



Universidade Federal da Bahia  
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

**DETECÇÃO DE NOVIDADES EM STREAM  
DE DADOS UTILIZANDOS FUNÇÕES DE  
BASE RADIAL E CADEIAS DE MARKOV**

Ruivaldo Azevedo Lobão Neto

QUALIFICAÇÃO DE MESTRADO

Salvador  
03 de Abril de 2019



RUIVALDO AZEVEDO LOBÃO NETO

**DETECÇÃO DE NOVIDADES EM STREAM DE DADOS  
UTILIZANDOS FUNÇÕES DE BASE RADIAL E CADEIAS DE  
MARKOV**

Esta Qualificação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Ricardo Araújo Rios

Salvador  
03 de Abril de 2019



## **RESUMO**

Com a grande quantidade de dados produzidos e coletados diariamente por diferentes sistemas, técnicas de aprendizado de máquina foram propostas com o intuito de auxiliar no processo de extração automática de informações. Dentre essas técnicas, pode-se destacar os algoritmos de agrupamento que buscam encontrar padrões e estruturas implícitas em conjuntos de dados sem qualquer conhecimento fornecido à priori. Neste trabalho de mestrado, busca-se desenvolver uma nova abordagem de agrupamento para dados que possuem uma dependência temporal entre suas observações, comumente chamados de séries temporais. A principal diferença dessa abordagem em relação aos trabalhos existentes na literatura baseia-se na hipótese de que dados coletados do mundo real possuem influências estocásticas e determinísticas que, se não forem individualmente analisadas, podem afetar o resultado do agrupamento. Neste sentido, a abordagem proposta realiza uma etapa de decomposição das séries temporais em componentes estocásticos e determinísticos. Em seguida, realiza o agrupamento dos dados analisando de maneira individual a similaridade entre cada componente. Experimentos preliminares foram realizados sobre séries temporais sintéticas, formadas a partir da combinação entre componentes estocásticos e determinísticos. A partir da decomposição das séries sintéticas, foi possível observar que medidas de distância largamente utilizadas na literatura apresentaram melhores resultados. Ao final deste trabalho de mestrado, espera-se obter um melhor resultado de agrupamento utilizando a abordagem proposta sobre dados reais.

**Palavras-chave:** Agrupamento, Séries Temporais, Decomposição, Estocasticidade, Determinismo.



# SUMÁRIO

<b>Capítulo 1—Introdução</b>	1
1.1 Considerações Iniciais . . . . .	1
1.2 Contextualização e Motivação . . . . .	2
1.3 Hipótese e Objetivo . . . . .	3
<b>Capítulo 2—Revisão Bibliográfica</b>	5
2.1 Considerações Iniciais . . . . .	5
2.2 Séries Temporais . . . . .	6
2.3 Análise de Séries Temporais . . . . .	6
2.4 Agrupamento de Séries Temporais . . . . .	8
2.4.1 Técnicas de Agrupamentos . . . . .	9
2.4.2 Cálculo da similaridade/distância entre séries temporais . . . . .	9
2.5 Trabalhos relacionados . . . . .	12
2.6 Considerações Finais . . . . .	13
<b>Capítulo 3—Plano de Pesquisa</b>	15
3.1 Considerações Iniciais . . . . .	15
3.2 Descrição do problema . . . . .	15
3.2.1 Atividades da pesquisa . . . . .	17
3.3 Considerações Finais . . . . .	18
<b>Capítulo 4—Experimentos Iniciais</b>	19
4.1 Considerações Iniciais . . . . .	19
4.2 Configuração dos experimentos . . . . .	19
4.3 Análise de Similaridade/Distância . . . . .	21
4.4 Considerações Finais . . . . .	26
<b>Apêndice A—Decomposição das séries temporais</b>	33
A.1 Considerações Iniciais . . . . .	33
A.2 Séries TIPO 1 . . . . .	33
A.3 Séries TIPO 2 . . . . .	35
A.4 Séries TIPO 3 . . . . .	38
A.5 Séries TIPO 4 . . . . .	40
A.6 Considerações Finais . . . . .	43



## **LISTA DE FIGURAS**

2.1 Séries temporais: (A) Sazonalidade. (B) Sazonalidade+ Componente aleatório. (C) Sazonalidade + Componente aleatório + Tendência. . . . .	7
3.1 Proposta do projeto de mestrado. . . . .	17
A.1 Série 1.1 e Série 1.2 . . . . .	33
A.2 Série 1.3 e Série 1.4 . . . . .	34
A.3 Série 1.5 e Série 1.6 . . . . .	34
A.4 Série 1.7 e Série 1.8 . . . . .	35
A.5 Série 1.9 e Série 1.10 . . . . .	35
A.6 Série 2.1 e Série 2.2 . . . . .	36
A.7 Série 2.3 e Série 2.4 . . . . .	36
A.8 Série 2.5 e Série 2.6 . . . . .	37
A.9 Série 2.7 e Série 2.8 . . . . .	37
A.10 Série 2.9 e Série 2.10 . . . . .	38
A.11 Série 3.1 e Série 3.2 . . . . .	38
A.12 Série 3.3 e Série 3.4 . . . . .	39
A.13 Série 3.5 e Série 3.6 . . . . .	39
A.14 Série 3.7 e Série 3.8 . . . . .	40
A.15 Série 3.9 e Série 3.10 . . . . .	40
A.16 Série 4.1 e Série 4.2 . . . . .	41
A.17 Série 4.3 e Série 4.4 . . . . .	41
A.18 Série 4.5 e Série 4.6 . . . . .	42
A.19 Série 4.7 e Série 4.8 . . . . .	42
A.20 Série 4.9 e Série 4.10 . . . . .	43



## **LISTA DE TABELAS**

3.1	Cronograma de atividades . . . . .	18
4.1	Grupos de séries temporais sintéticas utilizadas nos experimentos . . . . .	20
4.2	Cálculo da DTW entre séries do Tipo 1 e Tipo 3 (ruído aditivo). . . . .	21
4.3	Cálculo da DTW entre séries do Tipo 2 e Tipo 4 (ruído aditivo e tendência). . . . .	22
4.4	Cálculo da DTW entre séries do Tipo 1 e Tipo 3 com decomposição. . . . .	22
4.5	Cálculo da DTW entre séries do Tipo 2 e Tipo 4 com decomposição. . . . .	23
4.6	Cálculo da medida Euclidiana entre séries do Tipo 1 e Tipo 3 sem decomposição e com decomposição. . . . .	23
4.7	Cálculo da medida Euclidiana entre séries do Tipo 2 e Tipo 4 sem decomposição e com decomposição. . . . .	24
4.8	Cálculo da distância Manhattan entre séries do Tipo 1 e Tipo 3 sem decomposição e com decomposição. . . . .	24
4.9	Cálculo da distância Manhattan entre séries do Tipo 2 e Tipo 4 sem decomposição e com decomposição. . . . .	25
4.10	Cálculo da distância Minkowski entre séries sem decomposição e com decomposição. . . . .	26
4.11	Cálculo da distância CID entre séries sem decomposição e com decomposição. . . . .	27
4.12	Cálculo da distância Cross Correlation entre séries sem decomposição e com decomposição. . . . .	27



# Capítulo

# 1

## INTRODUÇÃO

### 1.1 CONSIDERAÇÕES INICIAIS

Atualmente, grandes volumes de dados são coletados e produzidos por diferentes sistemas. Segundo Nguyen, Woon e Ng (2015), diariamente, mais de 3,5 bilhões de buscas são realizadas nos repositórios do Google e cerca de 4TB de imagens são geradas por satélites da NASA.

Além das grandes corporações, com o surgimento das redes sociais e a popularização de dispositivos de acesso à Internet, usuários comuns passaram a produzir, de maneira efetiva, grandes volumes de dados por meio, por exemplo, da publicação e compartilhamento de fotos, textos e vídeos.

Por fim, é importante destacar que evoluções em áreas estratégicas da computação têm favorecido o crescente aumento no volume de dados produzidos e armazenados. Uma dessas áreas é chamada de Internet das Coisas (*Internet of Things*) (OH; KIM, 2017), a qual visa conectar e coordenar diferentes dispositivos sem a intervenção humana. Recentemente, uma pesquisa divulgada pela IDC (*International Data Corporation*) apresentou uma previsão de que cerca de 32 bilhões de dispositivos estarão interconectados até 2020 (OH; KIM, 2017).

Esse aumento significativo na quantidade de dados produzidos e armazenados tem dificultado a tarefa de especialistas de domínio na análise e extração novas informações. Visando superar essas dificuldades, técnicas de Aprendizado de Máquina (AM) têm sido propostas para desenvolver programas de computadores que sejam capazes de analisar dados e aprender, com base na experiência fornecida por especialistas, com intuito de melhorar o desempenho na realização de alguma tarefa (MITCHELL, 1997; FACELI et al., 2011).

De maneira geral, o aprendizado realizado por técnicas de AM ocorre visando induzir hipóteses que sejam capazes de descrever relações entre os dados analisados. Essa busca por tais hipóteses é determinada pelo viés de cada algoritmo, o qual representa a capacidade de generalização do modelo aprendido quando aplicado à novos dados não vistos previamente (FACELI et al., 2011).

Ao encontrar hipóteses que descrevem o comportamento dos dados, pode-se ajustar um modelo para, por exemplo, predizer o comportamento futuro de um sistema ou, simplesmente, descrever seu estado atual. O ajuste destes modelos ocorre de acordo com o paradigma de aprendizado, o qual pode ser supervisionado ou não-supervisionado. No paradigma supervisionado, modelos são estimados considerando um conjunto de instâncias (dados) que contém características (atributos) específicas de cada instância (dado) e um rótulo (atributo meta), o qual é normalmente fornecido por um especialista. Tarefas de aprendizagem neste caso são, geralmente, realizadas para classificação e regressão de novas instâncias.

Por outro lado, existem situações onde não é possível contar com um especialista para, previamente, fornecer rótulos para cada instância. Neste caso, a utilização de métodos de aprendizagem do paradigma não-supervisionado são normalmente utilizados, uma vez que nenhuma informação é conhecida a priori, exceto os atributos de cada instância.

Dentre as técnicas do paradigma não-supervisionado mais utilizadas na literatura, pode-se destacar os algoritmos de Agrupamento de Dados. Esses algoritmos analisam dados buscando estruturas de tal forma que dados pertencentes a um mesmo grupo sejam de alguma forma mais semelhantes do que dados de outros grupos (MITCHELL, 1997; FACELI et al., 2011; AGHABOZORGI; SHIRKHORSHIDI; WAH, 2015).

Os dados analisados pelos algoritmos de agrupamento podem ser caracterizados de diferentes formas como, por exemplo: textos em redes sociais, cadastro de clientes de uma organização, ou exames realizados em pacientes de um hospital.

Em geral, dados são coletados de maneira independente e identicamente distribuída ( $iid$ ), ou seja, são coletados seguindo alguma distribuição de probabilidade sem, necessariamente, ser caracterizado por uma dependência temporal. Entretanto, quando existe essa dependência, dados são organizados como séries temporais (ESLING; AGON, 2012; BOX et al., 2015). Por exemplo, séries temporais podem ser utilizadas para representar medidas sequenciais ao longo do tempo de médias diárias de temperatura de uma cidade, variações de preço de uma determinada ação na bolsa de valores, propagação de uma doença ou cantos de pássaros (ESLING; AGON, 2012).

Assim, o foco deste trabalho de mestrado é analisar a aplicação de técnicas de agrupamento de dados em séries temporais. Visando apresentar claramente as contribuições deste trabalho, a próxima seção detalha o contexto e a motivação para realização dessa análise.

## 1.2 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Séries temporais são representadas por uma coleção de valores obtidos a partir de medidas sequenciais ao longo do tempo e são utilizadas para análise do comportamento de sistemas em diversas áreas, tais como a medicina, meteorologia, captura de movimento, governo, engenharia, negócios, finanças, economia, entre outras (ESLING; AGON, 2012; BOX et al., 2015).

O agrupamento de séries temporais possibilita extrair informações em conjuntos de dados temporais, agrupando as séries (instâncias) em grupos (*clusters*) de tal forma que a variância *intercluster* seja maximizada e a *intraccluster* seja minimizada. Assim, séries

que compartilham características (informações) similares são organizadas em um mesmo grupo.

Para determinar variâncias entre séries temporais são utilizadas métricas que calculam similaridades (MORI; MENDIBURU; LOZANO, 2016). Na literatura de aprendizado de máquina não-supervisionado, instâncias são agrupadas considerando métricas de similaridades ou métricas de distância. Em algumas situações, a similaridade entre duas instâncias pode ser calculada utilizando o inverso da distância entre elas.

No entanto, a escolha de tais métricas não é trivial, uma vez que séries temporais podem apresentar diferentes comportamentos que influenciam o cálculo da similaridade. Por exemplo, a similaridade entre séries com comportamento determinístico pode ser calculada utilizando técnicas como *Dynamic Time Warping* (DTW) (TORMENE et al., 2009) ou *Mean Distance from the Diagonal Line* (MDDL) (RIOS; MELLO, 2016, 2013). Por outro lado, o cálculo da similaridade entre séries com comportamento estocástico pode ser realizado a partir da análise no domínio de frequência, i.e., comparando espectrogramas obtidos com a transformada de Fourier (MORETTIN; TOLOI, 2006).

Contudo, séries temporais podem apresentar uma mistura de ambos os comportamentos. Conforme discutido em Rios e Mello (2013, 2016), considerar apenas um dos comportamentos pode reduzir a acurácia no processo de modelagem e análise de séries temporais.

Visando resolver essa limitação, este projeto de mestrado discute uma abordagem que permite analisar individualmente os comportamento estocásticos e determinísticos no processo de cálculo de similaridade entre séries temporais e, consequentemente, melhorar o processo de agrupamento de dados.

### 1.3 HIPÓTESE E OBJETIVO

Com base nas observações citadas anteriormente, a seguinte hipótese foi formulada:

*“O agrupamento de séries temporais apresenta maior acurácia quando medidas de similaridade (ou distância) são, individualmente, calculadas sobre comportamentos estocásticos e determinísticos.”*

Assim, o objetivo deste trabalho de mestrado será validar esta hipótese. Para alcançar esse objetivo, será desenvolvida uma abordagem de agrupamento que decompõe séries temporais em dois componentes, um estocástico e um determinístico. Inicialmente, a decomposição das séries será realizada considerando as abordagens propostas em Rios e Mello (2013, 2016). A partir dessa decomposição, será possível, individualmente, medir a similaridade/distância entre cada componente. Para isso, diferentes métricas de similaridade e distância serão implementadas para validação da abordagem proposta. As séries utilizadas nos experimentos serão divididas em dois conjuntos. Um conjunto formado por séries sintéticas, que permitirá realizar uma análise detalhada da abordagem, uma vez que os comportamentos estocásticos e determinísticos serão previamente conhecidos. O outro conjunto será composto por séries coletadas de algum sistema do mundo real, visando apresentar uma aplicação prática para a solução proposta neste trabalho.

O restante deste projeto está organizado da seguinte maneira: O **Capítulo 2** possui uma revisão bibliográfica dos principais conceitos utilizados neste trabalho como, por exemplo, análise, decomposição e agrupamento de séries temporais; No **Capítulo 3** está o plano de pesquisa elaborado com o objetivo de validar a hipótese desta pesquisa, a metodologia que será utilizada na pesquisa e o cronograma de atividades; finalmente, no **Capítulo 4**, é apresentado um conjunto de experimentos preliminares, os quais foram desenvolvidos para demonstrar a importância de analisar os comportamentos estocásticos e determinísticos no agrupamento de séries temporais.

## Capítulo

# 2

## REVISÃO BIBLIOGRÁFICA

### 2.1 CONSIDERAÇÕES INICIAIS

Em Aprendizado de Máquina não-supervisionado, técnicas de agrupamento visam encontrar padrões e extrair estruturas sobre conjuntos de dados sem qualquer informação à priori, utilizando apenas os valores dos atributos de cada instância. A etapa mais importante dos algoritmos de agrupamento é a utilização de medidas que permitem calcular a similaridade ou a distância entre os dados. De maneira geral, esses algoritmos buscam manter em um mesmo grupo dados mais similares. Quando dados analisados são organizados como Séries Temporais, a complexidade no cálculo da similaridade/distância aumenta devido à presença de uma dependência temporal entre as observações (AGHA-BOZORGI; SHIRKHORSHIDI; WAH, 2015; ZHANG et al., 2011).

A dependência temporal entre observações de séries temporais é determinada pela presença de diferentes componentes como, por exemplo, a estocasticidade e o determinismo. Conforme discutido em Rios e Mello (2013, 2016), analisar séries temporais utilizando um único modelo pode produzir erros, uma vez que o componente estocástico é desconsiderado por técnicas de modelagem determinística e modelos estocásticos tendem a não analisar importantes informações no espaço topológico como atratores e repulsores. Este projeto de mestrado estende a abordagem apresentada em (RIOS; MELLO, 2013, 2016) para permitir que o agrupamento de dados seja realizado considerando, de maneira individual, as influências estocásticas e determinísticas das séries.

Este capítulo apresenta uma discussão geral sobre os principais temas abordados neste projeto. Inicialmente, serão apresentados conceitos básicos sobre análise de séries temporais. Em seguida, a técnica de decomposição de série temporal em componentes estocásticos e determinísticos, que será utilizada neste projeto, é discutida em detalhes. Por fim, conceitos de agrupamento de séries temporais são apresentados, fornecendo detalhes sobre as principais medidas de similaridade/determinismo utilizadas na literatura.

## 2.2 SÉRIES TEMPORAIS

Uma série temporal pode ser definida como uma sequência de observações coletados ao longo do tempo  $X_t = x_0, x_1, x_2, \dots, x_t$  (BOX et al., 2015; CHOUAKRIA; NAGABHUSHAN, 2007; MORETTIN; TOLOI, 2006). Uma série temporal univariada é composta por valores escalares, enquanto que as multivariadas possuem múltiplas dimensões dentro da mesma faixa de tempo (BOX et al., 2015; CHOUAKRIA; NAGABHUSHAN, 2007). Séries temporais podem ser caracterizadas, ainda, pelo intervalo entre coleta de observações que pode ser discreta ou contínua. Quando este intervalo é definido em tempos fixos, tem-se uma série discreta. Caso contrário, a série é dita ser contínua (BOX et al., 2015; MORETTIN; TOLOI, 2006).

Além disso, é importante enfatizar que séries temporais podem ser classificadas de acordo com a linearidade, estacionariedade e a estocasticidade de suas observações. Uma série é dita linear quando os valores de suas observações são determinados por uma combinação linear de ocorrências e ruídos passados. Quando a série é formada por sistemas cuja regra geradora são compostas por funções não-lineares, a série é dita ser não-linear (BOX et al., 2015).

As séries estacionárias assumem que as observações estão em equilíbrio estatístico com propriedades que não mudam ao longo do tempo. Em séries estacionárias, sua média e variância são constantes. Quando essas propriedades mudam ao longo do tempo, a série é dita ser não-estacionária e, possivelmente, apresenta um comportamento de tendência (BOX et al., 2015; CRYER; CHAN, 2008).

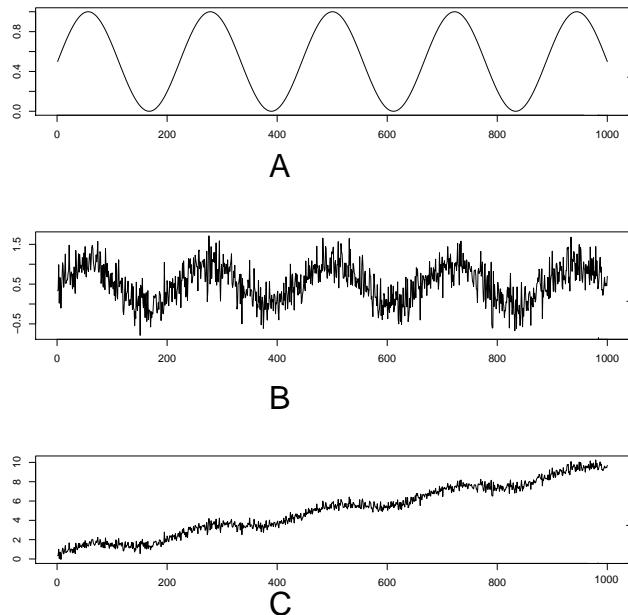
Por fim, séries podem ser classificadas em estocástica ou determinística. Quando apresentam um comportamento determinístico, o valor de suas observações possuem uma estrita relação de dependência com observações passadas. Séries estocásticas são constituídas por observações e relações aleatórias que seguem funções de densidade de probabilidade e podem se modificar ao longo do tempo, dificultando a modelagem de seus eventos (BOX et al., 2015; MORETTIN; TOLOI, 2006).

Segundo Box et al. (2015), observações de séries temporais podem ser escritas na forma  $x_t = T_t + S_t + \varepsilon_t$ , onde  $T_t$  representa a tendência,  $S_t$  representa a sazonalidade e  $\varepsilon_t$  representa os componentes aleatórios. Se o componente  $T$  está presente, a série é dita não-estacionária. Como pode ser visto nesta formulação, a sazonalidade é referenciada como sendo um componente determinístico e  $\varepsilon$  como sendo um componente estocástico. Na Figura 2.1 são mostradas três séries influenciadas por tais componentes.

Neste projeto de mestrado, serão consideradas séries univariadas, cujas observações são definidas por um ruído aditivo gerado a partir da combinação entre componentes estocásticos ( $\varepsilon$ ) e determinísticos ( $S$  e  $T$ ).

## 2.3 ANÁLISE DE SÉRIES TEMPORAIS

O processo de análise de séries temporais tem como objetivo estimar uma regra (ou função geradora) para modelar o comportamento de suas observações, visando estudar e/ou predizer o comportamento de sistemas reais (BOX et al., 2015; SHUMWAY; STOFFER, 2006).



**Figura 2.1** Séries temporais: (A) Sazonalidade. (B) Sazonalidade+ Componente aleatório. (C) Sazonalidade + Componente aleatório + Tendência.

No entanto, como já citado anteriormente, os sistemas do mundo real produzem séries com uma mistura de comportamentos estocásticos e determinísticos (HAN; LIU, 2009). Nesse caso, é importante utilizar modelos que analisem de forma individual a influência de cada comportamento.

Para alcançar esse objetivo, Rios e Mello (2013, 2016) desenvolveram uma técnica que permite decompor séries temporais em componentes estocásticos e determinísticos.

Essa técnica foi baseada no método de Decomposição de Modo Empírico (EMD) (*Empirical Mode Decomposition*) (HUANG et al., 1998). De maneira geral, EMD decompõe uma série temporal em um conjunto de monocomponentes chamados IMF's (*Intrinsic Mode Function*). Considerando que os IMF's revelam diferentes informações implícitas presente nos dados, Rios e Mello (2013, 2016) utilizaram diferentes abordagens para medir o nível de determinismo presente em cada IMF e, consequentemente, combinar IMFs gerando os componentes estocásticos e determinísticos.

No contexto deste trabalho de mestrado, essa técnica de decomposição permitirá separar séries temporais de acordo com suas influências estocásticas e determinísticas. Em seguida, essas influências serão individualmente analisadas pelos algoritmos de agrupamento, visando partitionar o conjunto de séries temporais com maior acurácia. Na próxima seção, serão discutidos os principais conceitos de agrupamento de séries temporais.

## 2.4 AGRUPAMENTO DE SÉRIES TEMPORAIS

O agrupamento de séries temporais busca encontrar padrões em dados, cujas observações são caracterizadas por uma dependência temporal, *i.e.*, a ordem de coleta das observações é importante para a análise. O agrupamento de séries temporais é um tema relevante e tem sido utilizado em diversas pesquisas tais como, Medicina (BLEIWEISS, 2015), Aviação (AYHAN; SAMET, 2016), Genética (IZAKIAN; PEDRYCZ; JAMAL, 2015), Financeira (DURANTE; PAPPADÀ; TORELLI, 2014) e detecção de anomalias (ESLING; AGON, 2012; AGHABOZORGI; SHIRKHORSHIDI; WAH, 2015).

Segundo Aghabozorgi, Shirkhorshidi e Wah (2015), o agrupamento de séries temporais pode ser dividido em três categorias: i) agrupamento sobre diferentes séries temporais; ii) agrupamento de janelas de observações visando encontrar padrões de comportamento em uma mesma série temporal; e iii) agrupamento de ponto temporal que visa encontrar, em uma mesma série temporal, observações semelhantes. As principal diferença entre a categoria (ii) e (iii) é que a última não precisa definir uma janela fixa de observações.

De acordo Bagnall e Janacek (2005), o agrupamento de séries temporais é realizado considerando 3 diferentes objetivos: i) tempo; ii) formato; e iii) mudança. Com relação ao tempo, algoritmos de agrupamento analisam séries calculando a similaridade entre observações coletadas em um mesmo instante de tempo. Com relação à forma, algoritmos visam agrupar séries temporais com comportamento geral semelhante, ou seja, com padrões recorrentes, visando encontrar um melhor alinhamento nos dados. Por fim, existem algoritmos cujo objetivo é avaliar a similaridade entre séries extrairindo novas características como, por exemplo, a presença de tendência ou componentes implícitos. Neste caso, os algoritmos de mudança são geralmente executados sobre as características extraídas e não diretamente sobre as séries.

Ainda segundo Liao (2005), séries temporais podem ser agrupadas considerando três abordagens: i) aplicação de agrupamento sobre dados brutos, ou seja, sem qualquer transformação nas observações coletadas; ii) extração de características e aplicação de algoritmos de agrupamento convencionais sobre as características; e iii) modelagem de cada série e aplicação do agrupamento sobre cada modelo obtido.

Neste projeto de mestrado, será estudada a categoria de agrupamento aplicada sobre bases de dados compostas por diferentes séries temporais. Com relação ao objetivo, este projeto visa permitir que agrupamentos no tempo, no formato e na mudança sejam realizados com maior acurácia ao analisar, individualmente, os componentes estocásticos e determinísticos de cada série.

Por fim, espera-se que essa análise individual dos componentes permita melhorar o resultado de métodos de agrupamento que executam diretamente sobre dados ou sobre características e modelos obtidos a partir das observações.

Na próxima seção, alguns algoritmos de agrupamento de séries temporais são descritos com maior acurácia.

### 2.4.1 Técnicas de Agrupamentos

Algoritmos de agrupamento visam identificar, sem informação de especialistas, a forma como os dados serão estruturados (FACELI et al., 2011). Com base nesta definição, as técnicas de agrupamento podem ser organizadas em cinco categorias: i) algoritmos baseados em particionamento; ii) algoritmos hierárquicos; iii) algoritmos baseados em densidade; iv) algoritmos baseados em grade; e v) algoritmos baseados em modelos (NGUYEN; WOON; NG, 2015; LIAO, 2005).

Algoritmos baseados em particionamento visam organizar os dados em um número pré-definido de grupos (NGUYEN; WOON; NG, 2015). Por outro lado, o agrupamento hierárquico visa organizar, de maneira aglomerativa ou divisiva, os dados em uma árvore hierárquica de grupos. Métodos aglomerativos, inicia sua execução considerando cada objeto como um grupo. Nas etapas seguintes, grupos são concatenados de acordo com a similaridade entre seus objetos, até que no último passo (nó raiz da árvore) todos os dados estão em um único grupo. Nos métodos divisivos, todos os dados são, inicialmente, organizados em um único grupo (nó raiz). Em seguida, grupo é dividido em dois outros grupos de acordo com a distância entre as observações do grupo original. Esse passo é repetido até que cada dado esteja presente em um único grupo nas folhas da árvore (AGHABOZORGI; SHIRKHORSHIDI; WAH, 2015; LIAO, 2005; NGUYEN; WOON; NG, 2015).

Em algoritmos baseados em densidade, grupos são formados considerados regiões densas entre dados do mesmo grupo e regiões de menor densidade destacando a separação entre dados de grupos diferentes (NGUYEN; WOON; NG, 2015). Algoritmos baseados em grade, inicialmente, organizam os dados em células individuais de uma grade. No próximos passos, células são agrupadas em subespaços, contendo informações resumidas de seus objetos. Esse passo pode ser repetido até que dos os dados sejam representados por um resumo geral em uma única célula (NGUYEN; WOON; NG, 2015). Algoritmos baseados em modelo tentam otimizar a probabilidade entre dados, utilizando modelos estatísticos. Em geral, esses algoritmos assumem um modelo para cada grupo. Em seguida, localizam o melhor ajuste de dados cada modelo (NGUYEN; WOON; NG, 2015; AGHABOZORGI; SHIRKHORSHIDI; WAH, 2015).

As técnicas descritas nesta seção foram desenvolvidas para dados gerais. A aplicação dessas técnicas em séries temporais depende do uso de medidas de similaridade e distância específicas para dados caracterizados por uma dependência temporal entre suas observações. Neste sentido, a próxima seção apresenta um conjunto de medidas desenvolvidas para calcular a similaridade e a distância entre séries temporais.

### 2.4.2 Cálculo da similaridade/distância entre séries temporais

Visando encontrar medidas de similaridade e distância que podem ser utilizadas para agrupar séries temporais, foi realizada uma revisão sistemática da literatura em parceria com um aluno do curso de Bacharelado em Ciência da Computação da Universidade Federal da Bahia (MOREIRA JUNIOR, 2016).

O resultado desta revisão evidenciou que os principais artigos relacionados utilizam as seguintes medidas: Distância Euclidiana (DE), *Dynamic Time Warping* (DTW) (ZHANG

et al., 2011), LB HUST (JUNKUI; YUANZHEN; XINPING, 2006), *Cross-Correlation* (HÖPPNER; KLAWONN, 2009) e *Complexity-Invariant Distance* (CID) (BATISTA et al., 2014).

A Distância Euclidiana (DE) é uma das medidas mais utilizadas na literatura de agrupamento de dados devido à sua baixa complexidade temporal e espacial ( $O(n)$ ). Esta medida pode ser definida como sendo uma especialização da Distância de Minkowski, conforme a Equação 2.1, quando o valor da variável  $p$  é igual a 2. De maneira geral, esta distância compara pares de observações entre duas séries temporais  $X$  e  $Y$  compostas por  $n$  observações, tal que  $x_i \in X$  e  $y_i \in Y$ .

$$D_{Minkowski}(X, Y) = \sqrt[p]{\sum_{i=1}^n (x_i - y_i)^p} \quad (2.1)$$

Em alguns experimentos publicados na literatura, observou-se que  $p$  assumia os valores 1 (Distância de Manhattan) e 3. Por esta razão, nos experimentos iniciais realizados neste trabalho, optou-se também por utilizar a distância de Minkowski com  $p = \{1, 2, 3\}$ .

Conforme discutido na literatura, calcular a distância entre séries analisando pares de observações, como a distância de Minkowski, pode ser uma desvantagem, principalmente quando não há um alinhamento perfeito entre as séries. Visando solucionar este problema, a DTW foi desenvolvida visando encontrar um alinhamento ótimo entre duas séries antes de calcular a distância entre suas observações (MEESRIKAMOLKUL; NIENNATTRAKUL; RATANAMAHATANA, 2012; CHEN; NG, 2004; ESLING; AGON, 2012; LIAO, 2005; MORI; MENDIBURU; LOZANO, 2016). A DTW entre duas séries temporais  $X$  e  $Y$  pode ser calculada pelas seguintes Equações :

$$DTW(X, Y) = \sqrt{dist(x_n - y_m)} \quad (2.2)$$

$$dist(x_i, y_j) = (x_i - y_j)^2 + \min \begin{cases} dist(x_{i-1}, y_j) \\ dist(x_i, y_{j-1}) \\ dist(x_{i-1}, y_{j-1}) \end{cases} \quad (2.3)$$

É importante destacar que o alinhamento das séries é realizada pela DTW, respeitando as seguintes condições: i) os primeiros e os últimos elementos das séries analisadas são alinhados; ii) as observações de uma mesma série devem manter a ordem da série original; e iii) definição do limite do passo para evitar replicações no alinhamento. A complexidade temporal da DTW é  $O(mn)$ , podendo restringir sua aplicação quando as séries possuem grandes volumes de observações.

Visando reduzir a complexidade temporal da DTW, Keogh (2002) propôs a medida LB Keogh, a qual adiciona uma restrição temporal que limita o número de etapas verticais e horizontais no alinhamento das séries (ESLING; AGON, 2012). Essa limitação faz com que a DTW execute com uma complexidade temporal  $O(n)$ . Contudo, LB Keogh é uma medida assimétrica, impedindo sua aplicação direta em agrupamento de séries temporais. Essa limitação foi superada pela a distância LB HUST, que define limites superiores ( $U_X$ ) e inferiores ( $L_X$ ) formados por uma janela deslizante de tamanho  $2w + 1$ :

i)  $U_{X_i} = \max(X_{i-w}, \dots, X_{i+w})$ ,  $1 \leq i \leq n$ ; e ii)  $L_{X_i} = \min(X_{i-w}, \dots, X_{i+w})$ ,  $1 \leq i \leq n$ . Sendo que  $U_{X_i} \geq X_i \geq L_{X_i}$ , com  $1 \leq i \leq n$ . Assim, a métrica para LB HUST é definida pela Equação 2.4.

$$LBHUST(X, Y) = \sqrt{\sum_{i=1}^n \begin{cases} (L_{X_i} - U_{Y_i})^2, & L_{X_i} < U_{Y_i} \\ (L_{Y_i} - U_{X_i})^2, & L_{Y_i} < U_{X_i} \\ 0, & \text{caso contrário} \end{cases}} \quad (2.4)$$

Outra medida muito utilizada para verificar a similaridade entre séries temporais é a *Cross-Correlation*, a qual permite analisar duas séries mesmo quando suas observações não estão alinhadas entre si (LIAO, 2005; HÖPPNER; KLAWONN, 2009). Ao analisar duas séries  $X$  e  $Y$  de tamanho  $n$ , esta medida calcula a correlação de uma janela  $l$  (*lag*) da série  $Y$  deslocada sobre a série  $X$  (LIAO, 2005; HÖPPNER; KLAWONN, 2009). De maneira geral, essa medida é calculada realizando um deslocamento da esquerda para a direita, até atingir um atraso máximo definido na sua execução. A distância *Cross-Correlation* possui complexidade temporal de  $O(nl)$  e é definida pela Equação 2.5. Nesta equação,  $CC(X, Y, k)$  calcula a correlação cruzada entre as séries  $X$  e  $Y$  considerando um lag  $k$ .

$$DCC(X, Y) = \sqrt{\frac{1 - CC(X, Y, 0)^2}{\sum_{k=1}^l 1 - CC(X, Y, k)^2}} \quad (2.5)$$

A última medida utilizada para analisar séries temporais estudada neste trabalho é a *Complexity-Invariant Distance* (CID). Esta medida foi inicialmente proposta para classificar dados de acordo com seus formatos. De maneira geral, essa medida calcula a distância entre duas séries de mesmo tamanho visando analisar informações sobre diferenças de complexidade entre suas observações. Neste caso, a complexidade de uma série temporal está relacionada ao formato do comportamento geral das suas observações. Para isso, CID permite comparar séries temporais, lidando com uma invariância na complexidade por meio de um fator de correção para as métricas existentes. A CID entre duas séries é obtida com a Equação 2.6, onde  $CF$  é o fator de correção de complexidade definido na Equação 2.7 e  $D(X, Y)$  é a distância entre as séries  $X$  e  $Y$  dada por alguma métrica de similaridade. É importante destacar que a complexidade temporal dessa medida é  $O(n)$  (BATISTA et al., 2014).

$$CID(X, Y) = D(X, Y) * CF(X, Y) \quad (2.6)$$

$$CF(X, Y) = \frac{\max(CE(X), CE(Y))}{\min(CE(X)), \min(CE(Y))} \quad (2.7)$$

Sendo que  $CE(X)$  é a complexidade da estimada série temporal  $X$  definida pela equação a seguir:

$$CE(X) == \sqrt{\sum_{i=1}^{n-1} (x_i - x_{i+1})^2}, \quad (2.8)$$

## 2.5 TRABALHOS RELACIONADOS

Além das referências básicas apresentadas nesta seção, foi realizada uma busca na literatura visando identificar trabalhos relacionados que propõem o agrupamento de séries considerando as influências dos componentes estocásticos e determinísticos.

No primeiro trabalho encontrado, Mori, Mendiburu e Lozano (2016) apresentaram um mecanismo para selecionar métricas de similaridades para agrupamento de séries temporais. Contudo, o componente estocástico foi tratado como sendo um ruído que deveria ser descartado da série.

De maneira semelhante, Bleiweiss (2015) realizou um agrupamento em séries temporais geradas a partir de um sinal de eletrocardiograma (ECG). Neste agrupamento, o autor identifica componente estocástico no intervalo entre batidas do coração. Este componente é removido pela projeção do sinal de ECG em um espaço de alta dimensionalidade que, em uma etapa posterior, é reduzido para um número menor de dimensões. Nesta redução, o componente estocástico é descartado.

Na Gestão do Tráfego Aéreo, é importante realizar uma previsão de trajetória confiável, uma vez que esta previsão pode melhorar não apenas a segurança, mas também garantir uma economia (AYHAN; SAMET, 2016). Considerando que as observações meteorológicas possuem incertezas, Ayhan e Samet (2016) adotou uma abordagem estocástica para trabalhar tais questões propondo um novo algoritmo de agrupamento de séries temporais. O algoritmo proposto permite gerar uma sequência ótima de observações auxiliando a predição de trajetória de voos. De maneira geral, o algoritmo busca encontrar um centroide ideal para séries com componentes estocásticos, sendo que, neste caso, o componente determinístico é descartado na análise.

Além destes trabalhos, foi encontrada também uma pesquisa que busca estudar a similaridade entre séries temporais não-estacionárias. O trabalho foi motivado pelo fato de que distâncias comumente utilizadas, como a euclidiana, podem sofrer interferência por fatores estocásticos (FEI et al., 2009). Para os autores, as séries analisadas são compostas por um componente estocástico e uma tendência. A tendência, então, é removida da série e o valor da parte estocástica é considerado como uma variável aleatória de alguma distribuição Fei et al. (2009). Fei et al. (2009) propõem utilizar parâmetros como coeficientes de equação de regressão para a tendência e de variância temporal autorregressivos (TVPAR) para componentes estocásticos. No trabalho, são definidos dois graus de similaridade baseados na tendência e no componente estocástico para calcular a proximidade entre as séries.

Esse último trabalho citado foi o mais similar com a proposta desta pesquisa. No entanto, a avaliação do trabalho é realizado com algoritmos de aprendizado supervisionado (classificação). Além disso, é importante salientar que o trabalho considera como componente determinístico apenas a tendência.

## **2.6 CONSIDERAÇÕES FINAIS**

Neste capítulo foram apresentados alguns conceitos básicos que serão utilizados como base para a execução deste trabalho. Foram descritos os conceitos de séries temporais e os componentes que descrevem o seu comportamento, conceitos de agrupamentos de séries temporais e suas principais medidas e comportamento diante dos ruídos. Por fim, foram discutidos trabalhos que consideram os componentes ruidosos das séries temporais e como buscam soluções para realizar o processo de aprendizado de máquina considerando a influência de tais componentes.



## PLANO DE PESQUISA

### 3.1 CONSIDERAÇÕES INICIAIS

Este capítulo descreve como a pesquisa proposta neste mestrado será desenvolvida para permitir que medidas de similaridade (distância) sejam aplicadas para analisar séries temporais com ruído aditivo. Espera-se que com a análise individual dos componentes estocásticos e determinísticos, o agrupamento de séries temporais seja realizado com maior acurácia. A seguir, são apresentados detalhes sobre cada etapa do desenvolvimento do projeto.

### 3.2 DESCRIÇÃO DO PROBLEMA

Seja  $X(t) = \{x_1, x_2, \dots, x_t\}$  uma série temporal univariada composta por  $t$  observações. Suponha que  $X(t)$  seja formada por um ruído aditivo onde cada observação de  $X(t)$  é definida por  $x_i = D_i + E_i + T_i$ , tal que  $1 \leq i \leq t$  e  $D_i$ ,  $E_i$  e  $T_i$  representam influências dos componentes determinístico, estocástico e tendência, respectivamente.

Considerando que as influências destes componentes podem ser decompostas utilizando abordagens disponíveis na literatura (RIOS; MELLO, 2013, 2016), este projeto de mestrado visa comprovar a hipótese que conjuntos de dados formados por séries temporais podem ser agrupados com maior acurácia quando medidas de similaridades (ou distância) são, individualmente, aplicadas sobre seus componentes estocásticos e determinísticos.

Neste contexto, um conjunto de dados temporais pode ser representado por  $Y = \{X_\alpha(\beta)\}$ ,  $\alpha = \{1, \dots, n\}$ , tal que  $n \in \mathbb{N}$  representa o número de séries temporais existentes na base de dados e  $\beta \in \mathbb{N}$  representa o número de observações em cada série temporal. É importante destacar ainda que  $n \geq 2$ , uma vez que não faz sentido encontrar estruturas em bases que contém apenas uma observação. Além disso, o tamanho das observações ( $\beta$ ) precisa ser suficiente para execução dos métodos de decomposição. Em análises empíricas, notou-se que é possível aplicar o método EMD em séries com ruído aditivo que contêm ao menos 10 observações.

Para exemplificar a execução desta proposta de mestrado, considere um agrupamento particional do tipo *hard* que visa dividir  $Y$  em  $k$  grupos,  $C = \{C_1, C_2, \dots, C_k\}$ , sendo

$k \leq n$ . Os grupos obtidos devem respeitar as seguintes restrições: i)  $C_i \neq \emptyset, 1 \leq i \leq k$ ; ii)  $\bigcup_{i=1}^k C_i = Y$ ; e iii)  $C_i \cap C_j = \emptyset, i, j = \{1, \dots, k\}$  e  $i \neq j$ .

Para determinar se duas séries devem pertencer a um mesmo grupo, utiliza-se medidas de similaridade ou distância ( $\mathbb{D}$ ). Neste caso, para um conjunto de dados com  $n$  séries temporais, uma matriz de similaridade  $\mathbb{M}$  ( $n \times n$ ) pode ser construída calculando  $\mathbb{M}_{p,q} = \mathbb{D}_{p,q}$ .

Para exemplificar, considere duas séries temporais de mesmo tamanho  $X_p(t)$  e  $X_q(t)$ . A distância entre essas séries pode ser calculada pela distância de Manhattan,  $\mathbb{D}_{p,q} = \sum_{j=1}^t |X_p(j) - X_q(j)|$ . Neste exemplo, considere que as tendências das séries foram removidas<sup>1</sup>, sendo que suas observações são compostas apenas pelas influências estocásticas e determinísticas, i.e.,  $\mathbb{D}_{p,q} = \sum_{j=1}^t |(D_{p,j} + E_{p,j}) - (D_{q,j} + E_{q,j})|$ . Logo, podemos reescrever o cálculo da distância pela Equação 3.1.

$$\begin{aligned}\mathbb{D}_{p,q} &= \sum_{j=1}^t |(D_{p,j} + E_{p,j}) - (D_{q,j} + E_{q,j})| \\ &= \sum_{j=1}^t |(D_{p,j} - D_{q,j}) + (E_{p,j} - E_{q,j})| \\ &= \sum_{j=1}^t |D_{p,j} - D_{q,j}| + \sum_{j=1}^t |E_{p,j} - E_{q,j}|.\end{aligned}\tag{3.1}$$

Portanto, pode-se afirmar que a distância<sup>2</sup> entre duas séries com ruído aditivo pode ser calculada considerando individualmente as distâncias entre seus componentes estocásticos e determinísticos.

Assim, aplicando a decomposição em série temporais, espera-se construir uma matriz de similaridade  $\mathbb{M}$  que pode ser utilizada por algoritmos de agrupamento, encontrando partições  $C$  com maior acurácia.

A Figura 3.1 apresenta de maneira resumida a execução deste projeto. Para o cálculo da distância entre duas séries temporais  $X_p(t)$  e  $X_q(t)$ , é realizada uma decomposição que extrai seus componentes estocásticos ( $E_p$  e  $E_q$ , respectivamente) e determinísticos ( $D_p$  e  $D_q$ , respectivamente).

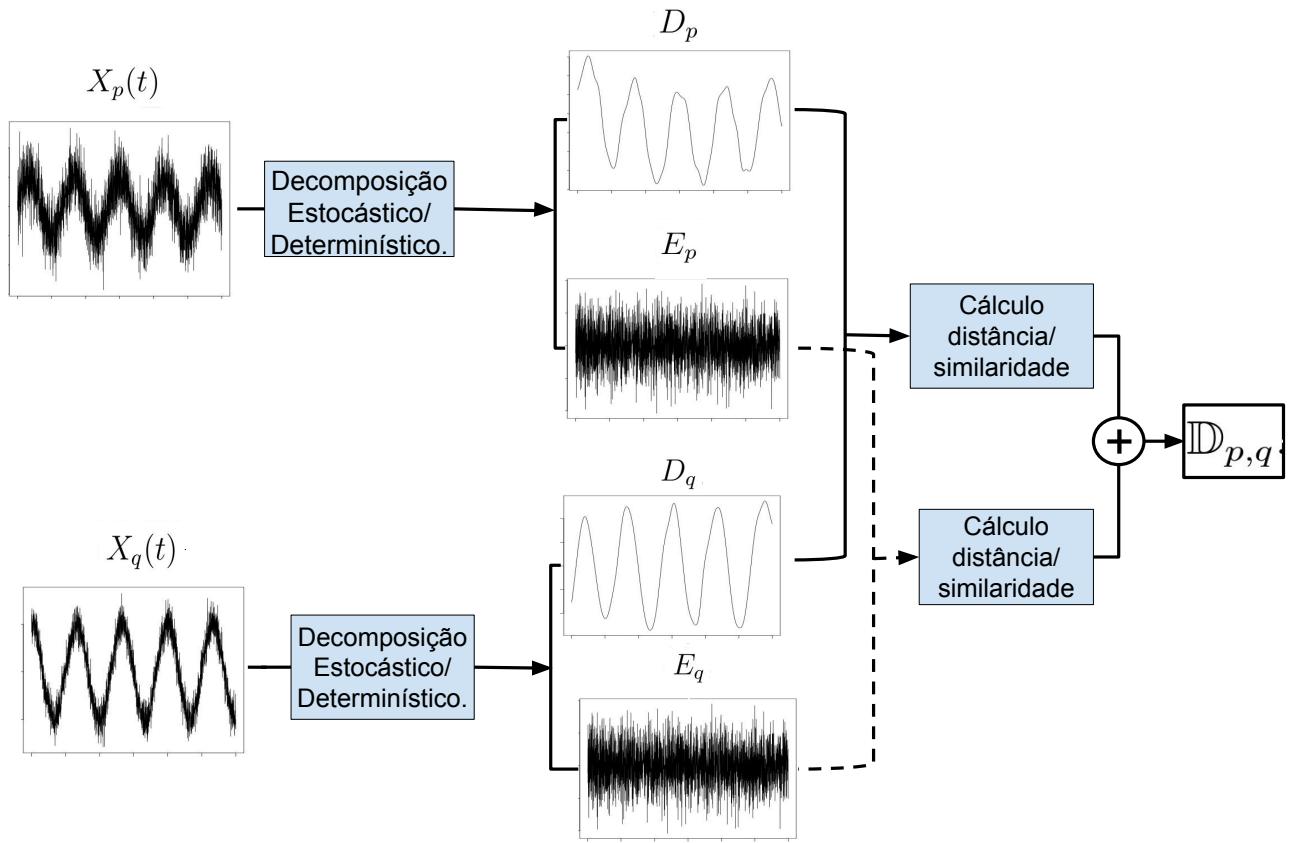
Em seguida, aplica-se uma medida de distância de maneira isolada entre os componentes, ou seja, calcula-se a distância entre os componentes estocásticos da série e a distância entre os componentes determinísticos, reduzindo o impacto da presença de ruído nas medidas utilizadas. Por fim, a distância entre os componentes é combinada para formar a distância geral entre as séries originais  $\mathbb{D}_{p,q}$ .

Em contraste com trabalhos encontrados na literatura, ao realizar uma decomposição de séries temporais, torna-se viável combinar medidas de distância que fornecem bons

---

<sup>1</sup>Em geral, a tendência pode ser considerada como um comportamento determinístico, uma vez que sua influência em um dado instante depende apenas das tendências de instantes anteriores.

<sup>2</sup>Em agrupamento de dados, pode-se utilizar tanto medidas de distância como medidas de similaridade. Dependendo do algoritmo utilizado, a distância pode ser obtida pelo inverso da similaridade.



**Figura 3.1** Proposta do projeto de mestrado.

resultados para séries estocásticas com medidas usualmente aplicadas sobre séries determinísticas.

### 3.2.1 Atividades da pesquisa

A Tabela 3.1 apresenta o cronograma das atividades planejadas para a realização da pesquisa. Atividades concluídas são representadas pelo símbolo  $X$  e as futuras por  $\bullet$ .

Para a conclusão da atividade 1, o Programa de Pós-Graduação em Ciência da Computação (PGCOMP) da Universidade Federal da Bahia (UFBA) exige que um mestrado obtenha um total de 18 créditos em disciplinas. Nos semestres 2016.1 e 2016.2, foram obtidos 15 créditos ao cursar as disciplinas: MATE64 – Seminários Científicos, MATE65 – Fundamentos de Pesquisa em Ciência da Computação I, MATD74 – Algoritmos e Grafos, MATD33 – Tópicos em Inteligência Computacional III e MATE70 – Computação Ubíqua e Sensível ao Contexto. No semestre corrente, 2017.1, está sendo cursada a disciplina MATE32 – Tópicos em Inteligência Computacional II, completando os 3 créditos restantes.

A segunda atividade planejada neste cronograma foi realizada em parceria com o

**Tabela 3.1** Cronograma de atividades

Atividades	Meses																									
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24		
1-Disciplinas	X	X	X	X	X	X	X	X	X	X	X	X	•	•												
2-Revisão da Literatura	X	X	X	X	X	X	X																			
3-Experimentos								X	X	X			•	•	•	•										
4-Análise dos Resultados											X			•	•	•	•									
5-Escrita da qualificação								X	X	X	X															
6-Estágio docente											X	X	•	•												
7-Pesquisa Orientada											X	X	•	•	•	•	•	•	•	•	•	•	•	•	•	
8-Apresentação da qualificação													•													
9-Escrita de artigos																									•	
10-Escrita da dissertação								X	X	X	X			•	•	•	•	•	•	•	•	•	•	•	•	
11- Defesa da dissertação																									•	

aluno de bacharelado em Ciência da Computação Evaldo Machado Moreira Junior. O objetivo da sua monografia foi entender como as métricas são influenciadas pela presença de ruídos. Neste trabalho, as principais medidas/métricas de similaridade/distância entre séries temporais utilizadas por pesquisas na área foram encontradas com a execução de uma Revisão Sistemática de Literatura (*Systematic Literature Review – SLR*).

As atividades 3 e 4 do cronograma consistem na realização dos experimentos e análise dos resultados. Essas atividades foram divididas em duas partes. A primeira contém apenas experimentos preliminares que foram realizados para validar esta proposta de trabalho. Nesta parte, séries temporais sintéticas com ruído aditivo foram criadas e analisadas conforme apresentado no Capítulo 4 e Apêndice A. A segunda parte dos experimentos e suas análises serão realizadas após a qualificação.

As atividades 5 e 8 estão relacionadas com o componente curricular MATD75 – Exame de qualificação. A atividade 5 refere-se à escrita deste texto e as atividades 6 e 7 representam os componentes curriculares MATA32 – Estágio Docente e MATA31 – Pesquisa Orientada, respectivamente. Além disso, durante a execução destas tarefas foi realizada a prova de proficiência em inglês. A atividade 8 está relacionada à apresentação desta qualificação de mestrado.

A escrita de artigos, listada no item 9 do cronograma, será realizada com base nos resultados gerados com os experimentos (Atividades 3 e 4) e nas contribuições obtidas com a apresentação da qualificação. Por fim, como requisito para a defesa de dissertação, fica pendente a atividade MATE93 -- Defesa de Proposta de Mestrado, a qual se refere aos itens 10 e 11 da Tabela 3.1.

### 3.3 CONSIDERAÇÕES FINAIS

Neste capítulo, apresentou-se de maneira detalhada o projeto de pesquisa, o plano de atividades e o cronograma planejado para a conclusão do mestrado. A seguir, no próximo capítulo, foram discutidos os resultados preliminares realizados com o objetivo de analisar a viabilidade da proposta de mestrado.

# Capítulo

# 4

## EXPERIMENTOS INICIAIS

### 4.1 CONSIDERAÇÕES INICIAIS

Este capítulo apresenta um conjunto de experimentos preliminares realizados sobre 40 séries temporais sintéticas, cujo principal objetivo foi demonstrar a viabilidade da proposta de mestrado. Neste sentido, resultados de análises empíricas justificaram a importância de calcular a distância entre séries temporais considerando, individualmente, as influências estocásticas e determinísticas de seus componentes.

Para execução dos experimentos, a distância entre pares de séries temporais foi calculada da seguinte forma: i) aplicação direta do cálculo da distância entre séries temporais ruidosas; ii) decomposição dos componentes estocásticos e determinísticos de cada série e, posterior, cálculo da distância apenas entre os componentes determinísticos.

Nestes experimentos, o componente estocástico foi descartado como um ruído que não adiciona qualquer nova informação relevante aos dados. Entretanto, no projeto proposto neste trabalho, a estocasticidade presente nas séries será analisada de maneira semelhante ao componente determinístico.

A decomposição das séries em componentes estocásticos e determinísticos foi realizada utilizando as etapas propostas por Rios e Mello (2013, 2016). A próxima seção apresenta como cada série sintética foi produzida para condução dos experimentos.

### 4.2 CONFIGURAÇÃO DOS EXPERIMENTOS

As séries temporais sintéticas utilizadas nos experimentos foram criadas combinando os seguintes componentes:

1. Determinístico: observações geradas a partir da função seno e cosseno com frequência angular igual a  $\pi$  e  $2\pi$ ;
2. Estocástico: observações geradas a partir de um ruído branco com média igual a 0 e desvio padrão entre 0,1 e 1,0;

3. Tendência: observações representando uma tendência linear positiva nos dados.

Estes componentes permitiram gerar grupos de séries temporais da seguinte forma:

- TIPO 1: 10 séries cossenoïdes com ruído branco.
- TIPO 2: 10 séries cossenoide com ruído branco e tendência.
- TIPO 3: 10 séries senoide com ruído branco.
- TIPO 4: 10 séries senoide com ruído branco e tendência.

Todas as séries temporais foram criadas com 3000 observações e detalhes sobre a adição dos ruídos são apresentados na Tabela 4.1.

**Tabela 4.1** Grupos de séries temporais sintéticas utilizadas nos experimentos

Desvio Padrão	TIPO 1	TIPO 2	TIPO 3	TIPO 4
0.1	Série 1.1	Série 2.1	Série 3.1	Série 4.1
0.2	Série 1.2	Série 2.2	Série 3.2	Série 4.2
0.3	Série 1.3	Série 2.3	Série 3.3	Série 4.3
0.4	Série 1.4	Série 2.4	Série 3.4	Série 4.4
0.5	Série 1.5	Série 2.5	Série 3.5	Série 4.5
0.6	Série 1.6	Série 2.6	Série 3.6	Série 4.6
0.7	Série 1.7	Série 2.7	Série 3.7	Série 4.7
0.8	Série 1.8	Série 2.8	Série 3.8	Série 4.8
0.9	Série 1.9	Série 2.9	Série 3.9	Série 4.9
1.0	Série 1.10	Série 2.10	Série 3.10	Série 4.10

Após a criação das séries temporais, suas observações foram normalizadas conforme a Equação 4.1, sendo que  $\hat{X}(t)$  representa a série temporal  $X(t)$  normalizada, e  $\max(\cdot)$  e  $\min(\cdot)$  representam o maior e o menor valor de  $X(t)$ , respectivamente.

$$\hat{X}(T) = \frac{X(t) - \min(X(t))}{\max(X(t)) - \min(X(t))} \quad (4.1)$$

Por fim, a decomposição dos componentes estocásticos e determinísticos das 40 séries normalizadas foi executada utilizando o método EMD (*Empirical Mode Decomposition*), conforme descrito em Rios e Mello (2013, 2016). Inicialmente, cada série foi decomposta em um conjunto de IMFs (*Intrinsic Mode Function*), as quais foram posteriormente combinadas para formar os componentes estocásticos e determinísticos. As IMFs (*Intrinsic Mode Function*) obtidas pelo método EMD (*Empirical Mode Decomposition*) para cada série temporal analisada estão listadas no Apêndice A.

### 4.3 ANÁLISE DE SIMILARIDADE/DISTÂNCIA

Esta seção apresenta um conjunto de experimentos realizados com o objetivo de validar esta proposta de mestrado. Embora os experimentos ainda não sejam suficientes para comprovar a hipótese deste trabalho, a análise empírica apresentada neste exame de qualificação é importante para evidenciar a relevância de analisar, de maneira, individual as influências dos componentes estocásticos e determinísticos no cálculo da similaridade e/ou da distância entre séries temporais.

Para execução dos experimentos, foram selecionadas 6 medidas de distância e similaridade listadas no Capítulo 2: i) DTW; ii) Euclidiana; iii) Manhattan; iv) Minkowski; v) CID; e vi) Cross-Correlation.

O primeiro experimento foi realizado executando usando a DTW para calcular a distância entre duas séries puramente determinísticas criadas usando uma função seno e uma função cosseno. Neste experimento, a distância entre as duas séries foi igual a 90.41829. Em seguida, calculou-se a distância entre pares de séries do Tipo 1 e do Tipo 3. Conforme detalhado na Seção 4.2, essas séries foram criadas a partir da adição de ruído em séries geradas a parir da função seno e cosseno. O resultado desta análise pode ser visto na Tabela 4.2.

**Tabela 4.2** Cálculo da DTW entre séries do Tipo 1 e Tipo 3 (ruído aditivo).

Séries	Distância
(Série 1.1, Série 3.1)	297.34
(Série 1.2, Série 3.2)	523.89
(Série 1.3, Série 3.3)	762.22
(Série 1.4, Série 3.4)	1013.72
(Série 1.5, Série 3.5)	1262.14
(Série 1.6, Série 3.6)	1497.14
(Série 1.7, Série 3.7)	1729.46
(Série 1.8, Série 3.8)	1990.24
(Série 1.9, Série 3.9)	2270.85
(Série 1.10, Série 3.10)	2459.25

Como pode ser observado na Tabela 4.2, o valor da distância entre as séries tende a aumentar à medida que suas observações são mais influenciadas pelo componente estocástico. O mesmo comportamento é observado quando a tendência é adicionada à série, além do componente estocástico, conforme apresentado na Tabela 4.3. Neste experimento, pôde-se notar que a inclusão da tendência não alterou significativamente o comportamento da DTW. Aplicações de um teste de homogeneidade de variância (F-test de Fisher) e do t-student retornaram  $p$  valores iguais a 0,8841 e 0,7002, respectivamente, evidenciando que as distâncias nos dois experimentos são similares.

Em seguida, aplicou-se o método de decomposição de séries em componentes estocásticos e determinísticos sobre as séries do Tipo 1 e do Tipo 3, cujas distâncias foram apresentadas na Tabela 4.2. Nos experimentos com a decomposição, o componente estocástico foi desconsiderado na análise. Por outro lado, as distâncias entre os compo-

**Tabela 4.3** Cálculo da DTW entre séries do Tipo 2 e Tipo 4 (ruído aditivo e tendência).

Séries	Distância
(Série 2.1, Série 4.1)	437.62
(Série 2.2, Série 4.2)	691.64
(Série 2.3, Série 4.3)	944.55
(Série 2.4, Série 4.4)	1159.95
(Série 2.5, Série 4.5)	1418.49
(Série 2.6, Série 4.6)	1634.87
(Série 2.7, Série 4.7)	1847.99
Série 2.8, Série 4.8)	2082.55
Série 2.9, Série 4.9)	2292.85
(Série 2.10, Série 4.10)	2556.13

nentes determinísticos foram calculadas usando DTW. Conforme resultados apresentados na Tabela 4.4, as distâncias entre as séries ruidosas reduziram significativamente com a aplicação da decomposição. Aplicando novamente os testes de hipótese de homogeneidade de variância (F-test de Fisher) e t-student, obteve-se  $p$  valores iguais a  $3,814 \cdot 10^{-08}$  e 0.0005, respectivamente, evidenciando que as médias das distâncias nos dois experimentos não são similares.

**Tabela 4.4** Cálculo da DTW entre séries do Tipo 1 e Tipo 3 com decomposição.

Séries (Componentes Determinísticos)	Distância
(Série 1.1, Série 3.1)	96.43
(Série 1.2, Série 3.2)	91.03
(Série 1.3, Série 3.3)	108.73
(Série 1.4, Série 3.4)	219.18
(Série 1.5, Série 3.5)	156.10
(Série 1.6, Série 3.6)	230.97
(Série 1.7, Série 3.7)	184.60
(Série 1.8, Série 3.8)	59.28
(Série 1.9, Série 3.9)	185.75
(Série 1.10, Série 3.10)	239.52

Por fim, as mesmas etapas de decomposição foram executadas sobre as séries com ruído aditivo e tendência (séries do Tipo 2 e 4). Em seguida, analisou-se a distância entre os componentes determinísticos. Assim como o resultado anterior, foi possível observar nesta análise que os valores obtidos para distância não tiveram comportamento explosivo à medida que a influência do componente estocástico aumentou.

Aplicando de maneira similar os testes de hipótese de homogeneidade de variância (F-test de Fisher) e t-student sobre as distâncias sem decomposição e as distâncias entre os componentes determinístico, obteve-se  $p$  valores iguais a  $4,005 \cdot 10^{-06}$  e 0,0002, respectivamente, evidenciando que as médias das distâncias entre os dois experimentos não são similares.

**Tabela 4.5** Cálculo da DTW entre séries do Tipo 2 e Tipo 4 com decomposição.

Séries (Componentes Determinísticos)	Distância
(Série 2.1, Série 4.1)	246.68
(Série 2.2, Série 4.2)	283.72
(Série 2.3, Série 4.3)	359.91
(Série 2.4, Série 4.4)	351.89
(Série 2.5, Série 4.5)	357.99
(Série 2.6, Série 4.6)	326.95
(Série 2.7, Série 4.7)	54.89
(Série 2.8, Série 4.8)	126.59
(Série 2.9, Série 4.9)	220.07
(Série 2.10, Série 4.10)	197.73

Os experimentos realizados anteriormente foram repetidos alterando apenas a medida de distância utilizada. Neste novo conjunto de experimentos, a DTW foi substituída pela distância Euclidiana. A Tabela 4.6 apresenta os resultados das distâncias calculadas diretamente sobre as séries do Tipo 1 e 3, e sobre seus componentes determinísticos após a etapa de decomposição. É importante destacar que a distância euclidiana calculada sobre as séries puramente determinísticas produzidas a partir das funções seno e cosseno foi igual a 27,38614.

**Tabela 4.6** Cálculo da medida Euclidiana entre séries do Tipo 1 e Tipo 3 sem decomposição e com decomposição.

Séries	Euclidiana (sem decomposição)	Euclidiana (decomposição)
(Série 1.1, Série 3.1)	28.56	27.24
(Série 1.2, Série 3.2)	32.04	27.75
(Série 1.3, Série 3.3)	36.24	28.23
(Série 1.4, Série 3.4)	41.03	27.51
(Série 1.5, Série 3.5)	48.14	24.87
(Série 1.6, Série 3.6)	53.48	29.18
(Série 1.7, Série 3.7)	59.69	27.01
(Série 1.8, Série 3.8)	66.93	27.65
(Série 1.9, Série 3.9)	74.61	26.84
(Série 1.10, Série 3.10)	82.13	26.09

Como pode ser observado nessa tabela, os resultados com a distância euclidiana foram semelhantes aos resultados obtidos com a DTW. Os resultados obtidos com a decomposição foram similares à distância obtida com as séries puramente determinísticas.

Aplicando os testes F-test de Fisher e t-student sobre as distâncias sem decomposição e as distâncias entre os componentes determinísticos, obteve-se  $p$  valores iguais a  $2,291 \cdot 10^{-9}$  e 0.001, respectivamente, evidenciando que as médias das distâncias entre os dois experimentos não são similares. Além disso, pode-se observar que os valores obtidos com a decomposição foram próximos dos valores esperados.

Na Tabela 4.7, os experimentos foram repetidos para as séries do Tipo 2 e 4, i. e., séries cujas observações são compostas por um ruído aditivo quanto pela presença de uma tendência positiva. Os resultados obtidos foram similares aos resultados apresentados na Tabela 4.6. Além disso, os  $p$  valores obtidos para o F-test de Fisher e o t-student foram, respectivamente,  $3,763 \cdot 10^{-07}$  e 0.001, enfatizando as diferenças entre os resultados obtidos.

**Tabela 4.7** Cálculo da medida Euclidiana entre séries do Tipo 2 e Tipo 4 sem decomposição e com decomposição.

Séries	Euclidiana (sem decomposição)	Euclidiana (decomposição)
(Série 2.1, Série 4.1)	28.54	25.14
(Série 2.2, Série 4.2)	31.48	26.55
(Série 2.3, Série 4.3)	35.58	25.72
(Série 2.4, Série 4.4)	41.61	27.21
(Série 2.5, Série 4.5)	48.24	30.85
(Série 2.6, Série 4.6)	54.74	28.60
(Série 2.7, Série 4.7)	60.60	23.41
(Série 2.8, Série 4.8)	68.74	27.41
(Série 2.9, Série 4.9)	74.60	27.56
(Série 2.10, Série 4.10)	82.33	29.18

O próximo conjunto de testes foi realizado utilizando a distância de Manhattan, cujo valor entre as duas séries puramente determinística foi igual a 1350,525. A Tabela 4.8 apresenta as distâncias utilizando essa medida para séries do Tipo 1 e 3 calculadas antes e após a decomposição. Assim como nos experimentos anteriores, além da decomposição permitir obter valores próximos ao esperado (antes da adição de ruído), os testes de hipóteses confirmaram que existe uma diferença entre os resultados obtidos (F-test de Fisher igual a  $7,876 \cdot 10^{-08}$  e o t-student igual a 0.002).

**Tabela 4.8** Cálculo da distância Manhattan entre séries do Tipo 1 e Tipo 3 sem decomposição e com decomposição.

Séries	Manhattan (sem decomposição)	Manhattan (decomposição)
(Série 1.1, Série 3.1)	1380.60	1341.64
(Série 1.2, Série 3.2)	1497.11	1373.13
(Série 1.3, Série 3.3)	1641.60	1394.78
(Série 1.4, Série 3.4)	1817.47	1326.44
(Série 1.5, Série 3.5)	2115.14	1160.13
(Série 1.6, Série 3.6)	2366.99	1401.63
(Série 1.7, Série 3.7)	2602.16	1316.90
(Série 1.8, Série 3.8)	2943.43	1312.15
(Série 1.9, Série 3.9)	3256.53	1278.53
(Série 1.10, Série 3.10)	3629.29	1242.96

Os resultados apresentados na Tabela 4.9, obtidos a partir da adição de tendência, foram similares aos resultados apresentados anteriormente. Neste experimento, os valores obtidos foram próximos do valor esperado e os  $p$  valores do F-test de Fisher e do t-student foram iguais a  $1.66 \cdot 10^{-6}$  e 0.002, respectivamente.

**Tabela 4.9** Cálculo da distância Manhattan entre séries do Tipo 2 e Tipo 4 sem decomposição e com decomposição.

Séries	Manhattan (sem decomposição)	Manhattan (decomposição)
(Série 2.1, Série 4.1)	1379.57	1217.93
(Série 2.2, Série 4.2)	1472.45	1299.72
(Série 2.3, Série 4.3)	1609.70	1228.13
(Série 2.4, Série 4.4)	1848.52	1308.81
(Série 2.5, Série 4.5)	2142.42	1429.08
(Série 2.6, Série 4.6)	2429.22	1370.57
(Série 2.7, Série 4.7)	2678.07	1063.83
(Série 2.8, Série 4.8)	3001.11	1320.86
(Série 2.9, Série 4.9)	3250.85	1328.37
(Série 2.10, Série 4.10)	3597.55	1390.20

Para os experimentos apresentados a seguir, as tabelas contendo as análise entre as séries do Tipo 1 e 3 e as séries do Tipo 2 e 4 foram concatenadas visando discutir seus resultados de maneira mais sucinta. Assim, analisando a Tabela 4.10, observou-se que a decomposição permitiu encontrar valores similares para a distância esperada de Minkowski, cujo valor era 7,663852. Os valores do F-test de Fisher e do t-student foram iguais a  $1,66 \cdot 10^{-6}$  e 0,002, respectivamente. Em seguida, foram analisadas as séries ruidosas com adição de tendência. Os resultados obtidos foram próximos do valor esperado e os  $p$  valores do F-test de Fisher e do t-student foram iguais a  $2.316 \cdot 10^{-6}$  e 0.0008, respectivamente.

O último experimento utilizando medidas de distância foi realizado com a medida CID. A Tabela 4.11 apresenta os resultados obtidos e é possível observar que a decomposição permitiu encontrar valores similares para a distância esperada para as séries puramente determinística, cujo valor era 27,38615. Os valores do F-test de Fisher e do t-student foram iguais a  $6,164 \cdot 10^{-6}$  e 0,004, respectivamente. Em seguida, foram analisadas as séries ruidosas com adição de tendência. Os resultados obtidos foram próximos do valor esperado e os  $p$  valores do F-test de Fisher e do t-student foram iguais a  $5,742 \cdot 10^{-5}$  e 0.004, respectivamente.

O último experimento apresentado neste capítulo foi realizado com a medida de distância Cross-correlation. Esta medida é altamente sensível à qualquer variação nos dados. Por exemplo, a distância entre duas senoides idênticas é igual a 0. Contudo, ao calcular a distância entre uma senoide e uma cossenoide, obtém-se um valor maior que 0, uma vez que as observações entre as duas séries não estão alinhadas. Sendo assim, visando verificar que a adição de ruído e tendência afeta diretamente esta medida, analisou-se apenas as séries Tipo 1 e 2. Com a abordagem estudada neste projeto, espera-se que a decomposição permita correlacionar séries com comportamento determinístico

**Tabela 4.10** Cálculo da distância Minkowski entre séries sem decomposição e com decomposição.

Séries (1 e 3)	Minkowski (sem decomposição)	Minkowski (decomposição)
(Série 1.1, Série 3.1)	8.14	7.64
(Série 1.2, Série 3.2)	9.38	7.76
(Série 1.3, Série 3.3)	10.84	7.90
(Série 1.4, Série 3.4)	12.48	7.81
(Série 1.5, Série 3.5)	14.73	7.21
(Série 1.6, Série 3.6)	16.30	8.33
(Série 1.7, Série 3.7)	18.40	7.62
(Série 1.8, Série 3.8)	20.53	7.93
(Série 1.9, Série 3.9)	22.95	7.69
(Série 1.10, Série 3.10)	25.10	7.53
Séries (2 e 4)	Minkowski (sem decomposição)	Minkowski (decomposição)
(Série 2.1, Série 4.1)	8.13	7.17
(Série 2.2, Série 4.2)	9.22	7.47
(Série 2.3, Série 4.3)	10.68	7.42
(Série 2.4, Série 4.4)	12.63	7.76
(Série 2.5, Série 4.5)	14.64	9.07
(Série 2.6, Série 4.6)	16.70	8.17
(Série 2.7, Série 4.7)	18.52	6.82
(Série 2.8, Série 4.8)	21.15	7.82
(Série 2.9, Série 4.9)	22.92	7.84
(Série 2.10, Série 4.10)	25.33	8.35

semelhantes. Os resultados obtidos estão apresentados na Tabela 4.12 e, de maneira similar aos experimentos anteriores, confirmam a viabilidade da proposta.

Por fim, é importante destacar que os resultados obtidos foram próximos do valor esperado (0) e os  $p$  valores do F-test de Fisher e do t-student foram iguais a  $3,045 \cdot 10^{-09}$  e  $4,809 \cdot 10^{-06}$ , respectivamente.

#### 4.4 CONSIDERAÇÕES FINAIS

Nesta seção, foram apresentados resultados dos primeiros experimentos executados neste trabalho, os quais visavam verificar a importância da utilização da decomposição de séries temporais na utilização de métricas de similaridade/distância.

Nos experimentos futuros, está previsto o estudo de outras medidas para análise do comportamento determinístico, como RQA-RP (*Recurrence Quantification Analysis – Recurrence plot*), e medidas para o comportamento estocástico. Dentre as medidas utilizadas para análise do comportamento estocástico, pode-se citar a transformação dos componentes no domínio de frequência (Transformada de Fourier) para posterior análise dos espectrogramas.

Por fim, em experimentos futuros, espera-se validar a hipótese desse trabalho com a

**Tabela 4.11** Cálculo da distância CID entre séries sem decomposição e com decomposição.

Séries (1 e 3)	CID (sem decomposição)	CID (decomposição)
(Série 1.1, Série 3.1)	28.86	27.83
(Série 1.2, Série 3.2)	33.01	27.92
(Série 1.3, Série 3.3)	36.64	28.36
(Série 1.4, Série 3.4)	41.07	34.30
(Série 1.5, Série 3.5)	48.21	27.80
(Série 1.6, Série 3.6)	54.35	35.18
(Série 1.7, Série 3.7)	61.72	27.39
(Série 1.8, Série 3.8)	67.01	28.80
(Série 1.9, Série 3.9)	77.95	32.64
(Série 1.10, Série 3.10)	83.45	28.54
Séries (2 e 4)	CID (sem decomposição)	CID (decomposição)
(Série 2.1, Série 4.1)	28.91	27.14
(Série 2.2, Série 4.2)	31.82	28.05
(Série 2.3, Série 4.3)	36.73	29.05
(Série 2.4, Série 4.4)	42.53	35.78
(Série 2.5, Série 4.5)	48.82	33.50
(Série 2.6, Série 4.6)	56.56	36.49
(Série 2.7, Série 4.7)	61.01	24.30
(Série 2.8, Série 4.8)	69.87	28.68
(Série 2.9, Série 4.9)	75.93	28.45
(Série 2.10, Série 4.10)	85.16	29.56

**Tabela 4.12** Cálculo da distância Cross Correlation entre séries sem decomposição e com decomposição.

Séries (1 e 2)	Cross (sem decomposição)	Cross (decomposição)
(Série 1.1, Série 2.1)	0.24	0.03
(Série 1.2, Série 2.2)	0.26	0.01
(Série 1.3, Série 2.3)	0.30	0.01
(Série 1.4, Série 2.4)	0.35	0.02
(Série 1.5, Série 2.5)	0.42	0.03
(Série 1.6, Série 2.6)	0.44	0.03
(Série 1.7, Série 2.7)	0.50	0.03
(Série 1.8, Série 2.8)	0.60	0.02
(Série 1.9, Série 2.9)	0.59	0.03
(Série 1.10, Série 2.10)	0.52	0.02

aplicação de todas as etapas do agrupamento de séries temporais, incluindo a execução de algoritmos de agrupamento e a aplicação de técnicas de validação. Espera-se, ainda, realizar uma aplicação prática da abordagem proposta, não apenas em dados sintéticos, mas em um conjunto de dados coletados a partir de um sistema do mundo real.



## REFERÊNCIAS BIBLIOGRÁFICAS

- AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering – a decade review. *Information Systems*, v. 53, p. 16 – 38, 2015.
- AYHAN, S.; SAMET, H. Time series clustering of weather observations in predicting climb phase of aircraft trajectories. In: *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science*. New York, NY, USA: ACM, 2016. (IWCTS '16), p. 25–30.
- BAGNALL, A.; JANACEK, G. Clustering time series with clipped data. *Machine Learning*, v. 58, n. 2, p. 151–178, 2005.
- BATISTA, G. E. A. P. A. et al. Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, v. 28, n. 3, p. 634–669, 2014.
- BLEIWEISS, A. Beat discovery from dimensionality reduced perspective streams of electrocardiogram signal data. In: *2015 12th International Joint Conference on e-Business and Telecommunications (ICETE)*. [S.l.: s.n.], 2015. v. 05, p. 39–48.
- BOX, G. et al. *Time Series Analysis: Forecasting and Control*. [S.l.]: Wiley, 2015. (Wiley Series in Probability and Statistics).
- CHEN, L.; NG, R. On the marriage of lp-norms and edit distance. In: *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*. [S.l.]: VLDB Endowment, 2004. (VLDB '04), p. 792–803.
- CHOUAKRIA, A. D.; NAGABHUSHAN, P. N. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, v. 1, n. 1, p. 5–21, 2007.
- CRYER, J.; CHAN, K. *Time Series Analysis: With Applications in R*. [S.l.]: Springer New York, 2008. (Springer Texts in Statistics).
- DURANTE, F.; PAPPADÀ, R.; TORELLI, N. Clustering of financial time series in risky scenarios. *Advances in Data Analysis and Classification*, Springer, v. 8, n. 4, p. 359–376, 2014.
- ESLING, P.; AGON, C. Time-series data mining. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 45, n. 1, p. 12:1–12:34, dez. 2012.
- FACELI, K. et al. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.: s.n.], 2011. 192 p.

- FEI, W. et al. Similarity analysis on nonstationary time series. In: *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. [S.l.: s.n.], 2009. v. 1, p. 286–290.
- HAN, M.; LIU, Y. Noise reduction method for chaotic signals based on dual-wavelet and spatial correlation. *Expert Systems with Applications*, v. 36, n. 6, p. 10060 – 10067, 2009.
- HÖPPNER, F.; KLAWONN, F. Compensation of translational displacement in time series clustering using cross correlation. In: ADAMS, N. M. et al. (Ed.). *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA, Lyon, France, August 31 - September 2*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 71–82.
- HUANG, N. E. et al. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, The Royal Society, v. 454, n. 1971, p. 903–995, 1998.
- IZAKIAN, H.; PEDRYCZ, W.; JAMAL, I. Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 39, p. 235–244, 2015.
- JUNKUI, L.; YUANZHEN, W.; XINPING, L. Lb hust: A symmetrical boundary distance for clustering time series. In: *Information Technology, 2006. ICIT '06. 9th International Conference on*. [S.l.: s.n.], 2006. p. 203–208.
- KEOGH, E. Exact indexing of dynamic time warping. In: *Proceedings of the 28th International Conference on Very Large Data Bases*. [S.l.]: VLDB Endowment, 2002. (VLDB '02), p. 406–417.
- LIAO, T. W. Clustering of time series data—a survey. *Pattern Recognition*, v. 38, n. 11, p. 1857 – 1874, 2005.
- MEESRIKAMOLKUL, W.; NIENNATTRAKUL, V.; RATANAMAHATANA, C. A. Shape-based clustering for time series data. In: TAN, P.-N. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD, Kuala Lumpur, Malaysia, May 29-June 1, Proceedings, Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 530–541.
- MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997.
- MOREIRA JUNIOR, E. M. *Estudo do Comportamento de Medidas de Similaridade Utilizadas no Agrupamento de Séries Temporais Ruidosas*. 2016. Monografia (Bacharel em Ciência da Computação), UFBA (Universidade Federal da Bahia), Salvador, Brasil.
- MORETTIN, P. A.; TOLOI, C. *Análise de séries temporais*. [S.l.]: Blucher, 2006.

- MORI, U.; MENDIBURU, A.; LOZANO, J. A. Similarity measure selection for clustering time series databases. *IEEE Transactions on Knowledge and Data Engineering*, v. 28, n. 1, p. 181–195, Jan 2016.
- NGUYEN, H.-L.; WOON, Y.-K.; NG, W.-K. A survey on data stream clustering and classification. *Knowledge and Information Systems*, v. 45, n. 3, p. 535–569, 2015.
- OH, S. R.; KIM, Y. G. Security requirements analysis for the iot. In: *2017 International Conference on Platform Technology and Service (PlatCon)*. [S.l.: s.n.], 2017. p. 1–6.
- RIOS, R. A.; MELLO, R. F. de. Improving time series modeling by decomposing and analyzing stochastic and deterministic influences. *Signal Processing*, v. 93, n. 11, p. 3001 – 3013, 2013.
- RIOS, R. A.; MELLO, R. F. de. Applying empirical mode decomposition and mutual information to separate stochastic and deterministic influences embedded in signals. *Signal Processing*, v. 118, p. 159 – 176, 2016.
- SHUMWAY, R.; STOFFER, D. *Time series analysis and its applications: With R examples, 2nd edn.* [S.l.]: New York: Springer, 2006.
- TORMENE, P. et al. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine*, Elsevier, v. 45, n. 1, p. 11–34, 2009.
- ZHANG, X. et al. A novel clustering method on time series data. *Expert Systems with Applications*, Elsevier, v. 38, n. 9, p. 11891–11900, 2011.



# Apêndice

# A

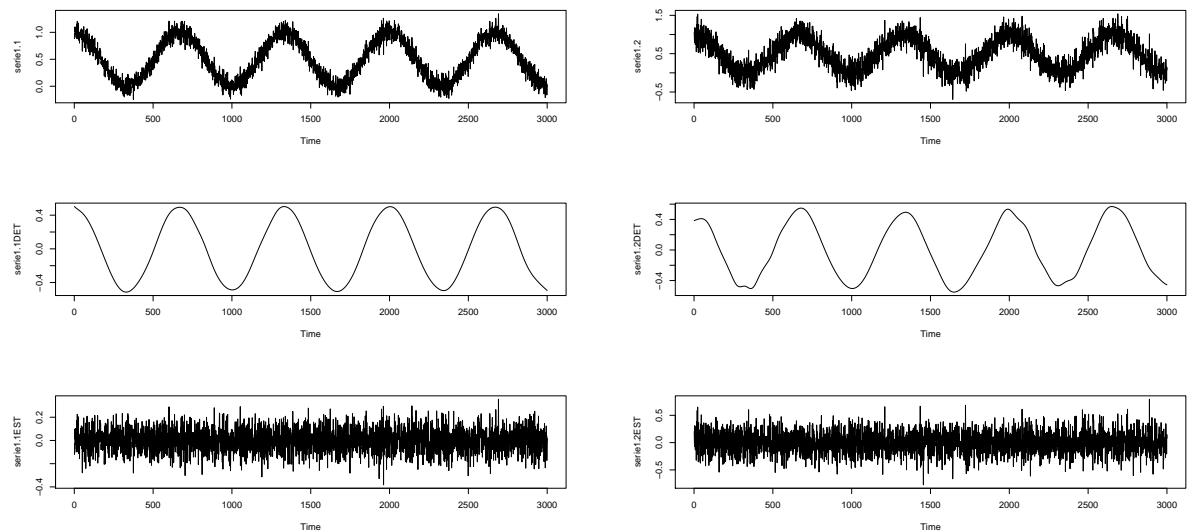
## DECOMPOSIÇÃO DAS SÉRIES TEMPORAIS

### A.1 CONSIDERAÇÕES INICIAIS

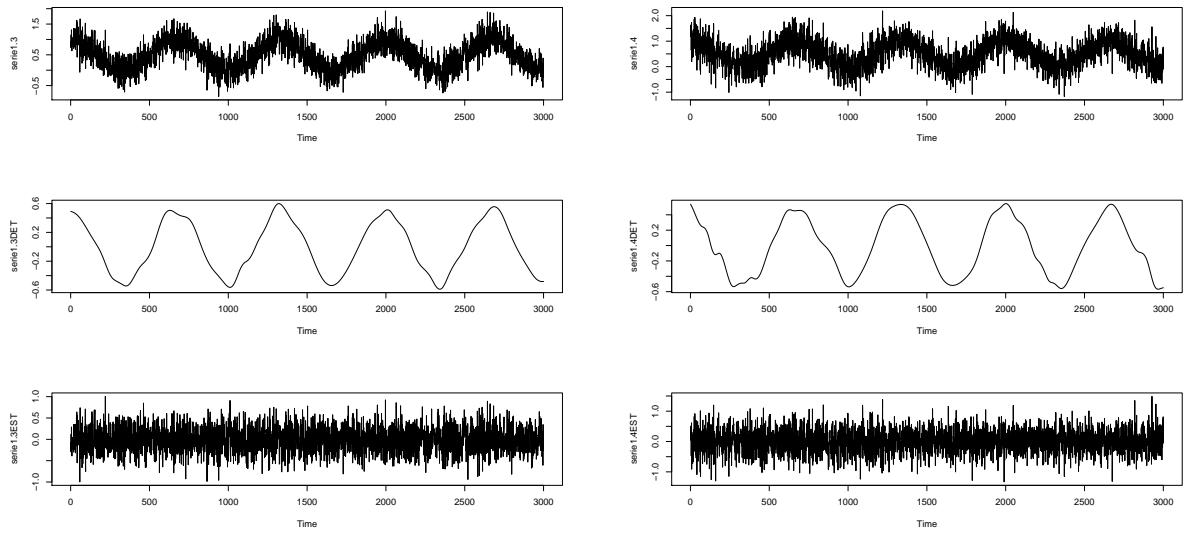
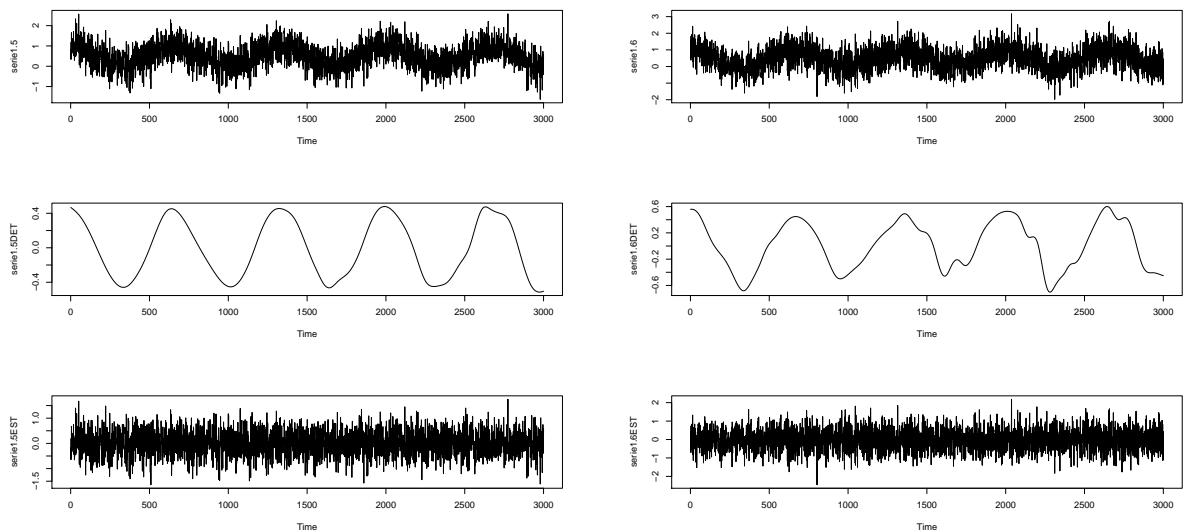
Neste apêndice consta as 40 séries temporais utilizadas nos experimentos mostrados no Capítulo 4. As séries foram divididas em 4 tipos conforme a Tabela 4.1, onde o tipo representa um conjunto de 10 séries senoide ou cossenoide, sendo acrescida de ruído ou acrescida de ruído e tendência. Nas imagens são representadas, a série original, seu componente determinístico e seu componente estocástico, os quais foram extraídos após a decomposição.

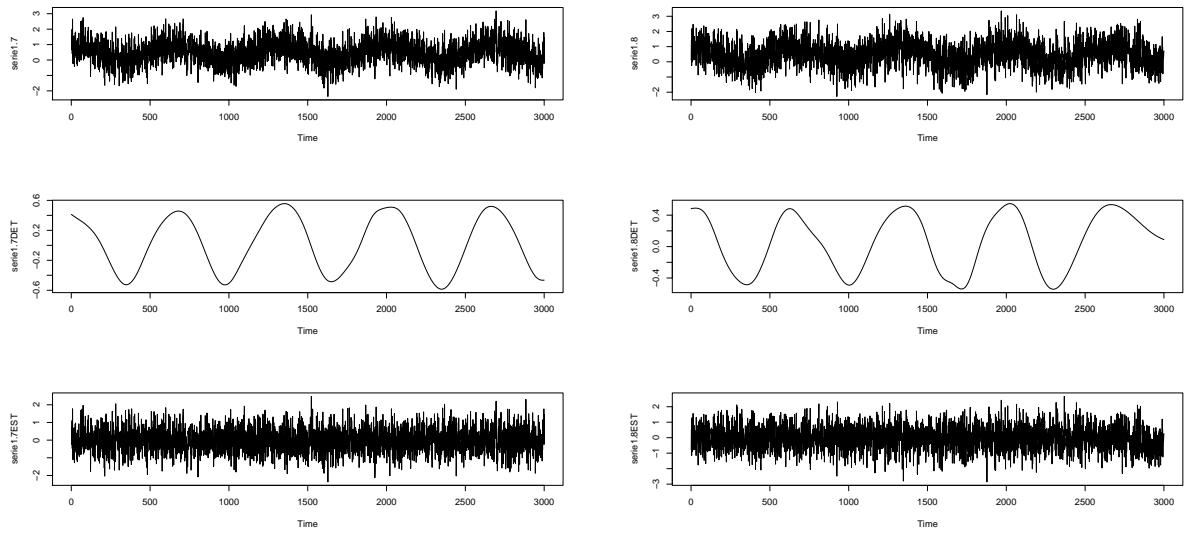
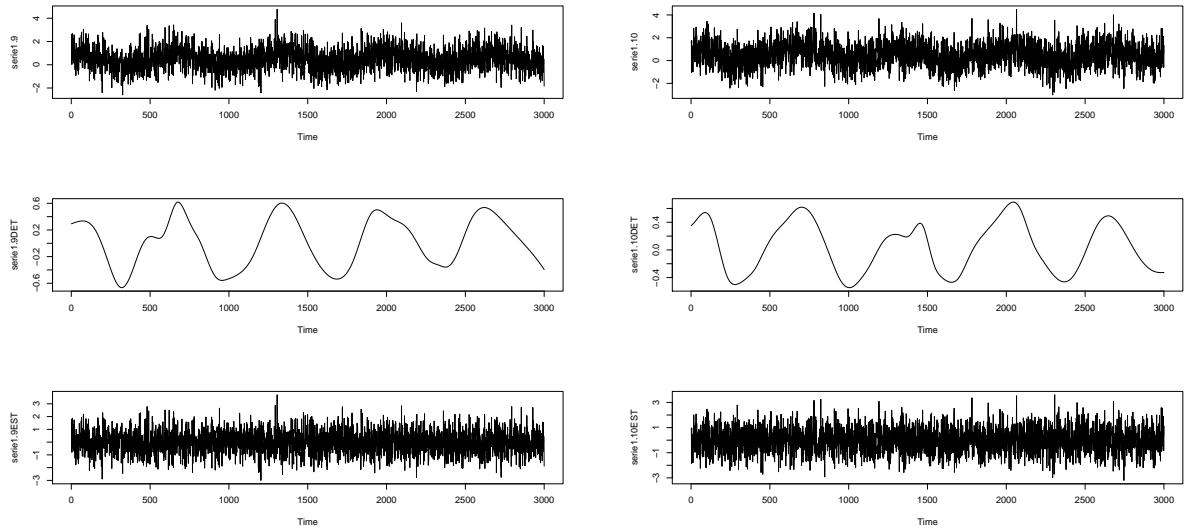
### A.2 SÉRIES TIPO 1

10 séries cossenoide com ruído ao longo da série.

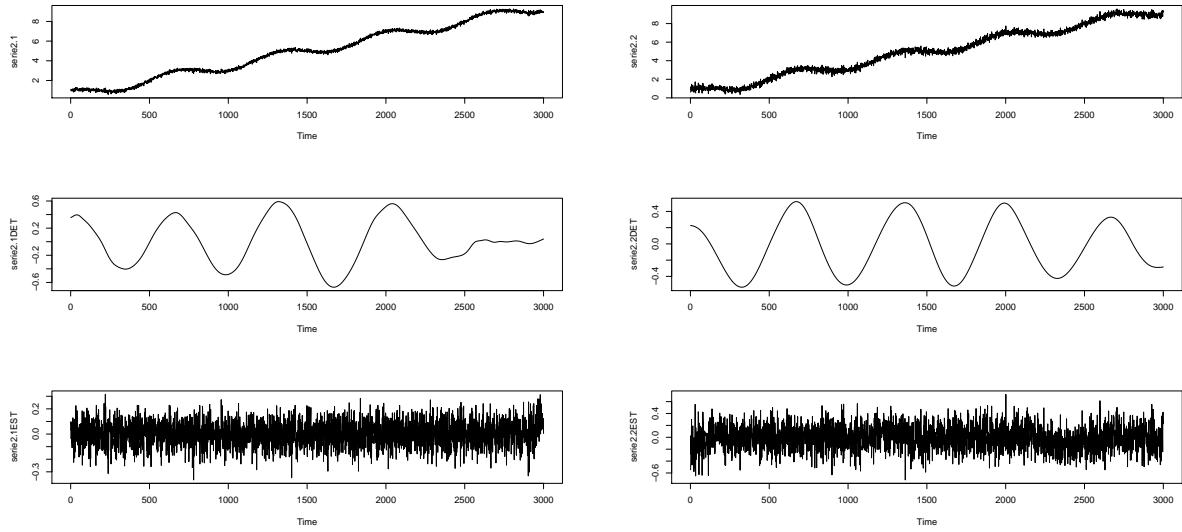
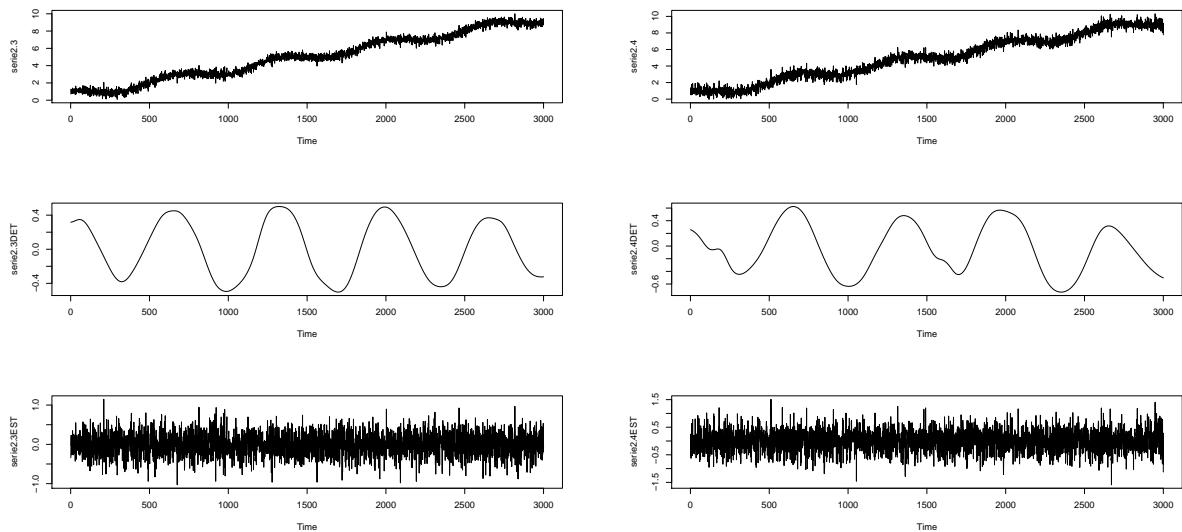


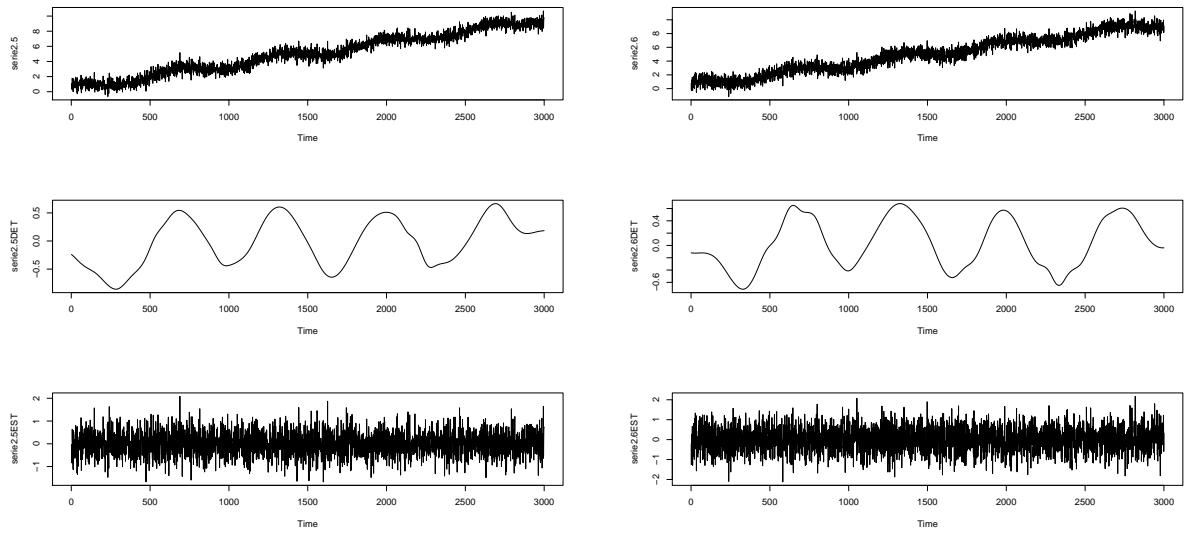
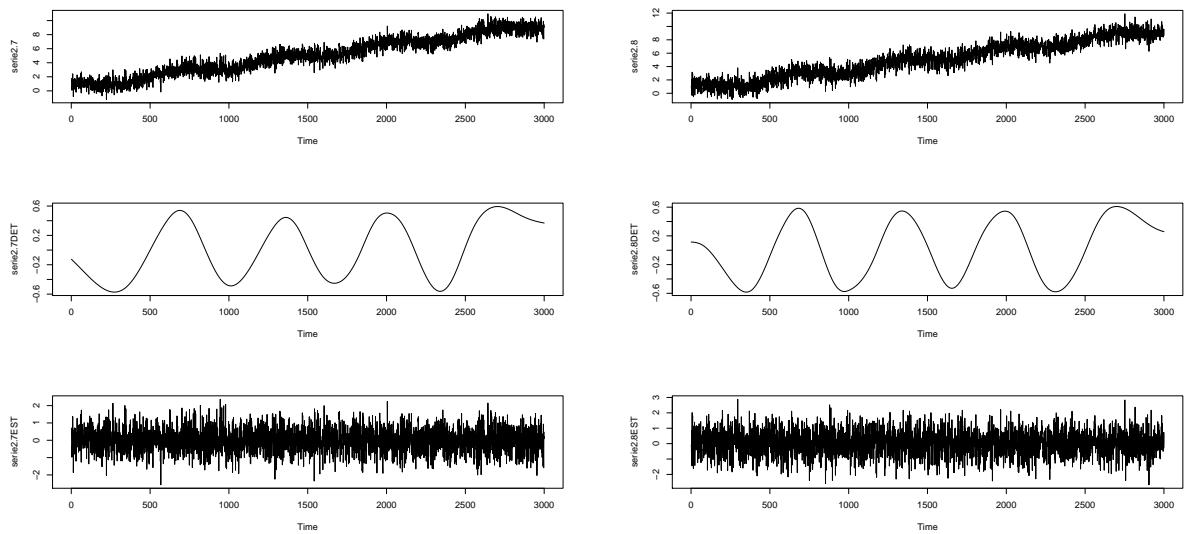
**Figura A.1** Série 1.1 e Série 1.2

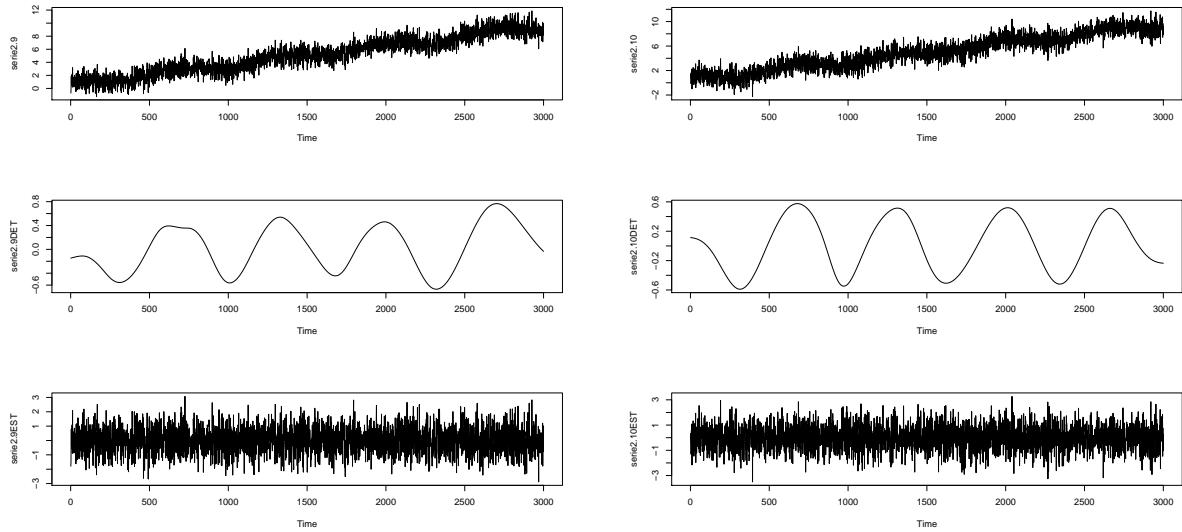
**Figura A.2** Série 1.3 e Série 1.4**Figura A.3** Série 1.5 e Série 1.6

**Figura A.4** Série 1.7 e Série 1.8**Figura A.5** Série 1.9 e Série 1.10**A.3 SÉRIES TIPO 2**

10 séries cossenoide com ruído ao longo da série e tendência.

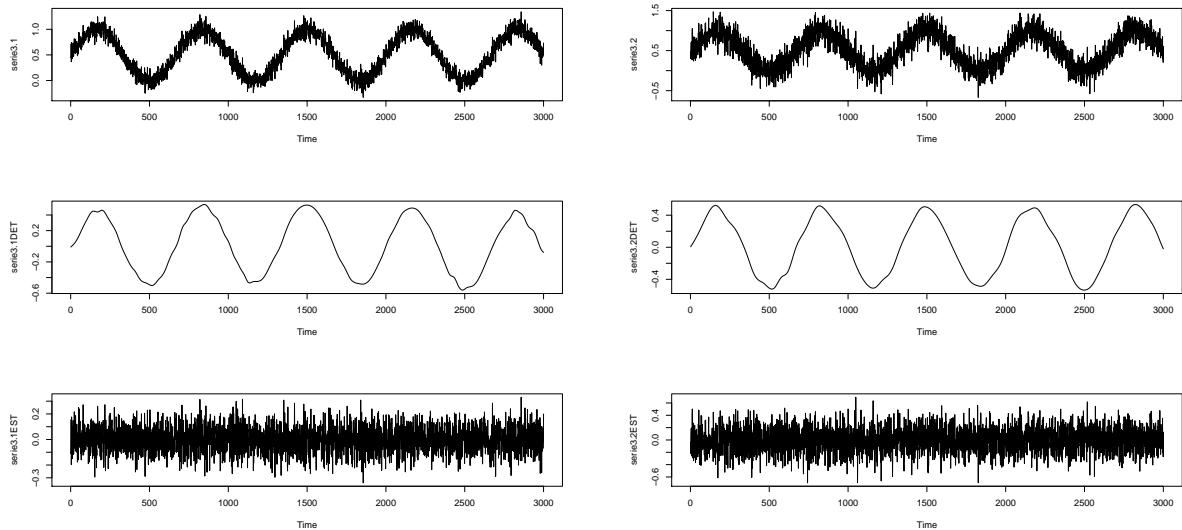
**Figura A.6** Série 2.1 e Série 2.2**Figura A.7** Série 2.3 e Série 2.4

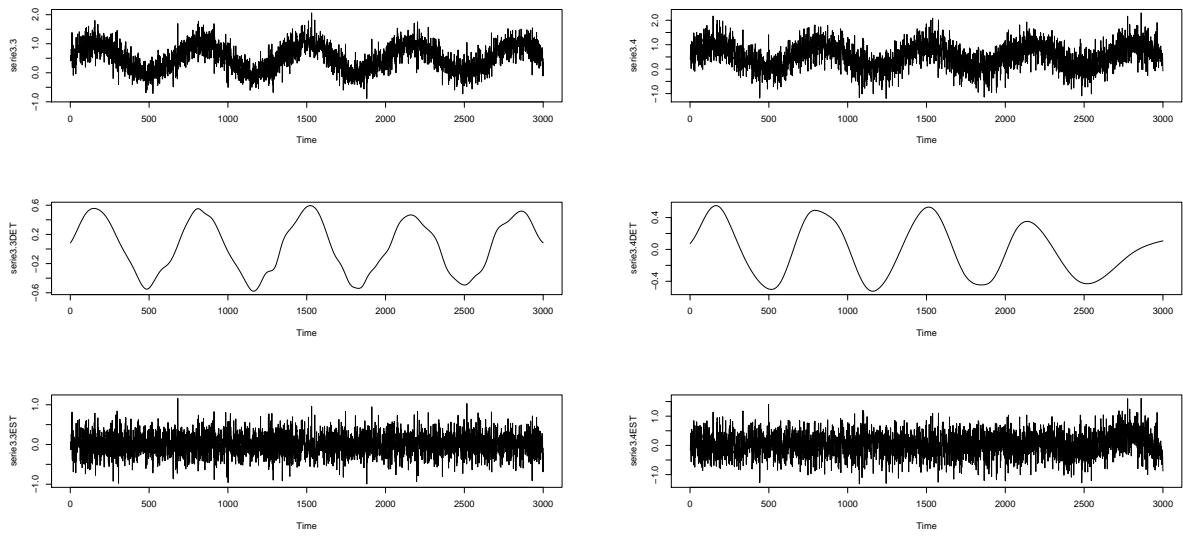
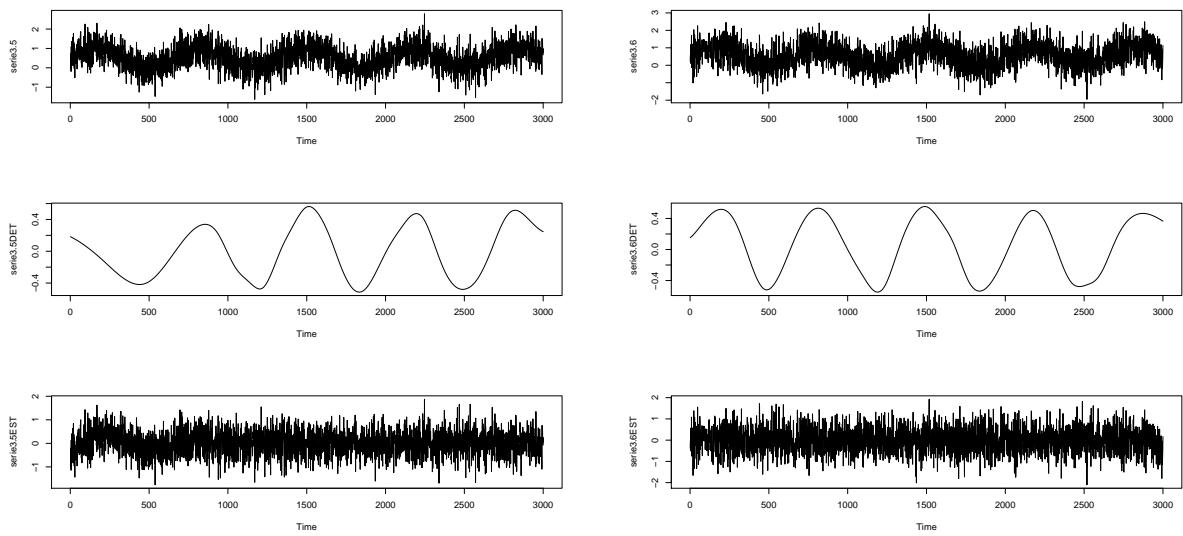
**Figura A.8** Série 2.5 e Série 2.6**Figura A.9** Série 2.7 e Série 2.8

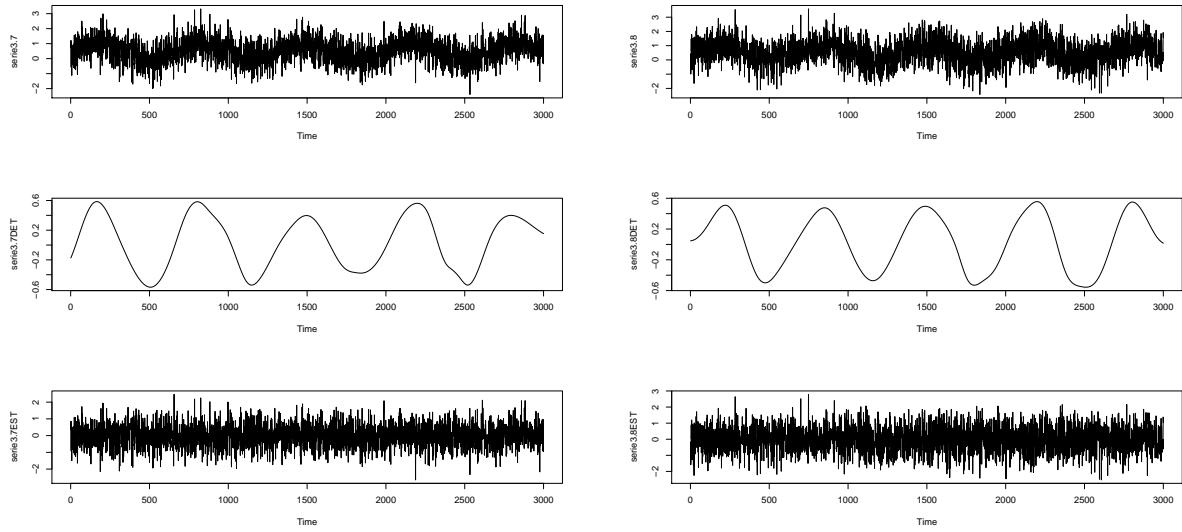
**Figura A.10** Série 2.9 e Série 2.10

#### A.4 SÉRIES TIPO 3

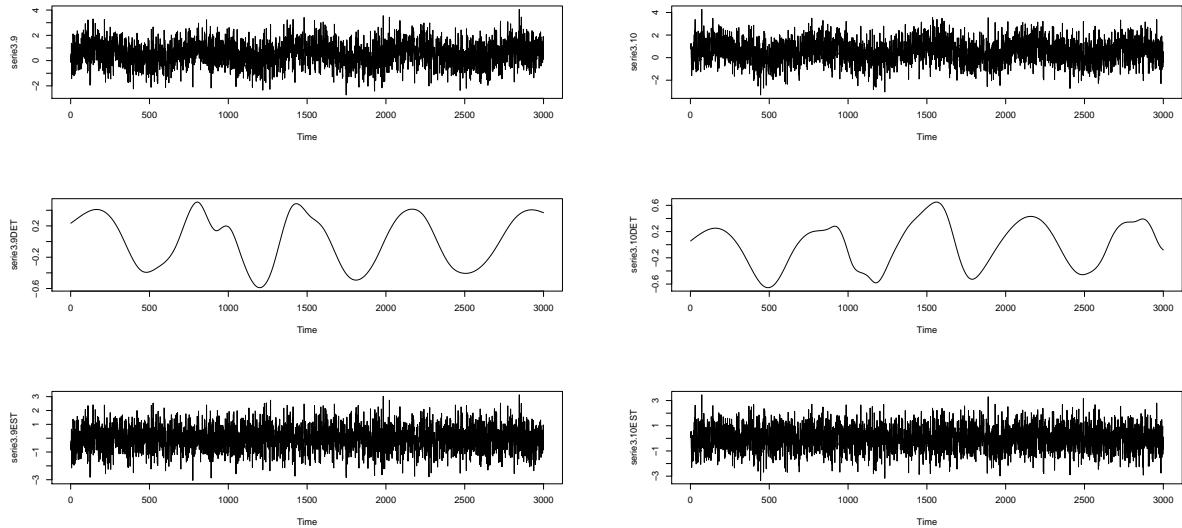
10 séries senoide com ruído ao longo da série.

**Figura A.11** Série 3.1 e Série 3.2

**Figura A.12** Série 3.3 e Série 3.4**Figura A.13** Série 3.5 e Série 3.6



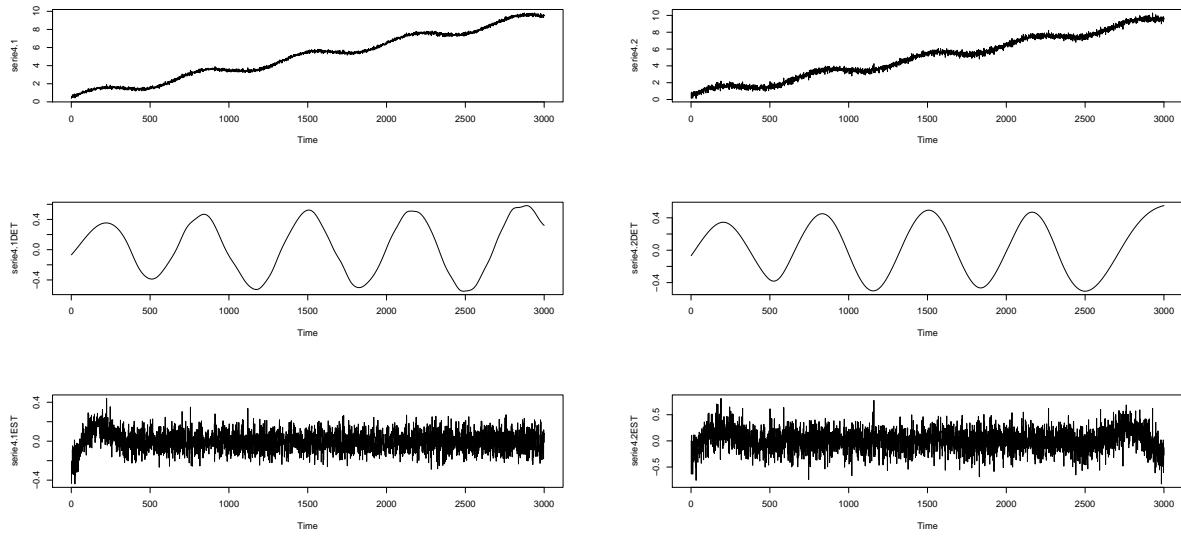
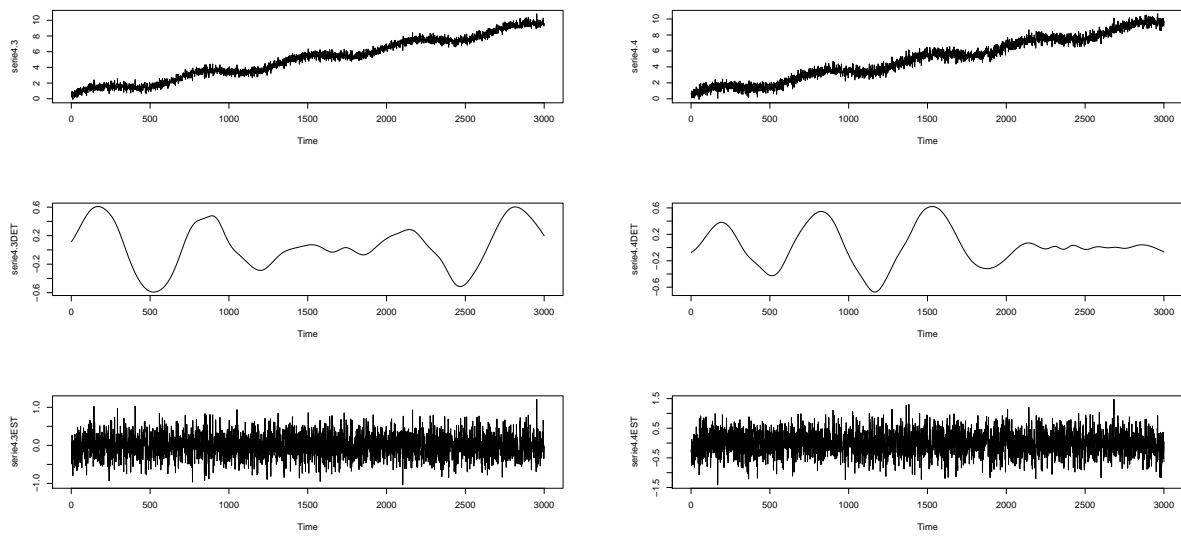
**Figura A.14** Série 3.7 e Série 3.8

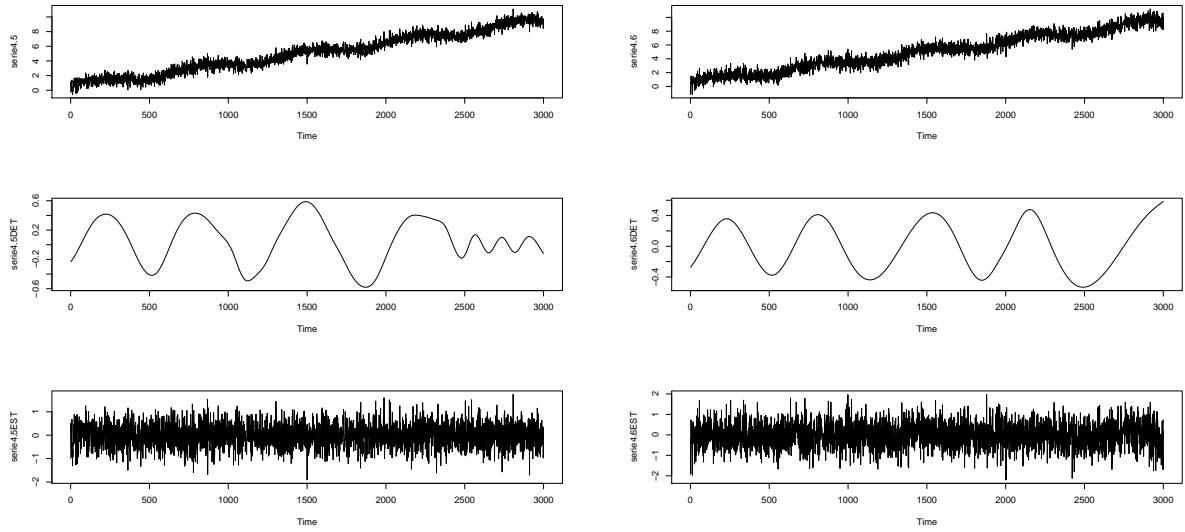
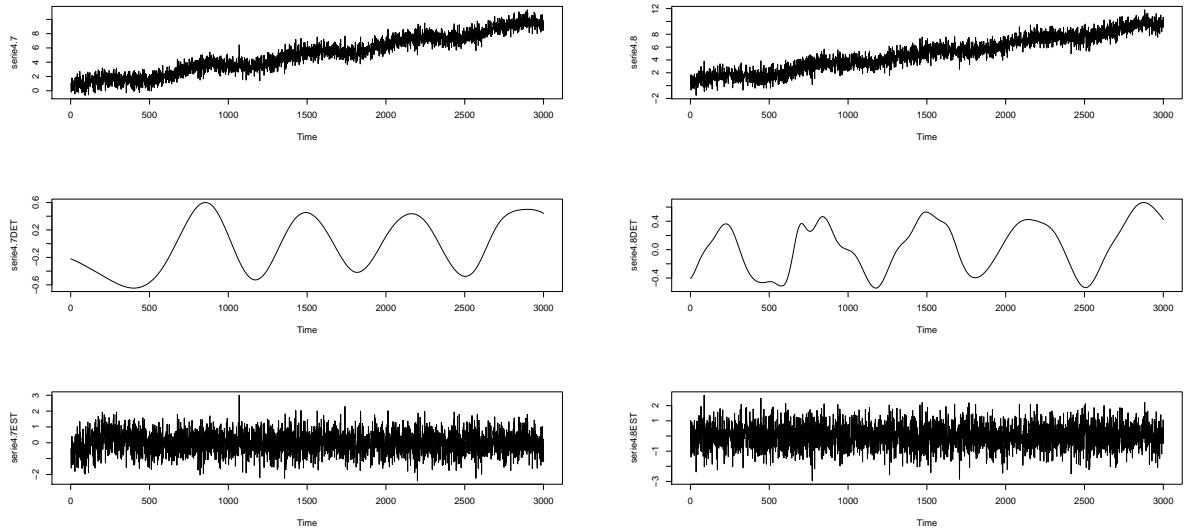


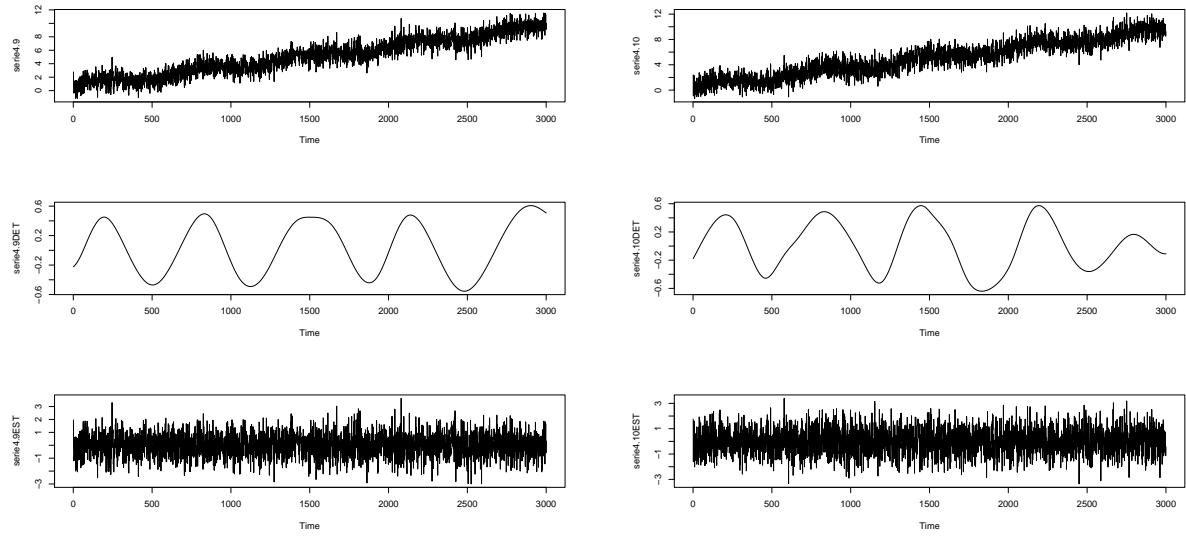
**Figura A.15** Série 3.9 e Série 3.10

## A.5 SÉRIES TIPO 4

10 séries senoide com ruído ao longo da série e tendência.

**Figura A.16** Série 4.1 e Série 4.2**Figura A.17** Série 4.3 e Série 4.4

**Figura A.18** Série 4.5 e Série 4.6**Figura A.19** Série 4.7 e Série 4.8

**Figura A.20** Série 4.9 e Série 4.10**A.6 CONSIDERAÇÕES FINAIS**

Foram apresentadas as séries temporais utilizadas neste trabalho experimental e suas respectivas decomposições.