelecer entre dois limites especificados, esse tipo modelo é denominado de modelo de probabilidade ou modelo estocástico. Com isso, uma série temporal  $Z_1, Z_2, \dots, Z_n$  de  $N$  observações consecutivas e vista como uma realização amostral de uma população de tais séries temporais que poderiam ter sido geradas por um processo estocástico (Box & Jenkins 2008).

#### 3.3.1- Modelos e operadores importantes

Modelos importantes para descrição de séries temporais são os modelos estacionários, nos quais se assume que o processo se mantém em equilíbrio estatístico, i.e., sem sofrer alterações no decorrer do tempo e que seus valores variam em torno de um nível médio fixo com variação constante. Contudo, existem fenômenos que não podem ser modelados por modelos estacionários, i.e., estas séries temporais (modelos) não têm nível médio constante natural ao longo do tempo (Box & Jenkins, 2008).

##### 3.3.1.1- Operadores deslocamento ao passado e deslocamento ao futuro

Utilizaremos amplamente o *operador de deslocamento para trás*  $B$ , definido por  $BZ_t = Z_{t-1}$ , com  $B^m Z_t = Z_{t-m}$ . A operação inversa é realizada pelo operador de *deslocamento ao futuro*  $F = B^{-1}$  dado por  $FZ_t = Z_{t+1}$  com isso,  $F^m Z_t = Z_{t+m}$ . Outro operador importante é

o operador diferença para trás, dado por  $\nabla Z_t = Z_t - Z_{t-1}$  e podendo ser escrito em termos de  $B$  (Box & Jenkins 2008):

$$\nabla Z_t = Z_t - Z_{t-1} = (1 - B)Z_t . \quad (3.1)$$

### 3.3.3.2- Modelos de filtro linear

Modelos de filtro linear são baseados na ideia de que uma série temporal observável  $Z_t$ , em que valores sucessivos são extremamente dependentes pode ser considerada como gerada diante de uma série de “choque” independentes  $a_t$ , são desenhos aleatórios de uma distribuição de probabilidade do tipo normal, com média zero e variância  $\sigma_a^2$ . Essa sequência de variáveis aleatórias independentes em  $a_t, a_{t-1}, a_{t-2}, \dots$  é denominada de processo de ruído branco (Box & Jenkins 2008).

Supõe-se que o processo de ruído branco  $a_t$  seja transformado no processo  $Z_t$  pelo que é denominado de filtro linear, esta operação considera uma soma ponderada dos choques aleatórios anteriores, de forma que

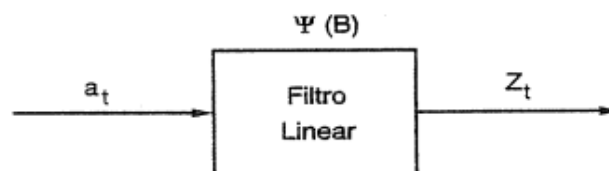
$$\begin{aligned} Z_t &= \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} \\ &= \mu + \psi(B)a_t . \end{aligned} \quad (3.2)$$

De forma geral,  $\mu$  é o parâmetro que determina/indica o nível do processo, e

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots \quad (3.3)$$

O operador  $\psi$  que transforma  $a_t$  em  $Z_t$  e é chamado de *função de transferência* do filtro. A representação em (3.2) pode fornecer uma gama flexível de padrões de dependência entre os valores do processo  $\{Z_t\}$  expressos como choque aleatório independentes (não observáveis)  $a_t$  (Box & Jenkins, 2008). Uma ilustração com o diagrama de um modelo de filtro linear é mostrada na figura 3.1.

**Figura 3.1:** Representação de uma série temporal como saída de um filtro linear. Entrada  $a_t$ , função de transferência  $\psi(B)$  e saída  $Z_t$ .



Fonte: (Morettin, 2008).

A sequência  $\psi_1, \psi_2, \dots$ , pode ser finita ou infinita. No caso dela ser finita ou infinita e completamente somável no sentido que  $\sum_{j=0}^{\infty} |\psi_j| < \infty$  o filtro é estável e o processo  $Z_t$  é estacionário e o parâmetro  $\mu$  é a média em torno da qual o processo varia, caso contrário,  $Z_t$  é

não estacionário e  $\mu$  acaba não tendo significado exato, servindo apenas como ponto de referência para o nível do processo (Box & Jenkins, 2008).

### 3.3.3.3- Modelos autorregressivos

É um tipo de modelo no qual o valor atual do processo é aparente como uma combinação linear finita de valores anteriores do processo e um choque aleatório  $a_t$ . Podemos escrever os valores de um processo em tempo de espaços iguais  $t, t - 1, t - 2, \dots$  por  $Z_t, Z_{t-1}, Z_{t-2}, \dots$  além disso, seja  $\tilde{Z}_t = Z_{t-\mu}$  a série de desvio de  $\mu$  (box & Jenkins 2008). Desta forma,

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \dots + \phi_p \tilde{Z}_{t-p} + a_t \quad (3.4)$$

é chamado um processo autoregressivo (AR) de ordem  $p$ . O motivo para esse nome é que, por ser um modelo linear da forma

$$\tilde{Z}_t = \phi_1 \tilde{X}_1 + \phi_2 \tilde{X}_2 + \dots + \phi_p \tilde{X}_p + a \quad (3.5)$$

relacionando uma variável dependente  $Z$  a um conjunto de variáveis independentes  $X_1, X_2, \dots, X_p$  mais um termo de erro  $a$ , é relacionado como um modelo de regressão, em outras palavras,  $Z$  é regredido em  $X_1, X_2, \dots, X_p$ . Em (3.4) temos que a variável  $Z$  é regredida em relação a valores passados de si mesma, daí a nomenclatura de autorregressivo. Podemos definir um operador autorregressivo de ordem  $p$  em termos do operador de deslocamento para trás  $B$  dado por

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p. \quad (3.6)$$

O modelo visto em (3.5) pode ser também escrito como

$$\phi(B)\tilde{Z}_t = a_t. \quad (3.7)$$

Este modelo tem  $p + 2$  parâmetros desconhecidos, a saber,  $\mu, \phi_1, \phi_2, \dots, \phi_p, \sigma_a^2$ , que devem ser estimados a partir dos dados. Tem-se que  $\sigma_a^2$  e a variância do processo de ruído branco  $a_t$ .

O modelo autorregressivo é um caso especial do modelo de filtro linear de (3.2). Pode-se cancelar  $\tilde{Z}_{t-1}$  ao lado direito de (3.5) substituindo.

$$\tilde{Z}_{t-1} = \phi_1 \tilde{Z}_{t-2} + \phi_2 \tilde{Z}_{t-3} + \dots + \phi_p \tilde{Z}_{t-p-1} + a_{t-1}. \quad (3.8)$$

De forma análoga, pode-se substituir  $\tilde{Z}_{t-2}$  e assim por diante, para gerar uma série infinita nos termos  $a$ . Considerando, particularmente, o processo AR simples de primeira ordem ( $p = 1$ ),  $\tilde{Z}_t = \phi \tilde{Z}_{t-1} + a_t$ , temos que, após  $m$  substituições consecutivas de  $\tilde{Z}_{t-j} = \phi \tilde{Z}_{t-j-1} + a_{t-j}, j = 1, \dots, m$ :

$$\tilde{Z}_t = \phi^{m+1} \tilde{Z}_{t-m-1} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \dots + \phi^m a_{t-m} \quad (3.9)$$

No limite, com  $m \rightarrow \infty$ , isso leva à representação convergente de série infinitas  $\tilde{Z}_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$ , com  $\psi_j = \phi^j, j \geq 1$ , uma vez que  $|\phi| < 1$  figurativamente (Box & Jenkins, 2008). No caso AR geral tem-se que:

$$\phi(B)\tilde{Z}_t = a_t \quad (3.10)$$

é equivalente a

$$\tilde{Z}_t = \phi^{-1}(B)a_t = \psi(B)a_t \quad (3.11)$$

com  $\psi(B) = \phi^{-1}(B) = \sum_{j=0}^{\infty} \psi_j B^j$

Modelos AR podem ser estacionários ou não estacionários. Para tal processo ser estacionário, os valores  $\phi$  tem que ser tais que os pesos  $\psi_1, \psi_2, \dots$  em  $\psi(B) = \phi_2 B^2 - \dots - \phi_p B^p$  formem uma série convergente. O critério necessário para a convergência dos modelos AR é que o operador autorregressivo  $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ , considerando como um polinômio em  $B$  de grau  $p$ , deve ter todas as raízes de  $\phi(B) = 0$  maior que 1 em valor absoluto, ou seja, cada raiz deve estar fora do circuito unitário (Box & Jenkins 2008).

### 3.3.3.4- Modelos de médias móveis

Outro modelo importante na representação de séries temporais, é o processo de média móvel finita, em que supomos  $\tilde{Z}_t$  linearmente dependente de um número finito  $q$  de termos  $a$  anteriores, dessa forma:

$$\tilde{Z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (3.12)$$

é chamado de média móvel (MA) de ordem  $q$ . Ao definimos um operador de média móvel de ordem  $q$  por

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q, \quad (3.13)$$

o modelo de média móvel pode ser escrito como

$$\tilde{Z}_t = \theta(B)a_t \quad (3.14)$$

e envolve  $q + 2$  parâmetros desconhecidos  $\mu, \theta_1, \dots, \theta_q, \sigma_a^2$  tem que ser estimado a partir dos dados

### 3.3.3.5- Modelos autorregressivos de médias móveis

Para obter maior versatilidade no ajuste de séries temporais reais, em certas vezes é necessário incorporar termos autorregressivos e de médias móveis no modelo, o que implica no modelo autorregressivo de média móvel (ARMA):

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \dots + \phi_p \tilde{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (3.15)$$

ou

$$\phi(B)\tilde{Z}_t = \theta(B)a_t. \quad (3.16)$$

Tal modelo utiliza  $p + q + 2$  parâmetros desconhecidos  $\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_a^2$ , que são estimados a partir dos dados (Box & Jenkins 2008).

### 3.3.3.6- Modelos não estacionários

Algumas séries expressam um comportamento não estacionário e, particularmente, não variam ao redor de uma média fixa. Elas podem, contudo, apresentar um comportamento homogêneo no decorrer do tempo, mesmo que o nível geral em relação as flutuações que estão a ocorrer possa ser diferente em momento distintos. Esse comportamento pode muitas vezes ser representado por um modelo em termos de um operador autorregressivo generalizado  $\varphi(B)$ , em que um ou mais dos zeros do polinômio  $\varphi(B)$  ficam no círculo unitário. No caso de haver  $d$  raízes unitárias e todas as outras raízes estiverem fora do círculo unitário (Box & Jenkins 2008) o operador  $\varphi(B)$  pode ser escrito

$$\varphi(B) = \varphi(B)(1 - B)^d \quad (3.17)$$

com  $\phi(B)$  sendo um operador autorregressivo estacionário. Com isso, um modelo que representa um comportamento não estacionário homogêneo é da forma

$$\varphi(B)Z_t = \varphi(B)(1 - B)^d Z_t = \theta(B)a_t. \quad (3.18)$$

Isto é,

$$\phi(B)w_t = \theta(B)a_t, \quad (3.19)$$

com

$$W_t = (1 - B)^d Z_t = \nabla^d Z_t. \quad (3.20)$$

Desta forma, o comportamento homogêneo não estacionário em algumas ocasiões é representado por um modelo que requer que a  $d$ -ésima diferença do processo seja estacionária (Box & Jenkins 2008). De costume,  $d$  é geralmente 0, 1 ou no máximo 2, com  $d = 0$  condizendo ao comportamento estacionário.

Os processos (3.19) e (3.20) proporcionam um modelo robusto para descrever séries temporais estacionárias e não estacionárias e é denominado de processo autorregressivo integrado de médias móveis (ARIMA), de ordem  $(p, d, q)$ . Este processo é definido por

$$\mathcal{W}_t = \phi \mathcal{W}_{t-1} + \dots + \phi_p \mathcal{W}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (3.21)$$

com peso  $\mathcal{W}_t = \nabla^d Z_t$ . Nota-se que ao trocar  $\mathcal{W}_t$  por  $Z_t - \mu$ , enquanto  $d = 0$ , o modelo (3.21) engloba o modelo ARMA estacionário (3.15) e, como caso especial, inclusive, o modelo AR puro (3.4) e o modelo MA puro (3.12). A relação, que é inversa a (3.20), é  $Z_t = S^d \mathcal{W}_t$ , onde  $S = \nabla^{-1} = (1 - B)^{-1} = 1 + B + B^2 + \dots$  é o operador de soma definido por

$$S\mathcal{W}_t = \sum_{j=0}^{\infty} \mathcal{W}_{t-j} = \mathcal{W}_t + \mathcal{W}_{t-1} + \mathcal{W}_{t-2} + \dots \quad (3.22)$$

Com isso, o processo ARIMA geral pode ser gerado somando ou “integrando” o processo ARMA estacionário  $\mathcal{W}_t$ ,  $d$  vezes (Box & Jenkins 2008).

### 3.4- Séries temporais

Uma série temporal é um conjunto de observações feitas sequencialmente no tempo. Essas observações podem ser contínuas ou discretas, dependendo se ocorrer de forma contínua ou em intervalos discretos. As observações de uma série temporal no caso discreto nos momentos  $\tau_1, \tau_2, \dots, \tau_t, \dots, \tau_N$ , podem ser denotados por  $Z(\tau_1), Z(\tau_2), \dots, Z(\tau_t), \dots, Z(\tau_N)$ , na análise das séries discreta as observações tem um intervalo fixo  $h$ . Quando temos  $N$  valores disponíveis para análise desta série o representamos por  $z_1, z_2, \dots, z_t, \dots, z_N$ , valores esses obtidos em intervalo de tempo equidistante  $\tau_0 + h, \tau_0 + 2h, \dots, \tau_0 + th, \dots, \tau_0 + Nh$ . Observando os valores  $\tau_0$  e  $h$ , consideramos  $\tau_0$  como origem e  $h$  como a unidade de tempo, podemos interpretar  $Z_t$  como observação no tempo  $t$  (box & Jenkins, 2008).

Em série temporal determinística e estatística, caso os valores futuros forem determinados por uma função matemática tal como

$$z_t = \cos(2\pi ft) \quad (3.23)$$

dizemos que a série temporal é considerada determinística, em caso dos valores futuros só puderem ser escritos em forma de distribuição de probabilidade, uma série temporal é dita não determinística ou também série temporal estatística.

### 3.5- Processos estocásticos estacionários

Em processos estocásticos temos uma classe especial chamada de processos estacionários, baseada na suposição de que o processo está num determinado estado de equilíbrio estatístico. No caso de as propriedades de um processo estocástico não serem afetados por uma mudança na origem do tempo dizemos que é um processo estritamente estacionário, quer dizer, se a distribuição de probabilidade conjunta associada a  $m$  observações  $Z_{t_1}, Z_{t_2}, \dots, Z_{t_m}$  feitas em qualquer conjunto de tempo  $t_1, t_2, \dots, t_m$  é associada a  $m$  observações  $Z_{t_1+k}, Z_{t_2+k}, \dots, Z_{t_m+k}$  feito nos tempos  $Z_{t_1+k}, Z_{t_2+k}, \dots, Z_{t_m+k}$ . Com isso, para um processo discreto estritamente estacionário suas observações não podem sofrer alterações pelo deslocamento dos tempos de observações para frente ou para trás por qualquer valor  $k$  (Box & Jenkins, 2008).

#### 3.5.1- Média e variância de um processo estacionário

Na ocasião em que  $m = 1$ , o pressuposto de estacionariedade implica que a distribuição de probabilidade  $p(Z_t)$  é a mesma para todos os tempos  $t$  que podemos escrever sendo  $p(Z)$ . Com isso, processo estocástico tem uma média constante dada por

$$\mu = E[Z_t] = \int_{-\infty}^{\infty} Z p(Z) dz, \quad (3.24)$$

que, por sua vez, define o nível sob qual ela “flutua” e uma variância constante

$$\sigma_Z^2 = E[(Z_t - \mu)^2] = \int_{-\infty}^{\infty} (Z - \mu)^2 p(Z) dz \quad (3.25)$$

que mede sua dispersão em torno deste nível.

Com a distribuição de probabilidade  $p(Z)$  sendo a mesma para todo tempo  $t$ , a sua forma pode ser deduzida por meio da elaboração do histograma das observações  $Z_1, Z_2, \dots, Z_n$ , que compõem a série temporal observada, onde a média  $\mu$  do processo estocástico pode ser calculada por meio da média amostral

$$\bar{Z} = \frac{1}{N} \sum_{t=1}^N Z_t. \quad (3.26)$$

A variância  $\sigma_Z^2$  do processo estocástico pode ser calculada pela variância amostral da série temporal dada por

$$\hat{\sigma}_Z^2 = \frac{1}{N} \sum_{t=1}^N (z_t - \bar{z})^2. \quad (3.27)$$

### 3.5.2- Coeficientes de autocovariância e autocorrelação

A condição de estacionariedade implica que a distribuição de probabilidade conjunta  $p(Z_{t_1}, Z_{t_2})$  é a mesma para os tempos  $t_1, t_2$  sendo eles separados por um intervalo constante. A covariância entre os valores  $Z_t$  e  $Z_{t+k}$  separados por  $k$  intervalos de tempo ou por um *lag*  $k$ , tem de ser igual para todo  $t$  sob a hipótese de estacionariedade. Chamamos esta covariância de autocovariância no atraso  $k$  (Box & Jenkins, 2008) que é definida por:

$$\gamma_t = COV[Z_t, Z_{t+k}] = E[(Z_t - \mu)(Z_{t+k} - \mu)]. \quad (3.28)$$

De forma análoga, a autocorrelação no *lag*  $k$  é

$$\begin{aligned} \rho_k &= \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sqrt{E[(Z_t - \mu)^2]E[(Z_{t+k} - \mu)^2]}} \\ &= \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sigma_Z^2} \end{aligned} \quad (3.29)$$

já que em um processo estacionário, a variância  $\sigma_Z^2 = \gamma_0$  é igual à do tempo  $t + k$  tal qual no momento  $t$ . Com isso, a autocorrelação no *lag*  $k$ , isto é, a correlação entre  $Z_t$  e  $Z_{t+k}$ , é

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (3.30)$$

e isto implica que  $\rho_0 = 1$ .

### 3.6- Estacionariedade fraca

Para um processo ser estritamente estacionário toda estrutura de probabilidade deve depender apenas das variações temporais. Um requisito menos restritivo, denominado de **estacionariedade fraca** de ordem  $f$ , no qual, os momentos até uma certa ordem  $f$  dependam apenas das variações temporais. Por exemplo, suponhamos que exista uma média fixa  $\mu$  e uma matriz de autocovariância  $\Gamma_n$  da forma

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \dots & \gamma_{n-1} \end{bmatrix} \quad (3.31)$$

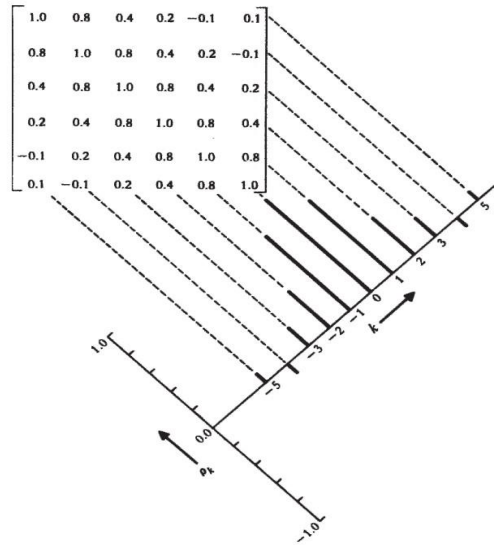
$$= \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix}$$

Isto é o bastante para garantir estacionariedade até a segunda ordem, ou seja, um processo  $\{Z_t\}$  é de natureza estacionaria fraca (de ordem 2), se a media  $E[Z_t] = \mu$  for constante fixa para todo  $t$  e as autocovariâncias  $COV[Z_t, Z_{t+k}] = \gamma_k$  apenas dependerem da diferença temporal ou do desfasamento (*lag*)  $k$  para todo  $t$ . Com isso, temos que a estacionariedade (de ordem 2), e a hipótese de normalidade, são o bastante para gerar estacionariedade estrita

### 3.7- Funções de autocovariância e autocorrelação

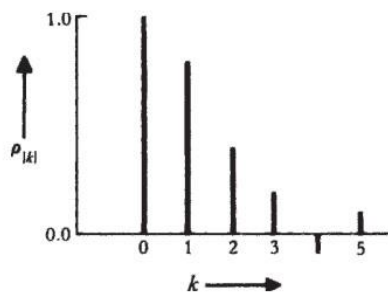
Sabe-se que o coeficiente  $\gamma_k$ , no *lag*  $k$ , mede a covariância de dois valores  $Z_t$  e  $Z_{t+k}$  a uma distância  $k$ . O gráfico de  $\gamma_k$  em função do *lag*  $k$  refere-se à função de autocovariância  $\{\gamma_k\}$  do processo estocástico. De forma análoga, o gráfico do coeficiente  $\rho_k$  em função do *lag*  $k$  refere-se à função de autocorrelação. Esta função não tem dimensão, o que significa que ela é independente da escala de medida de série temporal. Visto que,  $\gamma_k = \rho_k \sigma_Z^2$ , o conhecimento da função de autocorrelação  $\{\rho_k\}$ , e da variância  $\sigma_Z^2$  é similar ao conhecimento da função autocovariância  $\{\gamma_k\}$ . Logo abaixo temos a figura 3.2 que representa a função de autocorrelação como um gráfico das diagonais da matriz de autocorrelação em que revela como a correlação entre dois valores da série altera à medida que a sua separação muda, visto que,  $\rho_k = \rho_{-k}$  função de autocorrelação e simetria em relação a zero, em prática e necessário apresentar a metade positiva da função (Box & Jenkins, 2008), figura 3.3.

Figura 3.2. Matriz de autocorrelação e função de autocorrelação correspondente de um processo estacionário.



Fonte: Box & Jenkins, 2008.

Figura 3.3. Metade positiva da função de autocorrelação.



Fonte: Box & Jenkins, 2008.

Quando se fala da função de autocorrelação, refere-se à metade positiva desta função. Alguns autores entendem que a função de autocorrelação deva ser chamada de correlograma.

Em suma, um processo normal estacionário  $Z_t$  é definido pela média  $\mu$  e pela função de autocovariância  $\{\gamma_k\}$ , ou equivalentemente pela média  $\mu$ , variância  $\sigma_Z^2$ , e função de autocorrelação  $\{\rho_k\}$ .

### 3.8- Considerações finais

Neste capítulo, alguns fundamentos básicos de séries temporais foram apresentados com o intuito de nortear minimamente um leitor não iniciado no assunto às bases matemáticas das aplicações apresentadas no Capítulo 4.

## Capítulo 4 – Aplicações e seus resultados

### 4.1- Considerações Iniciais

Neste capítulo, as aplicações abordadas neste trabalho são apresentadas, isto é, uma aplicação referente ao coeficiente de correlação entre duas variáveis aleatórias e uma aplicação de um modelo ARIMA no fenômeno de propagação de ondas eletromagnéticas.

### 4.2- Modelo ARIMA aplicado à propagação de ondas eletromagnéticas – Parte 1

Nesta subseção, dois trabalhos com aplicação de modelos ARIMA à propagação eletromagnética no contexto das telecomunicações serão mostrados. Estes artigos formaram a primeira sequência de trabalhos nos quais o autor participou junto a seu orientador e ambos tratam de replicações da modelagem apresentada em (Fraiha, 2020) em outro cenário.

#### 4.2.1- Introdução

O trabalho (Barbosa, 2021) mostra a aplicação da modelagem proposta em (Fraiha, 2020) em um bairro de Salinópolis. As principais diferenças se dão no tamanho da região considerada e nos tipos de construção predominantes. Em (Barbosa, 2021) uma modelagem híbrida ARIMA – RNA (Rede Neural Artificial) foi aplicada, para a previsão de potência de sinal em um receptor (Rx), como uma função da distância em visada direta até o transmissor (Tx) na faixa de frequência entre 869 MHz e 880 MHz.

#### 4.2.2- Modelo híbrido ARIMA-RNA

Modelos do tipo ARIMA são baseados na teoria de séries temporais e são modelos lineares. Ao considerar um fenômeno natural qualquer (neste caso a propagação eletromagnética), pode-se decompô-lo em três componentes: uma linear, uma não linear e outra aleatória (ou composta de ruído branco). Por serem do tipo linear, modelos ARIMA conseguem prever uma destas componentes de forma confiável. Já uma rede neural artificial (RNA) é uma ferramenta de modelagem que consegue, dentre outras coisas, prever fenômenos não lineares (Haykin, 2000). Para esta finalidade é que se insere uma RNA na modelagem proposta por (Fraiha, 2020). A ideia é que o que o ARIMA não consegue modelar, a RNA consiga, de modo a obter um modelo mais preciso de um fenômeno que tem uma componente não linear com peso significativo. O modelo híbrido utilizado, proposto inicialmente em (Fraiha, 2020) pode ser matematicamente escrito como

$$Z = L + N, \quad (4.1)$$

com  $Z$  sendo a variável modelada,  $L$  a componente linear desta variável e  $N$  a componente não-linear desta variável. Comparações com outros modelos da literatura, a saber o modelo de

Okumura-Hata (OH) (Haykin, 2008) e o modelo ITU P.1546-6 (ITU, 2013), foram realizadas de modo a ratificar a escolha do modelo proposto em (Fraiha, 2020)

### 4.2.3- Campanhas de medição

Uma campanha de medição de intensidade de sinal foi realizada em um bairro residencial do centro de Salinópolis/PA, conforme mostrado na figura 4.1.

Figura 4.1. Cenário de Medições. Ruas destacadas em azul foram locais de medição. Ruas em Vermelho indicam a delimitação de toda a área idealizada para medições. Marcadores em amarelo indicam os pontos de coleta de dados. O quadrado preto e amarelo indica a posição do Tx.



Fonte: Barbosa, 2021, Google Earth, 2021.

De toda a região delimitada pela fronteira destacada em vermelho, apenas a sub-região com as ruas em azul com os marcadores em amarelo foi considerada para este trabalho. Nestas ruas, cada marcador em amarelo representa um ponto onde a intensidade de sinal foi registrada. Foi utilizado um celular Samsung J5 Prime<sup>®</sup> com a versão mais recente do sistema operacional Android<sup>®</sup> e o aplicativo Network Cell Info (Network Cell Info Lite, 2020) para capturar a informação da intensidade de sinal e gerar um arquivo *log* no formato *.csv*. Dados de apenas uma operadora de telefonia foram coletados, visto que o aplicativo registra apenas as informações recebidas pelo aparelho celular receptor (que tem apenas um *chip SIM* instalado).

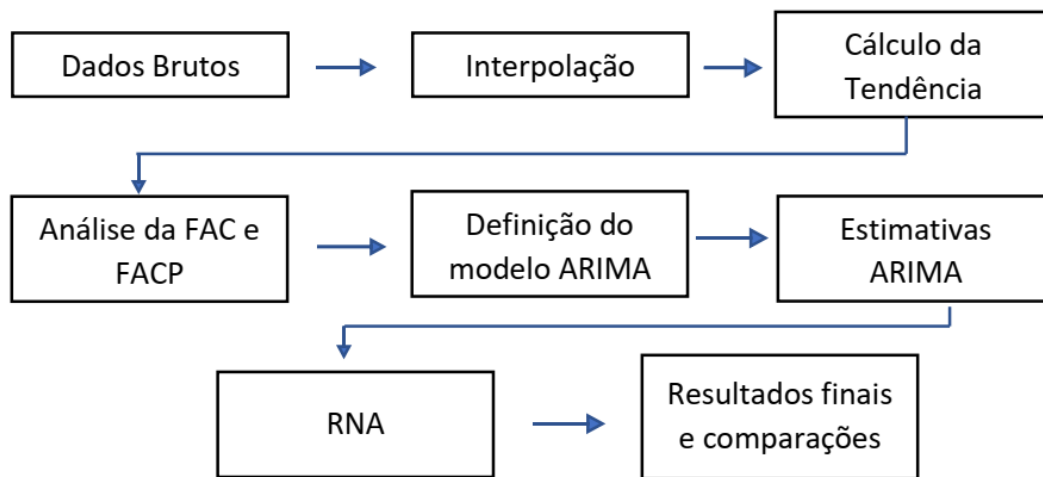
A campanha de medição foi realizada durante o dia e com tempo bom e com poucas nuvens. O receptor estava a uma altura de 1,5 m do chão (simulando a altura de um aparelho em uso por uma pessoa) e em cada ponto (marcador amarelo), foi coletada a informação de intensidade de sinal durante dois minutos. Os pontos de coleta estavam a aproximadamente 50 m de distância um do outro, considerando cada rua individualmente. Cada rua gerou um vetor de dados brutos.

### 4.2.4- Tratamento dos dados

Primeiramente, foi realizado a ordenação dos pontos medidos tendo como referência a distância em visada direta ao Tx (quadrado em preto e amarelo na figura 4.1). Em seguida, como detalhado em (Fraiha, 2020), fez-se necessário o acréscimo de amostras para 200 em cada vetor de dados brutos através de uma interpolação cúbica por partes com preservação de forma

(aqui abreviado como SPPCI). O acréscimo de amostras em cada um dos vetores de dados brutos representa uma analogia a terem sido medidos 200 pontos em cada uma das três ruas. Esta necessidade de acréscimo de amostras surge devido à quantidade de amostras recomendada para que um conjunto de dados (em estatística, população) possa ser representativo de um fenômeno qualquer ser de pelo menos 30 amostras (Spiegel, 2009). O fluxograma da metodologia de tratamento dos dados desde sua forma bruta até os resultados finais é mostrado na figura 4.2.

Figura 4.2. Fluxograma da metodologia desta aplicação.



Fonte: Inspirado em (Fraiha, 2020).

Dos três vetores de dados (após a interpolação), um deles será utilizado para treinar o modelo e os outros dois servirão como comparação às previsões do modelo híbrido. Todos os cálculos foram realizados no software MATLAB<sup>®</sup> (Matlab, 2012)

#### 4.2.5- Resultados e conclusões

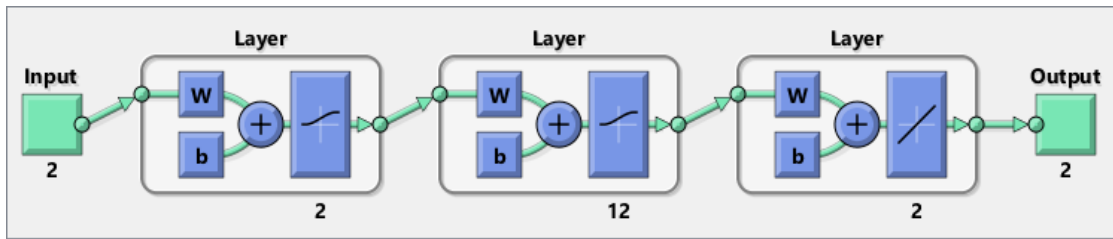
Os resultados desta seção foram obtidos de um modelo híbrido ARIMA-ANN cuja componente linear, i.e., a parte ARIMA do modelo é representada por um modelo ARIMA (4,0,0), ou seja, tem-se, na verdade, um modelo AR(4). A equação (4.2) indica este modelo

$$Y_h = \phi_1 Y_{h-1} + \phi_2 Y_{h-2} + \phi_3 Y_{h-3} + \phi_4 Y_{h-4} + c, \quad (4.2)$$

com  $\phi_1 = 3.35$ ,  $\phi_2 = -4.13$ ,  $\phi_3 = 2.18$ ,  $\phi_4 = -0.41$  e  $c = 0$ .

A RNA utilizada é do tipo *feedforward* e sua arquitetura de camadas e neurônios é mostrada no diagrama da figura 4.3.

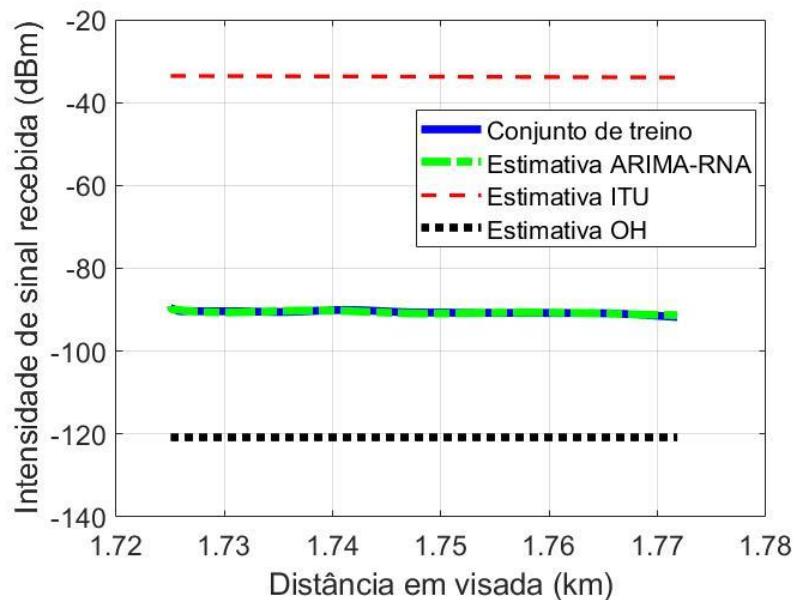
Figura 4.3. Diagrama da RNA.



Fonte: Barbosa, 2021.

Na figura 4.4, são mostradas as estimativas dos modelos OH e ITU e do modelo híbrido para o conjunto de dados 1, i.e., o conjunto de treino. O modelo OH para este cenário foi calibrado para a configuração de ambiente suburbano e cenário de pequena/média cidade.

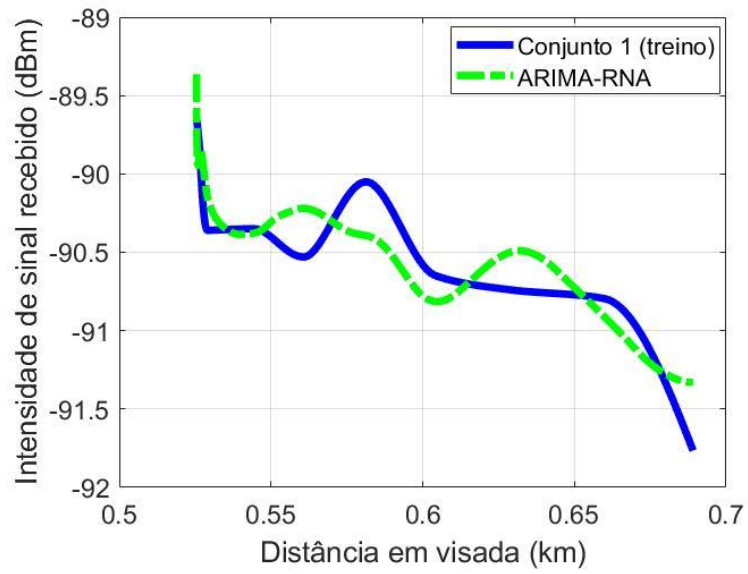
Figura 4.4. Conjunto de treino, seu respectivo ajuste e comparações com os modelos OH e ITU.



Fonte: Barbosa, 2021.

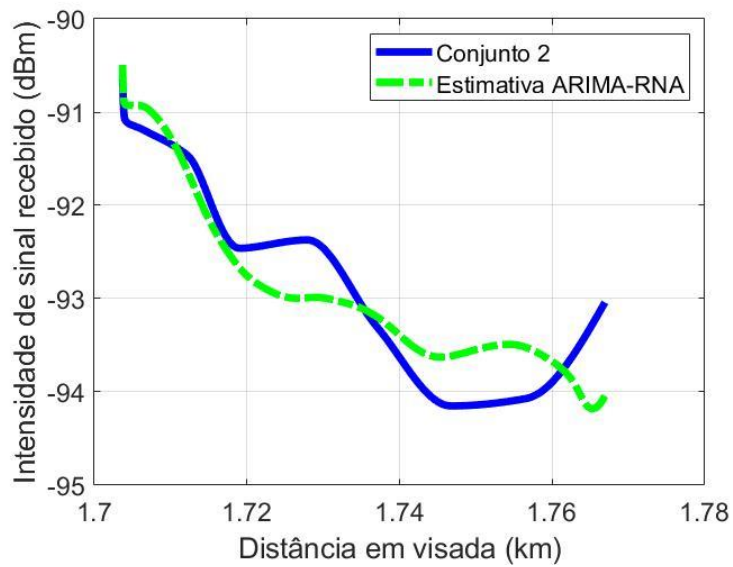
Pode-se perceber que os modelos da literatura escolhidos não se comportaram bem em relação aos valores efetivamente medidos de intensidade de potência de sinal no cenário analisado. Desta forma, nas figuras 4.5, 4.6 e 4.7 são mostradas as comparações do modelo híbrido com cada um dos três conjuntos de dados: o de treino e os dois de comparação. Chama-se de conjunto 1 ao conjunto de treino, e de conjunto 2 e conjunto 3 aos dados de comparação. Na Tabela 4.1 são mostrados os valores de erro médio quadrático (ERMS) das estimativas/ajustes em relação aos dados medidos (conjuntos 1, 2 e 3).

Figura 4.5. Conjunto de treino e ajuste do modelo híbrido.



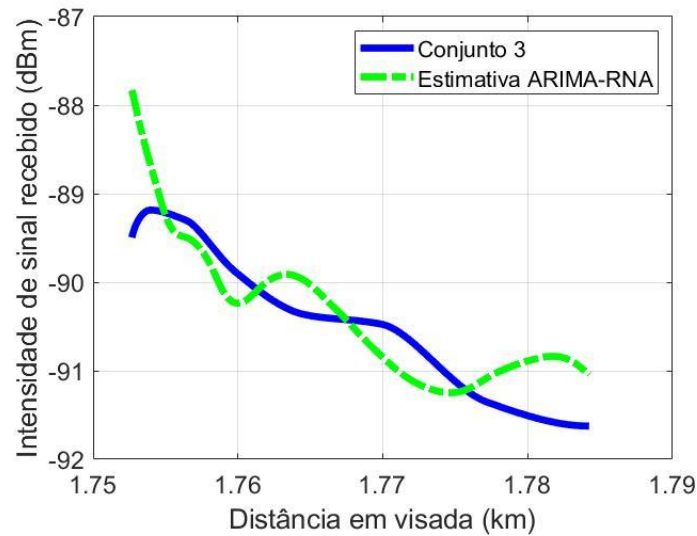
Fonte: Barbosa, 2021.

Figura 4.6. Conjunto 2 e estimativa do modelo híbrido.



Fonte: Barbosa, 2021.

Figura 4.7. Conjunto 3 e estimativa do modelo híbrido.



Fonte: Barbosa, 2021.

Tabela 4.1: Valores de ERMS entre as estimativas híbridas e os dados medidos

Conjunto	ERMS
Conjunto 1	0.18
Conjunto 2	0.39
Conjunto 3	0.54

Fonte: Barbosa, 2021.

Percebe-se que o modelo híbrido obteve bons resultados nos conjuntos de dados abordados nesta aplicação. Um destaque deve ser feito ao conjunto 3, que, devido ao ponto inicial da estimativa ter um valor maior que o valor medido, ilustrando um contraponto da componente linear, representada por um modelo do tipo ARIMA, que preservam os valores iniciais do conjunto de treino, o ERMS deste conjunto ficou perceptivelmente maior que os outros dois conjuntos, ainda que seja um valor absoluto de ERMS considerado baixo. Outro contraponto dos modelos ARIMA que esta modelagem carrega é que, em situações como esta, para retirar os pontos nitidamente discrepantes, perder-se-ia aproximadamente 25% das amostras comparativas. Conclui-se então que para este cenário deve-se ter mais dados (inclusive para treino) que permitam encontrar um ajuste (treino) mais representativo do ambiente estudado. Por outro lado, a variabilidade das previsões encontradas em trabalhos de campo (dados empíricos) é algo esperado, visto que por não se tratar de um ambiente controlado, a componente aleatória assume um caráter ainda mais significativo em medições de campo. Desta forma, considerando o cenário geograficamente pequeno estudado, entende-se que os três conjuntos de dados coletados dão uma boa representação da propagação de ondas eletromagnéticas na faixa de frequência entre 869 MHz e 880 MHz

## **4.3- Modelo ARIMA aplicado à propagação de ondas eletromagnéticas – Parte 2**

### **4.3.1- Introdução**

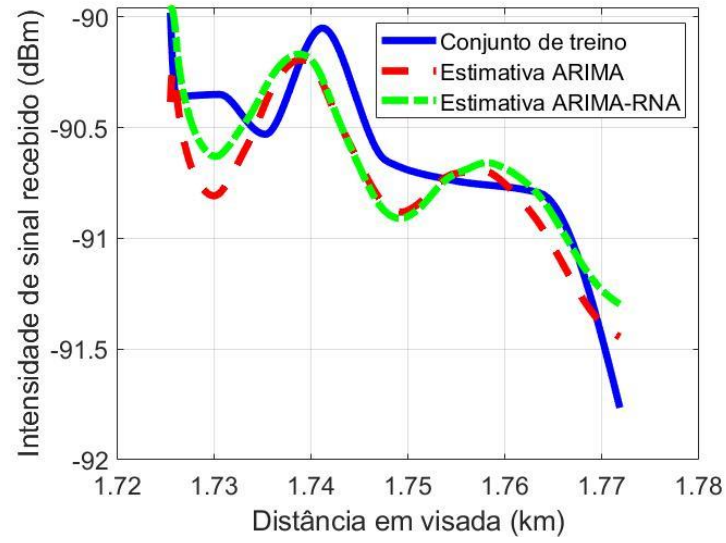
Esta segunda aplicação da modelagem ARIMA ao fenômeno da propagação eletromagnética é uma continuação da aplicação mostrada na seção 4.2. Aqui o objetivo é comparar se, para o cenário estudado na seção 4.2, apenas um modelo linear, i.e., sem a componente híbrida seria suficiente para representar o ambiente. O trabalho (BARBOSA, 2021) mostrou os resultados somente do modelo híbrido aplicado a este mesmo cenário. Contudo, uma comparação entre os modelos linear e híbrido é válida neste contexto a título de verificação da hipótese de que um modelo mais simples (com menor capacidade de modelar o fenômeno como um todo) é capaz de trazer bons resultados se aplicado a um cenário reduzido. Quando se modela um fenômeno qualquer, busca-se sempre uma boa relação de compromisso entre precisão do modelo e custo de implementação (custo computacional, custo financeiro para aquisição de dados ou de hardware).

Toda parte metodológica e de tratamento de dados seguiu os mesmos procedimentos de (Barbosa, 2021), desta forma, partir-se-á diretamente aos resultados obtidos.

### **4.3.2- Resultados e conclusões**

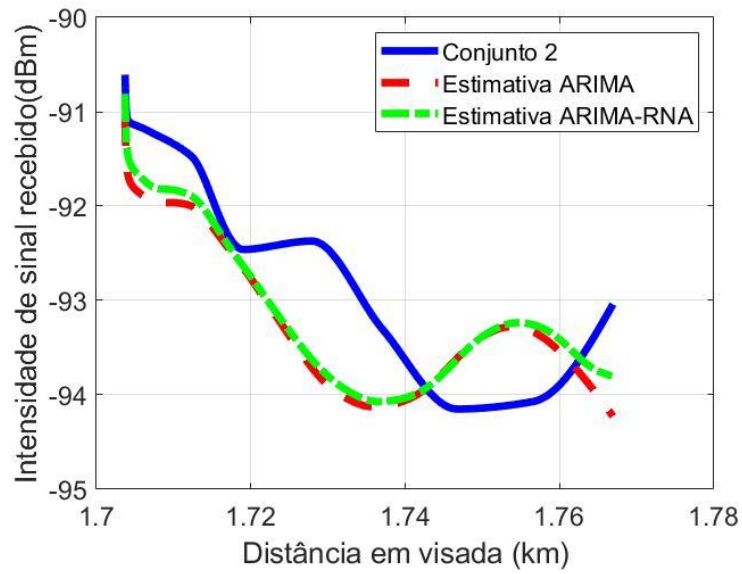
Nesta seção são comparados os resultados do modelo híbrido ARIMA-RNA com os de um modelo composto somente pela equação (4.2). Estes resultados são mostrados nas figuras 4.8, 4.9 e 4.10 e na Tabela 4.2.

Figura 4.8. Comparações com o conjunto de treino.



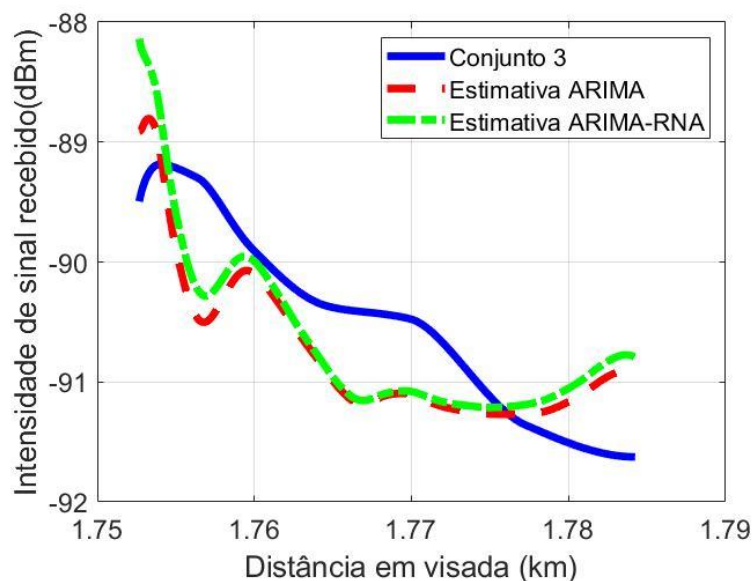
Fonte: Barbosa, 2022.

Figura 4.9. Comparações com o conjunto 2.



Fonte: Barbosa, 2022.

Figura 4.10 Comparações com o conjunto 3.



Fonte: Barbosa, 2022.

Tabela 4.2. Erros RMS do modelo linear e do modelo híbrido.

Conjunto de dados	ERMS ARIMA	ERMS ARTIMA-RNA
Conjunto 1 (treino)	0.2119	0.1880
Conjunto 2	0.7244	0.6496
Conjunto 3	0.5561	0.6163

Fonte: Barbosa, 2021.

Dos resultados expostos, conclui-se que ambos os modelos atingem resultados semelhantes. Logo, o modelo mais simples pode ser aplicado sem prejuízo de eficiência no cenário considerado (Barbosa, 2022). Outro ponto importante a ser destacado é a baixa amplitude dos valores dos dados coletados. Ainda que a perspectiva dos gráficos indique uma tendência decrescente notável, a amplitude dos valores diz o contrário. Isto pode ter contribuído para os resultados praticamente iguais entre o modelo ARIMA e o ARIMA-RNA. Trabalhos como (Fraiha, 2020) já mostram que em situações menos convenientes, o modelo ARIMA já não produz resultados satisfatórios, evidenciando um complemento a partir de outra metodologia, como as RNAs

#### 4.4- Cálculo do Coeficiente de Correlação entre duas VAs

##### 4.4.1- Introdução

O trabalho (Fraiha et. al, 2023) tem por objetivo analisar a existência de correlação significativa entre renda per capita (RPC) e as notas de português e matemática de alguns municípios do estado do Pará a partir de dados do Sistema de Avaliação da Educação Básica (Saeb) e do Instituto Brasileiro de Geografia e Estatística (IBGE). O estudo concentra dados

dos anos de 2015, 2017 e 2019, e realiza análises de correlação linear entre as notas das médias de proficiência em língua portuguesa e matemática e a renda per capita dos municípios selecionados. Duas rodadas de análise de dados foram conduzidas, a primeira com sete municípios da Região Metropolitana de Belém do Pará, e a segunda com 21 municípios, incluindo os mais populosos do estado (Fraíha et. al, 2023).

Nas partes iniciais do artigo é feita uma breve exposição sobre o histórico das avaliações dos sistemas educacionais no Brasil e no quadro geral da educação básica brasileira.

#### 4.4.2- Metodologia e tratamento dos dados

A escala de proficiência em língua portuguesa e matemática é dividida em níveis que vão de 1 a 9 para língua portuguesa e de 1 a 10 para matemática. Um exemplo da escala interpretada de Matemática adotada pelo Saeb de 2019 é mostrado na Tabela 4.3.

Tabela 4.3: Escala interpretada

<b>Nível</b>	<b>Descrição das habilidades desenvolvidas</b>
Nível 1 Desempenho maior do que ou igual à 125 e menor do que 150	Os estudantes provavelmente são capazes de: Grandezas e medidas – determinar a área de figuras desenhadas em malhas quadriculadas por meio de contagem
...	...
Nível 10 Desempenho maior do que ou igual a 350	... os estudantes provavelmente são capazes de: Espaço e Forma – Reconhecer entre um conjunto de quadriláteros aquele que possui lados perpendiculares com a mesma medida...

Fonte: Relatório de Resultados do Saeb referentes a 2019 (Saeb, 2023).

Em (Fraíha et. al, 2023), métricas de proficiência em língua portuguesa e matemática de alunos do 5º ano do ensino fundamental, obtidas dos dados do Saeb, são analisadas em relação à renda per capita de municípios do Estado do Pará. Os dados de proficiência em português e matemática foram obtidos das bases de dados do Saeb e os dados de renda per capita foram obtidos a partir do IBGE (IBGE, 2023).

Inicialmente coletou-se os dados da “Prova Brasil” na base de dados do Saeb. Em seguida organizou-se as notas de português e matemática por município, considerando as “médias totais” (média entre as notas das zonas urbanas e rurais) como valores de referência. Para cada município estudado, tem-se, então, um vetor com 9 coordenadas: notas de português de 2015, 2017 e 2019 e notas de matemática de 2015, 2017 e 2019 (Fraíha et. al, 2023). Estas matrizes foram inseridas em um código fonte elaborado utilizando o software MATLAB® para o cálculo da correlação linear (também conhecido como coeficiente de correlação Pearson) segundo a equação (4.3). Como resultado tem-se a matriz de correlação das variáveis analisadas. São 9 variáveis, portanto tem-se uma matriz de correlação 9x9. Os elementos indicando as correlações das notas de cada uma das matérias e das respectivas rendas per capita são os elementos alvo da análise deste trabalho.

$$\rho(A, B) = \frac{\text{cov}(A_i, B_j)}{\sigma_A \sigma_B}, \quad i, j = 1, 2, 3. \quad (4.3)$$

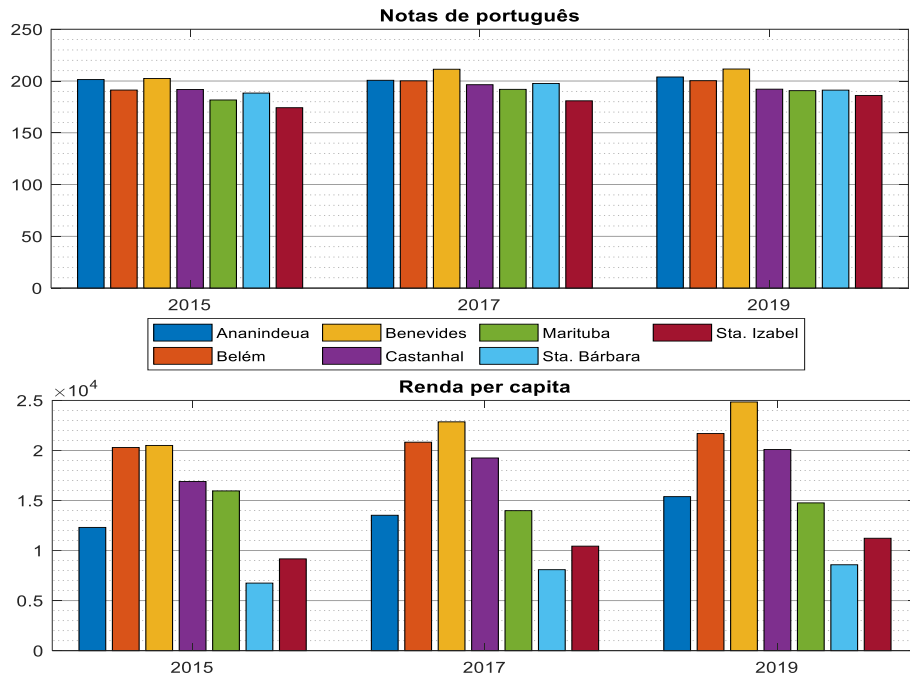
Ao se realizar os cálculos das correlações, fez também um teste de hipótese, sendo a hipótese nula a não existência de correlação entre as duas variáveis comparadas. Todo teste de hipótese tem um nível de confiança, representado por uma variável auxiliar. No caso deste trabalho, a variável auxiliar é a variável  $p$ . Idealmente seu valor deve ser menor do que, ou igual a, 0,05. Isto indica que há 5% de chance de que a hipótese nula se verifique, ou seja, há 5% de chance de não haver correlação entre as variáveis comparadas. Em suma, para um valor de correlação ser considerado estatisticamente significativo, o valor  $p$  associado a um valor de correlação  $\rho$  não deve ser superior a 0,05. Concomitantemente à matriz de correlações, foi gerada a matriz com os valores  $p$  também de dimensão 9x9 (Fraiha et. al, 2023).

#### **4.2.3- Resultados e conclusões**

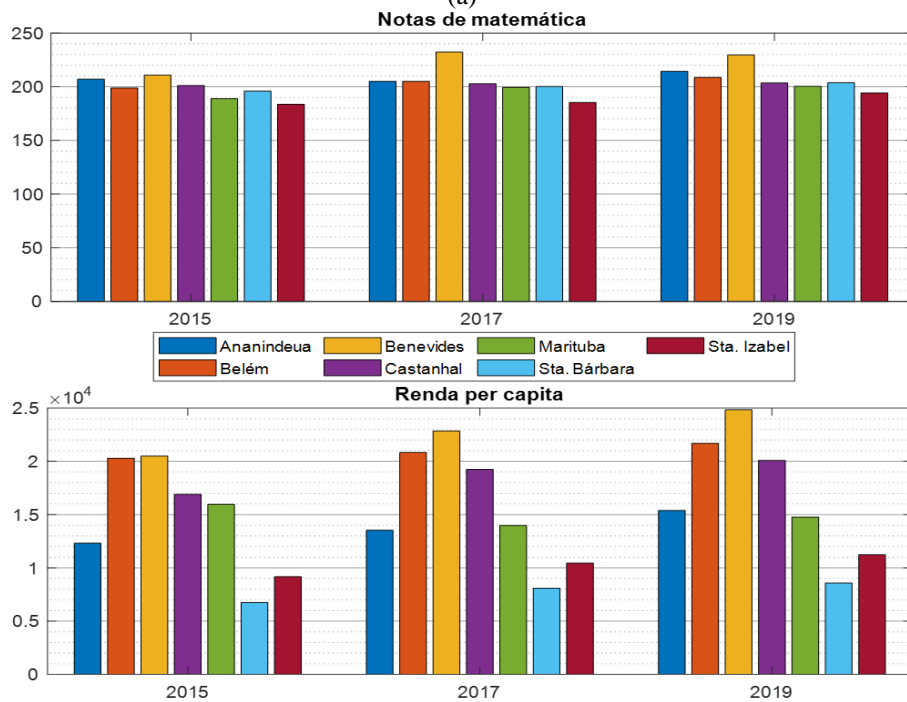
Na primeira rodada de análise foram considerados os municípios de Ananindeua, Belém, Benevides, Castanhal, Santa Bárbara e Santa Izabel, todos situados na Região Metropolitana de Belém (RMB). Os resultados para esta rodada de análise são mostrados na figura 4.11 e na Tabela 4.4.

Na segunda rodada de análise foram considerados os municípios da primeira rodada mais outros 14 municípios de outras regiões do estado. Os resultados para esta rodada de análise são mostrados na figura 4.2 e na Tabela 4.2.

Figura 4.11. Resultados notas e renda per capita de 7 municípios. (a) Notas de português (b) Notas de matemática.



(a)



(b)

Fonte: Fraiha et. al, 2023.

Tabela 4.4: Dados de correlação e parâmetro  $p$  para 7 municípios

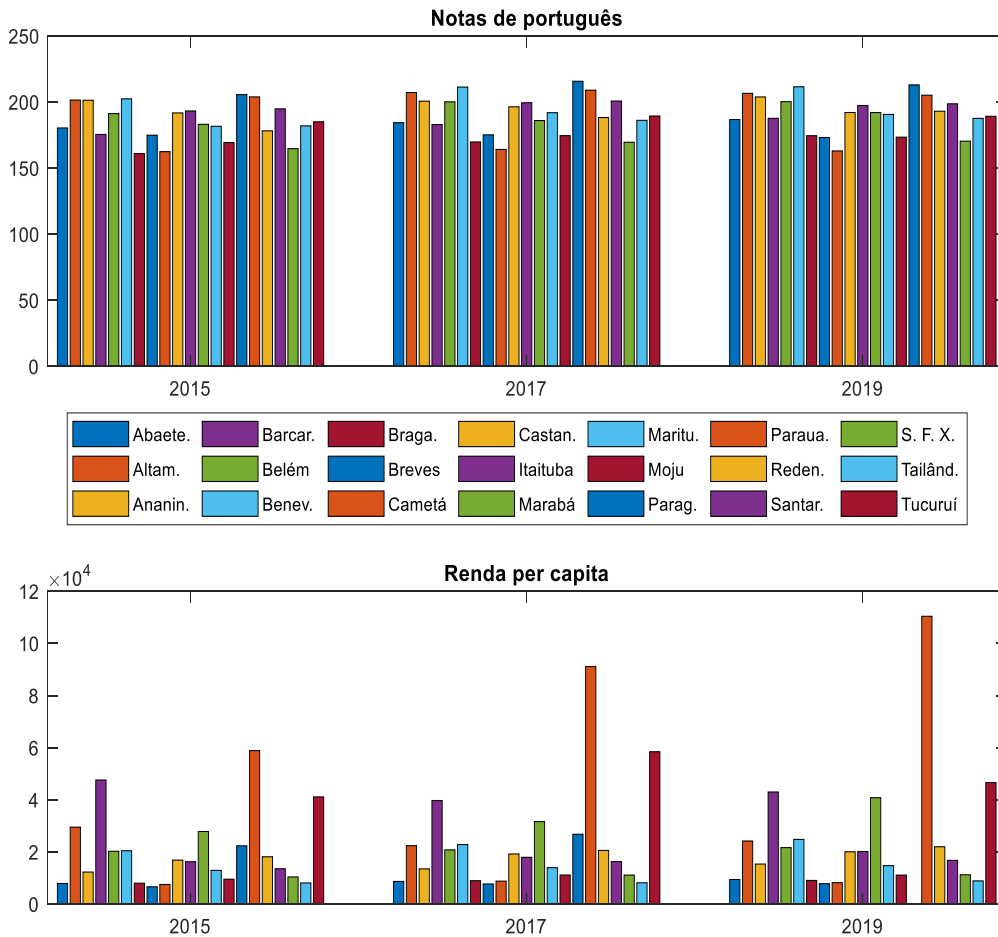
Correlação Português e RPC			Correlação Matemática e RPC		
2015	2017	2019	2015	2017	2019

0.4532	0.6432	0.7258	0.4853	0.7216	0.6957
Valores $p_k$ português e RPC			Valores $p_k$ Matemática e RPC		
2015	2017	2019	2015	2017	2019
0.3072	0.1192	0.0648	0.2696	0.0671	0.0826

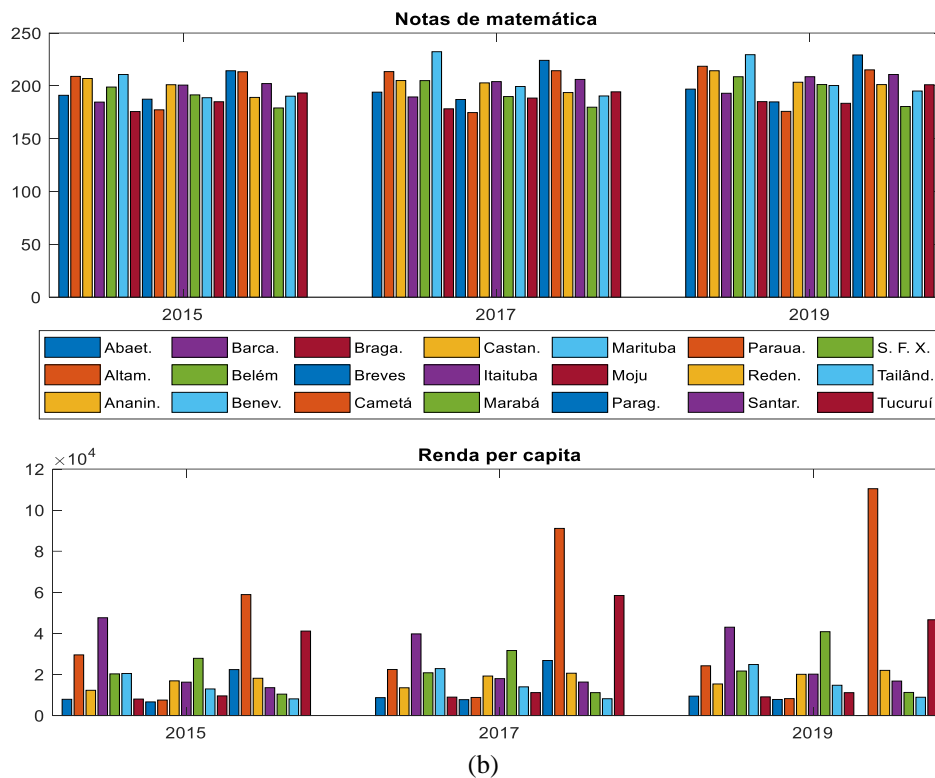
Fonte: Fraiha et. al. 2023.

Na segunda rodada de análise foram considerados os municípios da primeira rodada mais outros 14 municípios de outras regiões do estado. Os resultados para esta rodada de análise são mostrados na Figura 4.12 e na Tabela 4.5.

Figura 4.12. Resultados notas e renda per capita de 21 municípios. (a) Notas de português (b) Notas de matemática.



(a)



Fonte: Fraiha et. al, 2023.

**Tabela 4.5: Dados de correlação e parâmetro  $p$  para 7 municípios**

Correlação Português e RPC			Correlação Matemática e RPC		
2015	2017	2019	2015	2017	2019
0.4098	0.4067	0.3064	0.4024	0.3335	0.2472
Valores $p_k$ português e RPC			Valores $p_k$ Matemática e RPC		
2015	2017	2019	2015	2017	2019
0.065	0.0673	0.1767	0.0705	0.1395	0.2801

Fonte: Fraiha et. al. 2023.

A rodada com 7 municípios (figura 4.11 e tabela 4.4) os resultados sugeriram uma investigação ampliada, apresentando indícios de correlação forte entre as notas e a renda per capita. Neste trabalho, entende-se como correlação forte valores a partir de 0,7 (ou valores próximos a 0,7). Por outro lado, os valores  $p$  associados a estas correlações foram muito baixos. Entretanto, os resultados com 21 municípios (figura 4.12 e tabela 4.5) trouxeram essa hipótese inicial de volta a um patamar básico, mostrando uma diminuição da força das correlações identificadas na primeira rodada corroboradas por valores  $p$  próximos do ideal (0,05). Essa ampliação da amostra visava compreender se os indícios iniciais de correlação encontrados na primeira rodada se mantinham quando considerados contextos mais diversos e variados. Isso

sugere que, quando mais municípios e diferentes regiões foram considerados, a relação entre as notas do Saeb e as rendas per capita se tornou menos consistente (Fraiha et. al, 2023).

Utilizando a análise de correlação linear através do coeficiente de correlação Pearson, foi possível calcular a matriz de correlação 9x9 para as variáveis estudadas. Analisando as matrizes obtidas (correlação e valores  $p$ ) percebeu-se as diferenças que a quantidade de elementos de uma amostra pode levar aos resultados. Nesta etapa do estudo, duas possibilidades se destacam como desdobramentos futuros: (i) Buscar correlação entre outras variáveis socioeconômicas e as notas de português e matemática; (ii) Buscar correlação entre outras variáveis socioeconômicas e outras variáveis que possam mensurar o desempenho dos estudantes, além da nota.

#### **4.5- Considerações Finais**

As aplicações expostas neste capítulo ilustram algumas possibilidades para ferramentas matemáticas acessíveis a alunos de graduação que tenham alguma iniciação à pesquisa em probabilidade, estatística e/ou estudo de séries temporais. Conceitos normalmente abordados em aulas expositivas “tradicionais” são aplicados em um contexto diferente dos que normalmente ocorrem durante as aulas, propiciando ao estudante novos horizontes para sua formação e atuação.

## Capítulo 5 – Conclusões

Este capítulo finaliza a apresentação deste estudo e contém as conclusões acerca dos resultados apresentados em relação às aplicações aqui mostradas. Dois tipos de aplicações geraram três resumos/artigos publicados durante o período de orientação do autor juntamente a seu orientador.

A aplicação da modelagem do tipo ARIMA (e sua variação híbrida ARIMA-RNA) se deu como a iniciação do autor à pesquisa, auxiliando seu orientador em procedimentos como coleta de dados e manipulação básica dos dados brutos, além de proporcionar o contato do autor com outra área de conhecimento: a telecomunicação, em particular com o estudo da propagação de ondas eletromagnéticas no contexto das telecomunicações. Conforme previsto em (Fraiha, 2020), a modelagem híbrida adaptou-se muito bem ao novo cenário, principalmente se considerar que o cenário de Salinópolis tem menos obstáculos à propagação (pelo menos em termos de construções, como edifícios) que Belém, conforme exposto em (Barbosa, 2021). Contudo, justamente por ser um cenário mais simples, perguntas que devem sempre nortear uma pesquisa que envolva modelagem foram respondidas em (Barbosa, 2022), a saber: **Qual é o melhor custo-benefício para a situação? Existe uma modelagem mais simples que traga resultados similares?** Em (Barbosa, 2022), concluiu-se que sim, no cenário de Salinópolis estudado o modelo mais simples trouxe resultados satisfatórios. Isso não significa, contudo, que em outro cenário na mesma cidade contemplando uma região geográfica maior (o que pode implicar em uma amplitude e variabilidade maior dos valores de intensidade de sinal recebidos) o modelo puramente linear represente de forma fidedigna o fenômeno da propagação eletromagnética na faixa dos 800 MHz.

Com relação ao segundo tipo de aplicação apresentado, i.e., ao estudo da correlação entre duas VAs, o problema abordado surgiu de uma pesquisa iniciada por (Perez, 2016) que analisava os níveis de letramento e numeramento no Brasil. Ao longo desta primeira pesquisa, a autora inferiu a eventual existência de correlação forte entre as notas de português e matemática com a RPC de determinados municípios do estado do Pará. Isto posto, a inferência foi abordada no trabalho (Fraiha et. al, 2023) que conta com a colaboração do autor desta monografia. A dúvida inicialmente surgiu ao se analisar 7 municípios da região metropolitana de Belém, mas, ao se estender a análise para 21 municípios (agora escolhidos segundo critérios populacionais), concluiu-se que a correlação (caso exista na sua forma linear) tem alta probabilidade de não ser do tipo forte, segundo os critérios adotados em (Fraiha et. al, 2023).

Em comum, os dois tipos de aplicações apresentadas apresentam perguntas iniciais que, ao serem respondidas, resultam em outras dúvidas que aparentemente são contraditórias às primeiras. Contudo, este tipo de situação é comum em ambientes de pesquisa e contribuem para o uma formação mais refinada a nível de graduação daqueles que se deparam com este tipo de problema, pois é exigida uma revisão, ou redirecionamento dos objetivos da pesquisa inicialmente planejada, ou seja, adiciona um grau de complexidade ao trabalho do aluno pesquisador.

Finalmente, as situações problema abordadas pelo autor enquanto pesquisador ilustram que conteúdos abordados muitas vezes de forma automática e sem explicitar elos com a realidade e problemas reais da sociedade podem ser acessíveis enquanto objetos de pesquisa. De fato, o problema de analisar as correlações é acessível a qualquer aluno que tenha uma boa fundamentação em probabilidade e estatística. Já o problema da modelagem ARIMA-RNA, ainda que contemple assuntos que não costumam ser abordados em disciplinas de graduação do curso de Licenciatura em Matemática, são plenamente acessíveis a qualquer aluno que inicie pesquisa na área de processos estocásticos e séries temporais. Desta forma, entende-se que o autor desta monografia teve a oportunidade de ter uma trajetória acadêmica mais completa em virtude de sua atuação na pesquisa, enquanto aluno de iniciação científica.

## Referências Bibliográficas

- SPIEGEL, M. R., SCHILLER, J., SRINIVASAN, R. A., Schaum's Outlines Probability and Statistics. 3<sup>rd</sup> edition, McGraw-Hill, 2009.
- BOX JENKINS, G. E. P. et al. Time Series Analysis: Forecasting and Control. 4<sup>th</sup> edition, New Jersey, John Wiley & Sons, 2008.
- MORETTIN, P. A., TOLOI, C. M. C., Análise de séries temporais. Editora Blucher, 2008.
- FRAIHA L., R. L. Abordagem Multiobjetivo Outdoor Para Posicionamento de Estações Rádio Base e Metodologia Híbrida Aplicada à Avaliação da Exposição Humana à Radiação Não-Ionizante. Tese de doutorado. UFPA, 2020.
- MATLAB version 8.0.0.783. In Technical Documentation, 2012.
- GOOGLE EARTH PRO, version 7.3.3.778, 2020.
- Network Cell Info Lite, version 5.0.23, 2020.
- HAYKIN S., MOHER M. Modern Wireless Communications. Bookman, 2008.
- INTERNATIONAL TELECOMMUNICATOINS UNION (ITU). Recommendation ITU-R P. 1546-5. Method to point-to-area predictions for terrestrial services in the frequency range of 30 MHz to 3000 MHz. 2013.
- HAYKIN S., Redes Neurais: Princípios e Prática. 2<sup>a</sup> Edição. Bookman, 2000.
- IBGE. Cidades e estados, 2023, Disponível em: <https://www.ibge.gov.br/cidades-e-estados/>. Acessado em 06/02/2023.
- INEP, SAEB. Resultados, 2023. Disponível em: <https://www.gov.br/inep/pt-br/areas-deatuacao/avaliacao-e-exames-educacionais/Saeb/resultados>. Acessado em 06/02/2023.
- PEREZ, S.; OLIVEIRA, C. Z.; ANDRADE, E. F. V.; NOGUEIRA, D. P. S.; CANTÃO, E. A.; ROCHA, K. M. DE M; NUNES, F. S. R; Níveis de Letramento e Numeramento nos Exames em Larga Escala no Brasil. In: CONGRESSO IBERO AMERICANO EM INVESTIGAÇÃO QUALITATIVA, 5., (2016), Porto. Atas eletrônicas do 5<sup>o</sup> Congresso Ibero Americano em Investigação Qualitativa. Porto: Universidade Lusófona do Porto (2016).