Universidade Federal da Bahia
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

AGRUPAMENTO DE SÉRIES TEMPORAIS UTILIZANDO DECOMPOSIÇÃO DE COMPONENTES ESTOCÁSTICOS E DETERMINÍSTICOS

Mirlei Moura da Silva

DISSERTAÇÃO DE MESTRADO

Salvador
13 de Julho de 2018

MIRLEI MOURA DA SILVA

AGRUPAMENTO DE SÉRIES TEMPORAIS UTILIZANDO DECOMPOSIÇÃO DE COMPONENTES ESTOCÁSTICOS E DETERMINÍSTICOS

Esta Dissertação 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
13 de Julho de 2018

Sistema de Bibliotecas - UFBA

Silva, Mirlei Moura da.

Agrupamento de Séries Temporais Utilizando Decomposição de Componentes Estocásticos e Determinísticos / Mirlei Moura da Silva – Salvador, 2018.

74p.: il.

Orientador: Prof. Dr. Ricardo Araújo Rios.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática, 2018.

1. Primeira palavra-chave. 2. Segunda palavra-chave. 3. Terceira palavra-chave. I. Rios, Ricardo Araújo. II. Universidade Federal da Bahia. Instituto de Matemática. III Título.

CDD – XXX.XX

CDU – XXX.XX.XXX

TERMO DE APROVAÇÃO

MIRLEI MOURA DA SILVA

AGRUPAMENTO DE SÉRIES TEMPORAIS UTILIZANDO DECOMPOSIÇÃO DE COMPONENTES ESTOCÁSTICOS E DETERMINÍSTICOS

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 13 de Julho de 2018

Profa. Dra. Gecynalda Soares da Silva Gomes
Universidade Federal da Bahia

Prof. Dr. Rodrigo Fernandes de Mello
Universidade de São Paulo

Prof. Dr. Ricardo Araújo Rios
Universidade Federal da Bahia

AGRADECIMENTOS

Agradecer é o verbo que eu carrego ao concluir esse trabalho, eu sou e serei eternamente grata a cada pessoa que de alguma forma me trouxe até aqui. Agradeço a Deus por me dar forças e iluminação. Agradeço aos meus pais, pois graças aos seus esforços eu consegui chegar até aqui. À minha pequena, agora grande e inteligente, irmã Maria Eulina. Agradeço ao meu namorado e companheiro Rafael Raña, sempre me apoiando e incentivando com muita paciência e amor. Agradeço ao meu orientador Ricardo Rios pelo seu profissionalismo durante todo processo de execução do trabalho, seu excelente trabalho como orientador e docente fez com que esse trabalho acontecesse e o que ganhei foi aprendizado, cresci muito e tenho um exemplo a ser seguido. Agradeço à banca de qualificação pelas orientações referente ao trabalho. Aos meus colegas do mestrado pelos grupos de estudos nas disciplinas e pelos momentos de conforto. Ao aluno Junior por permitir acompanhar seu trabalho no inicio do meu curso. Ao grupo de pesquisa AI love coffee por me direcionar no momento final do trabalho.

Agradeço ao Programa de Pós Graduação em Ciência da Computação pela oportunidade e à FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia) pelo apoio financeiro à realização desse trabalho.

"A persistência é o menor caminho do êxito".

—CHARLES CHAPLIN

RESUMO

Com a grande quantidade de dados produzidos e coletados diariamente por diferentes sistemas, técnicas de aprendizado de máquina vêm sendo propostas com o intuito de auxiliar o 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. Este trabalho de mestrado apresenta uma nova abordagem de agrupamento para dados, que possuem uma dependência temporal entre suas observações, conhecidos como 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 avalia a importância da decomposição de séries temporais em componentes estocásticos e determinísticos no processo de agrupamento. Com isso, dados são agrupados analisando de maneira individual a similaridade entre cada componente. Os experimentos foram realizados em quatro etapas: i) a primeira etapa consistiu em verificar medidas para os componentes determinísticos; ii) na segunda etapa, foi proposta uma medida específica para os componentes estocásticos; iii) a terceira etapa consistiu em avaliar o uso das medidas considerando os componentes estocásticos e determinísticos de séries temporais; iv) e, por fim, foi realizado o processo de agrupamento de séries temporais com ruído aditivo. A partir da decomposição das séries temporais, foi possível observar que o processo de agrupamento melhorou significativamente os valores dos índices de validação externa.

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

ABSTRACT

As part of the unsupervised machine learning area, time series clustering aims at designing methods to extract patterns from temporal data in order to organize different series according to their similarities. According to the literature, most of researches either perform a preprocessing step to convert time series into an attribute-value matrix to be later analyzed by traditional clustering methods, or apply measures specifically designed to compute the similarity among time series. Based on such studies, we have noticed two main issues: i) clustering methods do not take into account the stochastic and the deterministic influences inherent of time series from real-world scenarios; and ii) similarity measures tend to look for recurrent patterns, which may not be available in stochastic time series. In order to overcome such drawbacks, we present a new clustering approach that considers both influences and a new similarity measure to deal with purely stochastic time series. Experiments provided outstanding results, emphasizing time series are better clustered when their stochastic and deterministic influences are properly analyzed.

Keywords: Time Series Clustering, Decomposition, Stochasticity, Determinism.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

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, todos os dias, mais de 3,5 bilhões de buscas são realizadas nos repositórios do Google e cerca de 4TB de imagens são gerados 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 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 e Kim 2017], a qual visa conectar e coordenar diferentes dispositivos sem a necessidade de 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 e Kim 2017].

Esse aumento significativo na quantidade de dados produzidos e armazenados tem dificultado a tarefa de especialistas na análise e extração de 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 coletados previamente com intuito de melhorar o desempenho na realização de alguma tarefa [Mitchell 1997, Faceli 2011, Bishop 2006].

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 a dados não vistos previamente [Bishop 2006].

Ao encontrar hipóteses que descrevem o comportamento dos dados, pode-se ajustar um modelo para, por exemplo, predizer o comportamento de um sistema ou descrever seu estado atual. O ajuste desses modelos ocorre de acordo com o paradigma de aprendizado, o qual pode ser supervisionado, baseado em reforço, semi-supervisionado e não-supervisionado [Bishop 2006, Mitchell 1997]. A pesquisa apresentada neste trabalho foi desenvolvida considerando o paradigma não supervisionado. Neste paradigma, métodos são ajustados sobre as características (atributos) dos dados, visando extrair padrões e novas informações, sem considerar qualquer informação previamente fornecida por especialistas.

Dentre as técnicas do paradigma não-supervisionado mais utilizadas na literatura, pode-se destacar os algoritmos de agrupamento. Esses algoritmos analisam dados buscando encontrar estruturas, nas quais instâncias mais semelhantes entre si são organizadas em um mesmo grupo [Mitchell 1997, Bishop 2006, Aghabozorgi, Shirkhorshidi e 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 amostrados a partir de uma distribuição de probabilidade independente e identicamente distribuída (iid), ou seja, seguem alguma distribuição de probabilidades sem, necessariamente, serem caracterizados por uma dependência ao longo do tempo. Entretanto, quando existe essa dependência, os dados são organizados como séries temporais [Esling e Agon 2012, Box 2015].

Resumidamente, séries temporais representam uma coleção de valores sequencialmente obtidos a partir de medidas realizadas em um sistema durante um determinado intervalo de tempo. Por exemplo, séries temporais têm sido amplamente utilizadas para representar médias diárias de temperatura medidas em uma cidade, variações de preço de uma determinada ação na bolsa de valores e propagação de doenças [Esling e Agon 2012].

Neste trabalho de mestrado, o objetivo principal é propor uma nova abordagem de agrupamento de séries temporais conforme detalhado na próxima seção.

1.2 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Nesta dissertação, assume-se como definição básica que técnicas de agrupamento visam extrair padrões de tal forma que dados com características semelhantes são organizadas em um mesmo grupo [Xu e Wunsch 2009].

Para estimar tais semelhanças, são utilizadas medidas que calculam similaridades entre objetos de uma base de dados [Mori, Mendiburu e Lozano 2016]. Na literatura de aprendizado de máquina não-supervisionado, a semelhança entre dados pode ser medida, ainda, pela distância (dissimilaridade)[1] entre suas características.

No contexto de séries temporais, a escolha de tais métricas não é trivial, uma vez que a dependência entre suas observações pode apresentar diferentes comportamentos que influenciam o cálculo da similaridade/distância. Por exemplo, a distância entre séries

[1]Em geral, a relação entre similaridade e distância é dada por $Distância = 1 - Similaridade$ se as medidas estiverem no intervalo $[0, 1]$.

com comportamento determinístico pode ser calculada utilizando técnicas como DTW (*Dynamic Time Warping*) [Tormene 2009] ou MDDL (*Mean Distance from the Diagonal Line*) [Rios e Mello 2013]. Por outro lado, o cálculo da distância entre séries com comportamento estocástico pode ser realizado a partir da análise no domínio de frequência, e.g., comparando espectrogramas obtidos por meio da transformada de Fourier [Morettin e Toloi 2006].

Contudo, séries temporais podem apresentar uma mistura de ambos os comportamentos, conforme discutido em Rios e Mello 2013. Ao considerar apenas um dos comportamentos para calcular a distância/similaridade entre as séries, o desempenho no processo de agrupamento de séries temporais pode ser consideravelmente afetado.

Visando resolver essa limitação, esta dissertação de mestrado apresenta uma abordagem que permite analisar individualmente as influências estocásticas e determinísticas presentes em séries temporais, a fim de melhorar o resultado de técnicas de agrupamento.

1.3 HIPÓTESE E OBJETIVO

Com base na limitação citada anteriormente, a seguinte hipótese foi formulada para condução deste trabalho:

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

O objetivo deste trabalho é validar essa hipótese melhorando o agrupamento de séries temporais por meio da decomposição. Para alcançar esse objetivo, os experimentos foram executados seguindo os seguintes passos: i) Inicialmente, todas as séries temporais foram decompostas em dois componentes, um estocástico e outro determinístico; ii) Em seguida, a distância entre os componentes determinísticos foram calculadas utilizando medidas comumente usadas em séries temporais, enquanto que a distância entre os componentes estocásticos foram calculadas com a medida proposta neste trabalho; iii) Finalmente, as medidas estocásticas e determinísticas foram combinadas para gerar um valor de distância que considera não apenas a influência determinística, mas também a influência estocástica. Os resultados obtidos enfatizaram a importância de utilizar o processo de decomposição para agrupar séries temporais ruidosas.

1.4 CONSIDERAÇÕES FINAIS

Esta dissertação de mestrado está organizada da seguinte maneira: O **Capítulo 2** possui uma revisão bibliográfica dos principais conceitos utilizados neste trabalho como, por exemplo, decomposição, agrupamento de séries temporais e as principais medidas utilizadas em agrupamento de séries temporais; No **Capítulo 3** é detalhada a abordagem proposta neste trabalho ainda como a medida desenvolvida para componentes estocásticos; No **Capítulo 4**, é apresentado o conjunto de experimentos, resultados e discussões; Por fim, o **Capítulo 5** apresenta as conclusões obtidas neste trabalho.

Capítulo

2

SÉRIES TEMPORAIS E AGRUPAMENTO DE SÉRIES

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 com características mais similares. Quando dados analisados são organizados como Séries Temporais, a complexidade no cálculo da similaridade/distância aumenta devido à presença de dependência temporal entre as observações [Aghabozorgi, Shirkhorshidi e Wah 2015, Zhang 2011].

A dependência temporal entre observações de séries temporais é diretamente influenciada pela presença de diferentes componentes como, por exemplo, estocasticidade e determinismo. Conforme discutido em Rios e Mello 2013, Rios e Mello 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 e Mello 2013, Rios e Mello 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 de pesquisa abordados neste projeto. Inicialmente, serão apresentados conceitos básicos sobre 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, como os algoritmos de agrupamento, técnicas de validação e as principais medidas de distância/similaridade utilizadas na literatura.

2.2 SÉRIES TEMPORAIS

Neste trabalho, uma série temporal é definida como uma sequência de observações coletadas ao longo do tempo na forma $X_t = \{x_0, x_1, x_2, ..., x_T\}$ [Box 2015, Chouakria e Nagabhushan 2007, Morettin e Toloi 2006]. Uma série temporal univariada é composta por valores escalares, enquanto que as multivariadas possem múltiplas dimensões dentro da mesma faixa de tempo [Box 2015, Chouakria e Nagabhushan 2007]. Séries temporais podem ser caracterizadas, ainda, pelo intervalo entre coleta de observações que pode ser discreta ou contínua. Geralmente, uma série temporal contínua no intervalo $[0, T]$ pode ser amostrada em intervalos de tempos fixos Δt, gerando uma série discreta [Box 2015, Morettin e 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 obervações são determinados por uma 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 2015].

As séries estacionárias assumem que as obervações 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 2015, Cryer e Chan 2008]. Por fim, séries podem ser classificadas em determinística ou estocástica.

Quando uma série apresenta um comportamento determinístico, os valores de suas observações possuem uma estrita relação de dependência com observações passadas [Box 2015] como, por exemplo, as séries senoidais que são produzidas por oscilações repetitivas de uma onda. Séries determinísticas podem, ainda, apresentar comportamento caótico onde uma pequena perturbação no sistema pode modificar completamente o valor das observações como, por exemplo, o mapa logístico e o atrator de Lorenz, que são sistemas complexos e não-lineares [Alligood, Sauer e Yorke 1997].

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. Por exemplo, o ruído branco é caracterizado por uma sequência de variáveis aleatórias com média zero, $E(X_t) = 0$ e variância constante $var(X_t) = \sigma^2$ [Box 2015, Morettin e Toloi 2006].

A relação entre influências estocásticas e determinísticas em uma mesma série temporal tem sido abordada na literatura, principalmente com o uso de modelos do tipo ARMA (*Autoregressive Moving Average*), que consiste da combinação de dois processos: AR (*Autoregressive*) e MA (*Moving Average*). O primeiro processo consiste em modelar uma série em função de suas p observações passadas $x_t = \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + \cdots + \phi_p \cdot x_{t-p} + \varepsilon_t$, sendo ϕ_n operadores auto-regressivos e ε ruído. No processo MA, modelos são obtidos calculando q ruídos passados $x_t = x_{t-1} + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + ... + \theta_q \cdot \varepsilon_{t-q}$, onde θ_q são operadores de média móvel e ε_t ruídos [Hamilton 1994, Box e Pierce 1970, Box 2015]. O modelo ARMA consiste em modelar séries temporais estacionárias, a fim de modelar

séries não-estacionárias é utilizado o modelo ARIMA (ARMA Integrado), que é uma generalização do modelo ARMA. No modelo ARIMA, a parte integrada (I) representa uma série temporal não-estacionária, a qual é inicialmente integrada e logo é removida com intuito de tornar a série em uma série temporal estacionaria [Morettin e Toloi 2006].

Considerando esta relação, Box 2015 definem a forma aditiva de séries temporais como $X_t = T_t + S_t + E_t$, onde T_t representa a tendência, S_t representa a sazonalidade e ε_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 E como sendo um componente estocástico.

Na Figura 2.1 são mostrados os três principais componentes que podem estar em uma série temporal. A primeira série da figura representa sazonalidade, a qual foi gerada por uma função seno com frequência angular igual a $\omega = \pi$. Logo abaixo, é apresentada uma tendência positiva. A última série da Figura 2.1 foi gerada a partir de um ruído branco com média igual a 0 e desvio padrão igual 0,5.

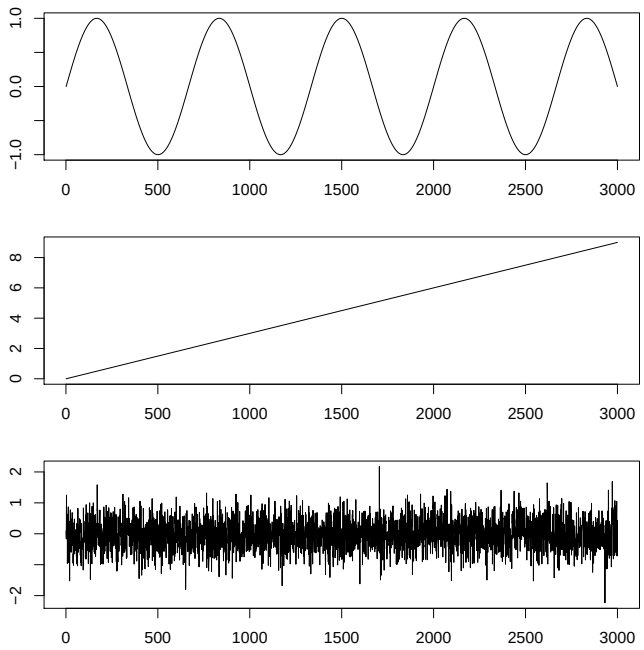
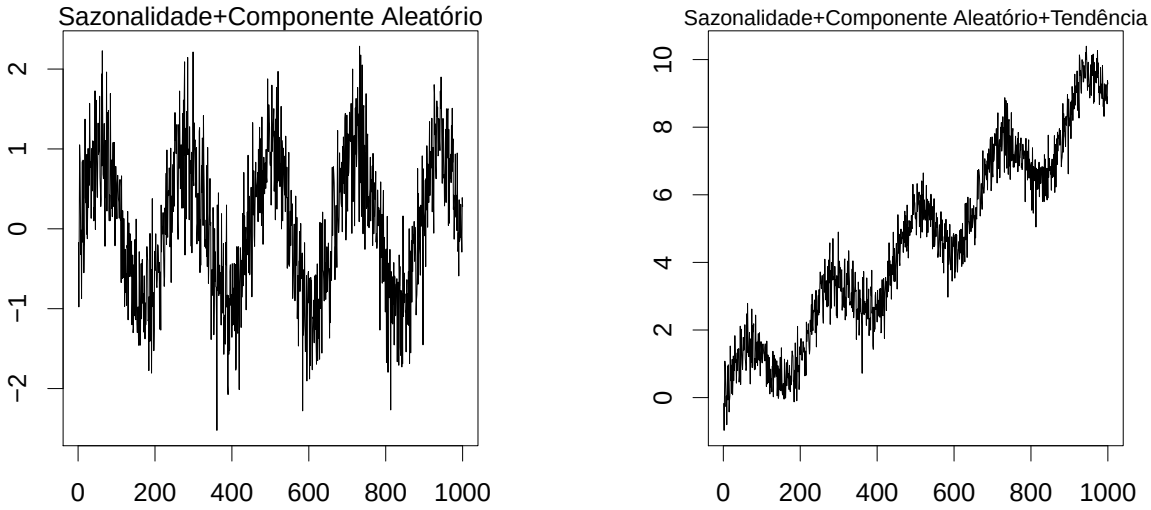


Figura 2.1: Exemplos de componentes de séries temporais: sazonalidade (figura superior), tendência (figura do meio) e componente aleatório (figura inferior).

A Figura 2.2 apresenta duas séries geradas a partir da combinação dos componentes citados acima: (a) série senoidal ruidosa estacionária ($X_t = S_t + E_T$) e (b) série senoidal ruidosa não estacionária ($X_t = T_t + S_t + E_t$).

Considerando que a proposta deste trabalho de mestrado busca agrupar séries, estimando a semelhança existente entre suas influências estocásticas e determinísticas, a próxima seção apresenta a técnica de decomposição de séries utilizada para alcançar esse objetivo.



(a) $x_t = S_t + E_t$ (b) $x_t = T_t + S_t + E_t$

Figura 2.2: Séries temporais ruidosas

2.3 DECOMPOSIÇÃO DE SÉRIES TEMPORAIS

A técnica utilizada neste trabalho para decompor séries temporais de acordo com seus componentes estocásticos e determinísticos foi baseada na abordagem proposta por Rios e Mello 2013, chamada EMD-RP.

Inicialmente, essa abordagem utiliza o método de Decomposição de Modo Empírico (EMD) (*Empirical Mode Decomposition*) [Huang 1998] para decompor uma série temporal em um conjunto de monocomponentes chamados de IMF's (*Intrinsic Mode Functions*). Esse conjunto é gerado através de um processo chamado *sifting*, que inicialmente identifica os valores máximos e mínimos da série $X(t)$. Em seguida, os valores máximos são conectados através de um método de interpolação (e.g. *cubic spline*) para compor os envelopes superior $s(t)$ e inferior $i(t)$. Em seguida, obtém-se o envelope médio com a Equação 2.1.

$$m(t) = \frac{s(t) + i(t)}{2} \tag{2.1}$$

A primeira IMF candidata é extraída usando $h_{1,1}(t) = x(t) - m(t)$. Esse processo é repetido, agora utilizando $h_{1,1}$, até que a IMF candidata satisfaça uma condição de parada, como por exemplo $m(t) = 0 \quad \forall t$. Na próxima etapa, a IMF obtida é subtraída da série original e a série resultante é utilizada novamente no processo de *sifting*. Quando nenhuma outra IMF puder ser removida, o método EMD finaliza retornando E IMFs $(h_j(t))$ mais um resíduo monotônico $(r(t))$ como apresentado na Equação 2.2.

$$X(t) = \sum_{e=1}^{E} h_e(t) + r(t) \tag{2.2}$$

Considerando que o conjunto de IMF's de uma série temporal revela diferentes informações implícitas nos dados, Rios e Mello 2013 utilizaram a ferramenta RP (*Recurrence Plot*) [Marwan 2007] com o intuito de medir o nível de determinismo de cada IMF extraído de uma série temporal.

Em resumo, RP desdobra uma série em um número maior de dimensões, chamado de espaço fase ou de coordenada de atraso, por meio da seguinte relação $X_n(m, \tau) = \{x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau}\}$. Nesta relação, m representa a dimensão embutida e τ a dimensão de separação. Os valores para essas dimensões podem ser estimados usando os métodos FNN (*False Nearest Neighbors*) e AMI (*Average Mutual Information*), respectivamente [Alligood, Sauer e Yorke 1997].

Considerando o espaço de coordenada de atraso, RP cria uma matriz de recorrência binária, cujos valores representam a proximidade entre as observações da série, considerando uma distância ϵ. Utilizando essa matriz, diversas medidas, chamadas RQA (*Recurrence Quantification Analysis*), podem ser calculadas fornecendo importantes informações da série temporal como, por exemplo, a sua taxa de determinismo (mais detalhes sobre RP e RQA são fornecidos na Seção 2.5.6).

Em resumo, a abordagem de decomposição EMD-RP calcula a taxa de determinismo para cada IMF extraída da série temporal e compara o valor obtido com um limiar (e.g. 95%) para decidir se a IMF deve ser considerada determinística ou estocástica. Por fim, as IMFs consideradas estocásticas são somadas para formar o componente estocástico e as demais são somadas para formar o componente determinístico.

A Figura 2.3 exemplifica a aplicação da abordagem EMD-RP. Inicialmente, uma série ruidosa é decomposta em um conjunto de IMFs que são combinadas para formar os componentes estocásticos e determinísticos.

Na próxima seção, são 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. Diversos problemas reais têm sido abordados com o uso de agrupamento de séries temporais conforme pode ser visto em: Medicina [Bleiweiss 2015], Aviação [Ayhan e Samet 2016] , Genética [Izakian, Pedrycz e Jamal 2015], Financeira [Durante, Pappadà e Torelli 2014] e detecção de anomalias [Esling e Agon 2012, Aghabozorgi, Shirkhorshidi e Wah 2015].

Segundo Aghabozorgi, Shirkhorshidi e Wah 2015, o agrupamento de séries 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 pode ser realizado considerando 3 diferentes tipos de objetivo de similaridade: i) tempo; ii) formato; e iii) parâmetros. Com relação ao tempo, algoritmos de agrupamento analisam séries

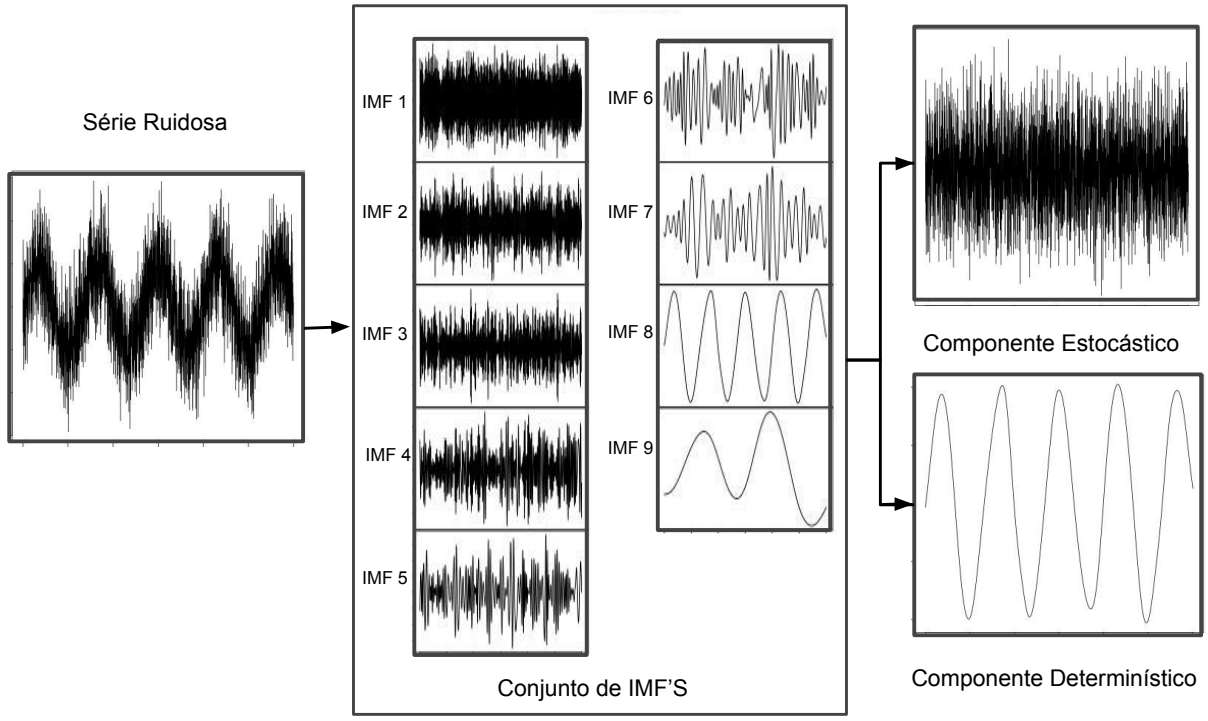


Figura 2.3: Processo de decomposição

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 parâmetros de modelos ajustados sobre séries temporais. Neste caso, algoritmos são geralmente executados sobre os parâmetros 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 trabalho de mestrado, foi estudada a categoria de agrupamento aplicada sobre bases de dados compostas por diferentes séries temporais. Com relação ao objetivo, este projeto visou 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.

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

2.4.1 Algoritmos de Agrupamentos

Algoritmos de agrupamento visam identificar, sem informação de especialistas, a forma como dados podem ser estruturados [Xu e Wunsch 2009,Faceli 2011,Bishop 2006]. Dependendo da definição utilizada para agrupamento, diferentes estruturas podem ser extraídas dos dados. Essas definições podem ser organizadas em 5 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 e Ng 2015, Liao 2005].

Os algoritmos particionais executam particionando, de maneira iterativa, os dados em um conjunto de grupos previamente definido. Exemplos de algoritmos particionais: Kmeans [MacQueen 1967], PAM (*Partitioning Around Medoids*) ou Kmedoids [Kaufman e Rousseeuw 1990] e CLARA (*Clustering Large Applications*) [Ng e Han 2002].

O agrupamento hierárquico visa organizar, de maneira aglomerativa ou divisiva, os dados em uma árvore de grupos, sendo que a raiz representa um grupo contendo todos os objetos e as folhas são grupos contendo apenas um único objeto. Métodos aglomerativos iniciam sua execução considerando cada objeto como um grupo (*bottom-up*). 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.

Durante cada etapa, dois grupos são concatenados com base em um método de vinculação. Os métodos mais usados são *single-link*, *complete-link* e *average-link*. O *single-link* concatena dois grupos que apresentam a distância mínima entre os objetos mais próximos. O *complete-link* concatena dois grupos que apresentam a distância mínima entre seus objetos mais distantes. O *average-link* concatena dois grupos que apresentam a distância mínima considerando a distância média entre todos os objetos. O algoritmo divisivo executa de maneira semelhante ao aglomerativo, porém todos os dados são, inicialmente, organizados em um único grupo (*top-down*). 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 nos nós folha da árvore. Exemplos de algoritmos hierárquicos são: BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) [Zhang, Ramakrishnan e Livny 1996] e CURE (*Clustering Using Representatives*) [Guha, Rastogi e Shim 1998].

Algoritmos baseados em grade funciona de forma similar ao algoritmo hierárquico. De maneira geral, os dados são projetados em um espaço n-dimensional o qual é dividido em células. Inicialmente, um objeto pertence a uma única célula. Nos próximos passos, células são agrupadas em subespaços, contendo informações resumidas de seus objetos. Esse passo pode ser repetido até que os dados sejam representados por um resumo geral em uma única célula [Nguyen, Woon e Ng 2015]. Exemplos de algoritmos baseado em grade são o STING (*Statistical Information Grid-clustering*) [Wang 1997] e o CLIQUE (*CLutering In QUEst*) [Agrawal 1998].

Em algoritmos baseados em densidade, grupos são formados considerando regiões de alta densidade que são separadas por regiões de baixa densidade. Com base nesta suposição, os grupos compostos por regiões de alta densidade podem ser detectados mesmo apresentando formas arbitrárias [Nguyen, Woon e Ng 2015]. Os algoritmos DBSCAN

(*Density-Based Spatial Clutering of Applications with Noise*) [Ester 1996] e DENCLUE (*Density-based Clustering*) [Hinneburg, Keim 1998] são amplamente utilizados na literatura.

Algoritmos baseados em modelos realizam uma etapa de modelagem inicialmente nos dados. Em seguida, modelos similares são organizados em um mesmo grupo [Nguyen, Woon e Ng 2015, Aghabozorgi, Shirkhorshidi e Wah 2015]. Os modelos podem ser obtidos usando diferentes técnicas como estatísticas descritivas, modelos dinâmicos ou aprendizado de máquina. Como exemplo dos algoritmos baseado em modelo é possível citar o algoritmo SOM (*Self-Organizing Maps*) [Kohonen 2001] e o COWEB [Fisher 1987].

Como pode ser observado, a aplicação dessas técnicas em séries temporais depende do uso de medidas específicas para cálculo da similaridade e/ou distância. Neste sentido, a próxima seção apresenta um conjunto de medidas desenvolvidas para calcular a similaridade e a distância entre séries temporais.

2.5 MEDIDAS DE DISTÂNCIA/SIMILARIDADE

Um dos passos mais importantes durante o processo de agrupamento é a aplicação das medidas de similaridade/distância. Usando essas medidas, é possível determinar se dois objetos devem ou não ser inseridos no mesmo grupo. As medidas utilizadas na literatura calculam a similaridade ou distância entre dois objetos que são independentes e identicamente distribuídos. No entanto, aplicar tais medidas sobre séries temporais pode produzir resultados insatisfatórios devido à dependência temporal entre as observações [Aghabozorgi, Shirkhorshidi e Wah 2015, Zhang, Ramakrishnan e Livny 1996].

No geral, as medidas de similaridade/distância precisam satisfazer algumas propriedades. A primeira propriedade consiste em garantir que os objetos não são diferentes de si próprios, ou seja, a distância $\text{Dist}(x_i, x_i) = 0 \; \forall \; x_i$. A segunda propriedade visa garantir a simetria, $\text{Dist}(x_i, x_j) = \text{Dist}(x_j, x_i)$. Por fim, é importante garantir a positividade $\text{Dist}(x_i, x_j) \geq 0 \; \forall \; x_i$ e x_j. Para uma medida ser considerada métrica é preciso satisfazer mais duas propriedades, na qual, $\text{Dist}(x_i, x_j) = 0$ somente se $x_i = x_j$ e a segunda consiste em garantir a desigualdade triangular, ou seja, $\text{Dist}(x_i, x_l) \leq \text{Dist}(x_i, x_j) + \text{Dist}(x_j, x_l) \; \forall$ x_i, x_j e x_l [Xu e Wunsch 2009].

Visando encontrar medidas de similaridade e distância que possam ser utilizadas para agrupar séries temporais, foi realizada uma revisão sistemática da literatura. Para isso, as palavras-chave "*clustering*", "*time series*", "*distance*" e "*similarity*" foram combinadas formando uma *string* de busca que foi utilizada para pesquisar por trabalhos relacionados em três diferentes repositórios: IEEE[1], ACM[2] e Scopus[3].

Como resultado dessa revisão, neste projeto foram utilizadas as seguintes medidas: i) Minkowski, ii) DTW, iii) *Cross-Correlation*, iv) CID (*Complexity-Invariant Distance*), v) MDDL e vi) DET-CRQA. As próximas seções apresentam de forma detalhada como funciona cada medida.

[1] http://ieeexplore.ieee.org/Xplore/home.jsp
[2] http://dl.acm.org/
[3] http://www.scopus.com/home.uri

2.5.1 Minkowski

Minkowski [Danielsson 1980] é uma das métricas mais utilizadas na literatura de agrupamento de dados devido à sua baixa complexidade temporal e espacial ($O(n)$) [Groenen, Mathar e Heiser 1995]. Essa medida pode ser definida pela Equação 2.3, tal que $X_t = x(1), x(2), ..., x(n)$ e $Y_t = y(1), y(2), ..., y(n)$ são séries temporais compostas por n observações.

$$D_{\text{Minkowski}}(X, Y) = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p} \tag{2.3}$$

Essa métrica pode ser especializada dependendo do valor de p nesta equação. Se $p = 1$, utiliza-se a distância de Manhattan. Por outro lado, a distância euclidiana é calculada com $p = 2$. Outra distância comumente usada para agrupar séries temporais é obtida usando $p = 3$. Na área de Mineração de Dados, essa métrica também é usada configurando $p = \infty$, referenciada como distância *supremum*, que é basicamente a maior diferença entre todas as dimensões de coordenadas.

2.5.2 DTW

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 [Berndt e Clifford 1994] foi desenvolvida executando uma etapa de alinhamento entre duas séries antes de calcular a distância entre suas observações. A DTW entre duas séries temporais X e Y pode ser calculada pelas seguintes Equações 2.4 e 2.5. É importante destacar que o alinhamento das séries, chamado de *warping path*, é realizada pela DTW, respeitando as seguintes condições: i) as primeiras e as últimas observações das séries analisadas são alinhadas; 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 Figura 2.4 apresenta o *warping path* ou caminho de deformação entre duas séries temporais ruidosas.

A complexidade temporal da DTW é $O(mn)$, podendo restringir sua aplicação quando as séries possuem grandes volumes de observações [Meesrikamolkul, Niennattrakul e Ratanamahatana 2012, Chen e Ng 2004, Esling e Agon 2012, Liao 2005, Mori, Mendiburu e Lozano 2016].

$$DTW(X, Y) = \sqrt{\text{dist}(x_n - y_m)} \tag{2.4}$$

$$\text{dist}(x_i, y_j) = (x_i, y_j)^2 + \min \begin{cases} \text{dist}(x_{i-1}, y_j) \\ \text{dist}(x_i, y_{j-1}) \\ \text{dist}(x_{i-1}, y_{j-1}) \end{cases} \tag{2.5}$$

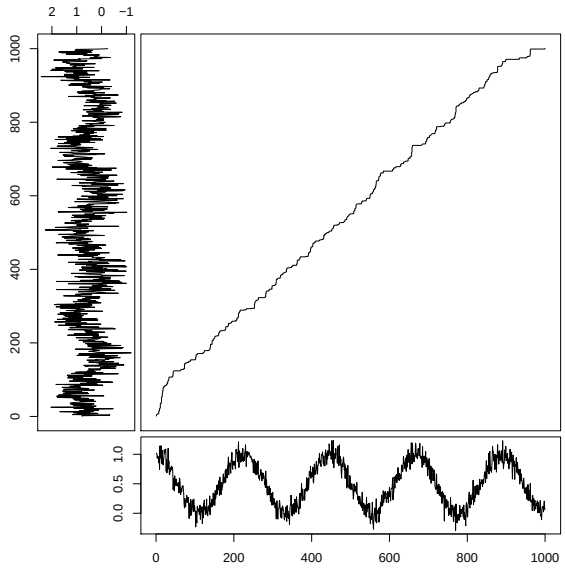


Figura 2.4: Caminho de deformação (*warping path*) entre duas séries temporais

2.5.3 Cross-Correlation

Além da distância, séries podem ser comparadas usando medidas de similaridade como a *Cross-Correlation* [Höppner e Klawonn 2009], a qual permite correlacionar duas séries mesmo quando suas observações não estão alinhadas entre si [Liao 2005, Höppner e 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 e 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 2.6. Nesta equação, $CC(X, Y, k)$ calcula a correlação cruzada entre as séries X e Y considerando um lag k.

$$\text{Dist}CC(X,Y) = \sqrt{\frac{1 - CC(X,Y,0)^2}{\sum_{k=1}^{l} 1 - CC(X,Y,k)^2}} \tag{2.6}$$

2.5.4 CID

A medida CID [Batista, Wang e Keogh 2011] foi, inicialmente, proposta para classificar dados de acordo com seus formatos. De maneira geral, essa medida calcula a distância entre duas séries 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, a 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 2.7, onde CF é o fator de correção de complexidade definido na Equação 2.8 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, Wang e Keogh 2011].

$$CID(X, Y) = D(X, Y) * CF(X, Y) \tag{2.7}$$

$$CF(X, Y) = \frac{\max(CE(X), CE(Y))}{\min(CE(X), (CE(Y))} \tag{2.8}$$

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

$$CE(X) = \sqrt{\sum_{i=1}^{n-1}(x_i - x_{i+1})^2}, \tag{2.9}$$

2.5.5 MDDL

A MDDL [Rios e Mello 2013] é uma medida de distância também desenvolvida para estimar a diferença entre duas séries temporais, levando em consideração seu comportamento geral, em vez de analisar apenas pares de observações.

Inicialmente, a MDDL calcula a DTW para as séries temporais esperadas e previstas, produzindo o *warping path*. O primeiro e último elementos do caminho de deformação estão conectados através de uma linha diagonal. Por fim, a distância MDDL é obtida calculando a área absoluta entre o *warping path* e a linha diagonal. À medida que a área aumenta, a similaridade entre as duas séries temporais reduz [Rios e Mello 2013].

Conforme apresentado por Rios e Mello 2013, MDDL varia no intervalo $[0, \int_a^b \mid f(x) \mid dx]$, onde $f(x)$ é a função MDDL usada para calcular o *warping path*, e a e b são o primeiro e último elementos, respectivamente. Assim, pode-se afirmar que MDDL varia de zero, que representa uma correspondência perfeita (ambas as séries temporais são iguais), até a área máxima formada pelo *warping path* e pela linha diagonal. Em relação a DTW, a complexidade desta medida apenas adiciona uma comparação linear entre o caminho de deformação e a linha diagonal, ambos com p observações: $O(mn) + O(p)$.

2.5.6 DET-CRQA

Por fim, a última medida apresentada nesta seção é apresentada por Marwan 2007 e foi desenvolvida levando em consideração conceitos da área de Sistemas Dinâmicos e Teoria do Caos. Esta medida é calculada primeiramente transformando uma série temporal para o espaço de fase (também referido como espaços de coordenadas de atraso de tempo) [Alligood, Sauer e Yorke 1997]. Estudos de séries no espaço fase baseiam-se no Teorema de Imersão de Takens [Takens 1981], que afirma que os atratores são melhor compreendidos

quando as séries temporais são desdobradas em um espaço de alta dimensão. Considerando esse teorema, pode-se reconstruir uma série temporal $\{x_0, x_1, \cdots, x_{n-1}\}$ no espaço fase $x_n(m, \tau) = \{x_n, x_{n+\tau}, \cdots, x_{n+(m-1)\tau}\}$, sendo m dimensão embutida e τ representa representando o atraso de tempo (ou dimensão de atraso ou dimensão de separação). A dimensão embutida basicamente define o número de eixos necessários para desdobrar uma série temporal no espaço fase. A dimensão de atraso, por outro lado, é importante para representar o comportamento sazonal da série, indicando o deslocamento necessário entre observações passadas.

Após reconstruir as séries temporais no espaço de fase, suas observações são organizadas em uma matriz binária bidimensional, denominada Matriz de Recorrência, conforme definido na Equação 2.10, na qual ε é um limiar de distância, $\|\cdot\|$ é uma norma usada para calcular a distância entre observações, e $\Theta(\cdot)$ é definido pela Equação 2.11.

$$CR_{i,j}^{\vec{x},\vec{y}}(\varepsilon) = \Theta(\varepsilon - \|\vec{x_j} - \vec{y_i}\|), \qquad i = 1, \ldots, N, \quad j = 1, \ldots, M, \tag{2.10}$$

$$\Theta(\alpha) = \begin{cases} 0, & \alpha < 0 \\ 1, & \alpha \geq 0 \end{cases} \tag{2.11}$$

Com base nessa matriz, a CRP *Cross Recurrence Plot* é gerada por pontos pretos quando $CR_{i,j} = 1$ e brancos quando $CR_{i,j} = 0$. Estruturas presentes no CRP fornecem informações importantes sobre os sistemas dinâmicos em estudo [Marwan e Kurths 2004]. Por exemplo, pontos isolados significam que estados do sistema raramente são repetidos, enfatizando um comportamento estocástico. Por outro lado, linhas diagonais ocorrem quando há comportamento persistente, mostrando que os estados do sistema são altamente recorrentes e determinísticos.

As estruturas produzidas pela CRP são analisadas pelas medidas CRQA (*Cross Recurrence Quantification Analysis*), que quantificam, por exemplo, a taxa de determinismo, a taxa de recorrência e a linha diagonal máxima.

Uma forma de calcular a similaridade entre duas séries temporais pode ser realizada estimando a taxa de determinismo (DET) conforme definida na Equação 2.12, sendo $P(\varepsilon, l)$ a frequência de linhas diagonais de comprimento l presente nas estruturas RP. Essa frequência é calculada usando a Equação 2.13.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(\varepsilon, l)}{\sum_{l=1}^{N}\sum_{l=1}^{N} R_{i,j}, \forall i \neq j} \tag{2.12}$$

$$P(\varepsilon, l) = \sum_{i,j=1}^{N} (1 - R_{i-1,j-1(\varepsilon)})(1 - R_{i+l,j-l(\varepsilon)}) \prod_{k=0}^{l-1} R_{i+k,j-k(\varepsilon)} \tag{2.13}$$

A principal vantagem do uso do DET-CRQA é a possibilidade de calcular a similaridade entre os atratores das séries temporais reconstruídas no espaço fase, o que fornece mais informações do que apenas analisá-las em uma dimensão (tempo).

Após o processo de agrupamento de séries temporais, técnicas de validação são aplicadas com o intuito de verificar se a estrutura gerada é relevante ou se foram produzidas

de maneira aleatória. Nesse sentido, a próxima seção apresenta algumas técnicas de validação de agrupamento de séries temporais.

2.6 VALIDAÇÃO DE AGRUPAMENTO

As diferentes técnicas de agrupamento disponíveis são heurísticas e fornecem uma aproximação para o resultado ideal, podendo gerar resultados diferentes a partir de um único conjunto de dados de entrada [Halkidi, Batistakis e Vazirgiannis 2001, Zeng 2002].

Verificar quantitativamente a diferença entre a aproximação e a resposta real é uma tarefa complexa porque depende do conhecimento do domínio, da aplicação e das técnicas de agrupamento utilizadas. Com isso, determinar se a estrutura obtida a partir dos algoritmos pode não ser trivial [Zeng 2002].

Considerando tais questões, existem técnicas de avaliação de agrupamento, as quais são baseadas em índices [Rendón 2011]. Um índice de validação indica a qualidade de um determinado agrupamento e seu cálculo pode ser realizado considerando diferentes funções como o erro quadrático ou a compactação existente entre elementos de um mesmo grupo [Jain e Dubes 1988, Zeng 2002].

Na validação de agrupamento, os índices podem ser classificados conforme três critérios: i) relativo; ii) interno; e iii) externo [Theodoridis e Koutroubas 1999]. Os índices relativos são usados para comparar as técnicas de agrupamento e/ou determinar a quantidade de grupos adequada considerando certos aspectos, ou seja, encontrar um agrupamento que melhor se ajuste aos dados. Dentre os índices relativos, pode-se citar a família de índices Dunn [Dunn 1974] e Silhueta [Rousseeuw 1987].

Os índices internos avaliam a qualidade de um agrupamento baseando-se apenas nos dados originais, ou seja, depende somente de recursos e informações do próprio conjunto de dados [Rendón 2011]. A Estatística Gap [Tibshirani, Walther e Hastie 2001] é comumente utilizada no processo de validação de índices internos [Aghabozorgi, Shirkhorshidi e Wah 2015].

Os índices externos calculam a qualidade de um agrupamento em função de uma estrutura referência [Rendón 2011]. Considerando que neste trabalho são utilizados dados previamente conhecidos (séries temporais sintéticas), foram utilizados os seguintes índices de validação externo: i) Rand [Rand 1971]; ii) Jaccard [Jaccard 1908]; e iii) Folkes-Mallows [Fowlkes e Mallows 1983]. Esses índices comparam a partição resultante, α, com uma partição real β.

Com base nestas partições, 4 contagens são realizadas considerando pares de objetos: i) a representa o total de pares de objetos que pertencem ao mesmo grupo em α e β; ii) b representa o total de pares de objetos colocados no mesmo grupo na partição α e em grupos separados na partição β; iii) c representa o total de pares de objetos colocados em grupos diferentes na partição α e em mesmo grupo na partição β; e iv) d representa o total de pares de objetos colocados em grupos diferentes na partição α e em β.

Com isso, os índices Rand, Jaccard e Folkes-Mallows são calculados pelas seguintes Equações 2.14, 2.15, e 2.16, respectivamente.

$$Rand(\pi^e, \pi^r) = \frac{(a+d)}{(a+b+c+d)} \tag{2.14}$$

$$Jaccard(\pi^e, \pi^r) = \frac{a}{a+b+c} \tag{2.15}$$

$$FolkesMallows(\pi^e, \pi^r) = \frac{a}{\sqrt{(a+b)\cdot(a+c)}} \tag{2.16}$$

As técnicas descritas nesta seção permitem não apenas identificar a quantidade adequada de grupos, mas também avaliar a relevância do agrupamento realizado. A próxima seção apresenta trabalhos que buscam agrupar séries considerando as influências de seus componentes estocásticos e determinísticos.

2.7 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 e Samet 2016]. Considerando que as observações meteorológicas possuem incertezas, Ayhan e Samet 2016 adotaram 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 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 2009]. No trabalho, os autores 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.

2.8 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 das séries temporais e como buscam soluções para realizar o processo de aprendizado de máquina considerando a influência de tais componentes.

Capítulo

3

ABORDAGEM PROPOSTA

3.1 CONSIDERAÇÕES INICIAIS

Este capítulo descreve como a pesquisa proposta neste mestrado foi 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 E PROPOSTA

Seja $X(t) = \{x_1, x_2, \ldots, x_t\}$ uma série temporal univariada composta por t observações. Conforme definido por Box 2015, toda observação x_i, $1 \leq i \leq t$, pode ser criada a partir de uma combinação linear de três componentes: $x_i = D_i + E_i + T_i$, tal que D_i e E_i representam os componentes determinístico e estocástico, respectivamente e T_i, representa tendência. Com isso, $X(t)$ é referida como uma séries temporal com ruído aditivo.

Este trabalho baseia-se na hipótese de que medidas de similaridade/distância proporcionam melhores resultados ao comparar diferentes séries temporais ruidosas, quando seus componentes são analisados individualmente. O primeiro passo para validar essa hipótese é usar uma abordagem para decompor as séries. Embora existam diferentes técnicas para dar suportar a esta decomposição, a análise da melhor técnica está fora do escopo desta pesquisa e, neste trabalho, foi utilizado a abordagem EMD-RP [Rios e Mello 2013], a qual foi, especificamente, projetada para decompor séries temporais em componentes estocásticos e determinísticos [1].

A análise realizada para validar a hipótese considerou um conjunto de dados $Y = \{X_\alpha(\beta)\}$, $\alpha = \{1, ..., n\}$, em que $n \in N$ representa o número de séries temporais e $\beta \in \mathbb{N}$, o número de observações na série temporal.

[1] Mais detalhes sobre abordagem podem ser vistos na Seção 2.3

Para avaliar a importância da decomposição em séries temporais antes de calcular as medidas de similaridade/distância, foi utilizado um método particional para realizar um agrupamento do tipo *hard*, o qual visa extrair uma estrutura (partição) com k grupos, $C = \{C_1, C_2, \ldots, C_k\}$, 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) $C_i \cap C_j = \emptyset$, $i, j = \{1, \ldots, k\}$ e $i \neq j$.

Para determinar se duas séries devem pertencer a um mesmo grupo, utiliza-se medidas de similaridade/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 $\mathbb{M}_{p,q} = \mathbb{D}_{p,q}$,onde $p, q \in \{1, n\}$.

Para exemplificar o uso da decomposição durante o processo de agrupamento, considere duas séries temporais $X_p(t)$ e $X_q(t)$ com o mesmo número de observações t. Considere, ainda, uma medida de distância como Manhattan, que é dada pela equação $\mathbb{D}_{p,q} = \sum_{j=1}^{t} |X_p(j) - X_q(j)|$. Neste trabalho, após realizar a decomposição das séries, as tendências são tomadas como comportamento determinístico, uma vez que sua influência em um dado instante depende apenas das tendências em instantes anteriores. Com isso, a distância de Manhattan pode ser reescrita pela Equação 3.1.

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

Portanto, pode-se afirmar que a distância[2] 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, pode-se construir uma matriz de similaridade determinística ($\mathbb{M}_D$) e outra estocástica ($\mathbb{M}_D$) que podem ser combinadas para encontrar estruturas (grupos) nos dados com maior acurácia.

A Figura 3.1 apresenta de maneira resumida a execução deste projeto. Nesta figura, duas séries temporais $X_p(t)$ e $X_q(t)$ são decompostas, extraindo as influências estocásticas (E_p e E_q) e determinísticas (D_p e D_q). Os componentes determinísticos são analisados pelas medidas apresentadas na Seção 2.5. Da mesma forma, a distância entre os componentes estocásticos também é calculada e combinada com a determinística, resultando em $\mathbb{D}_{p,q}$ que fornecerá uma distância mais precisa entre as séries ruidosas $X_p(t)$ e $X_q(t)$.

[2]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.

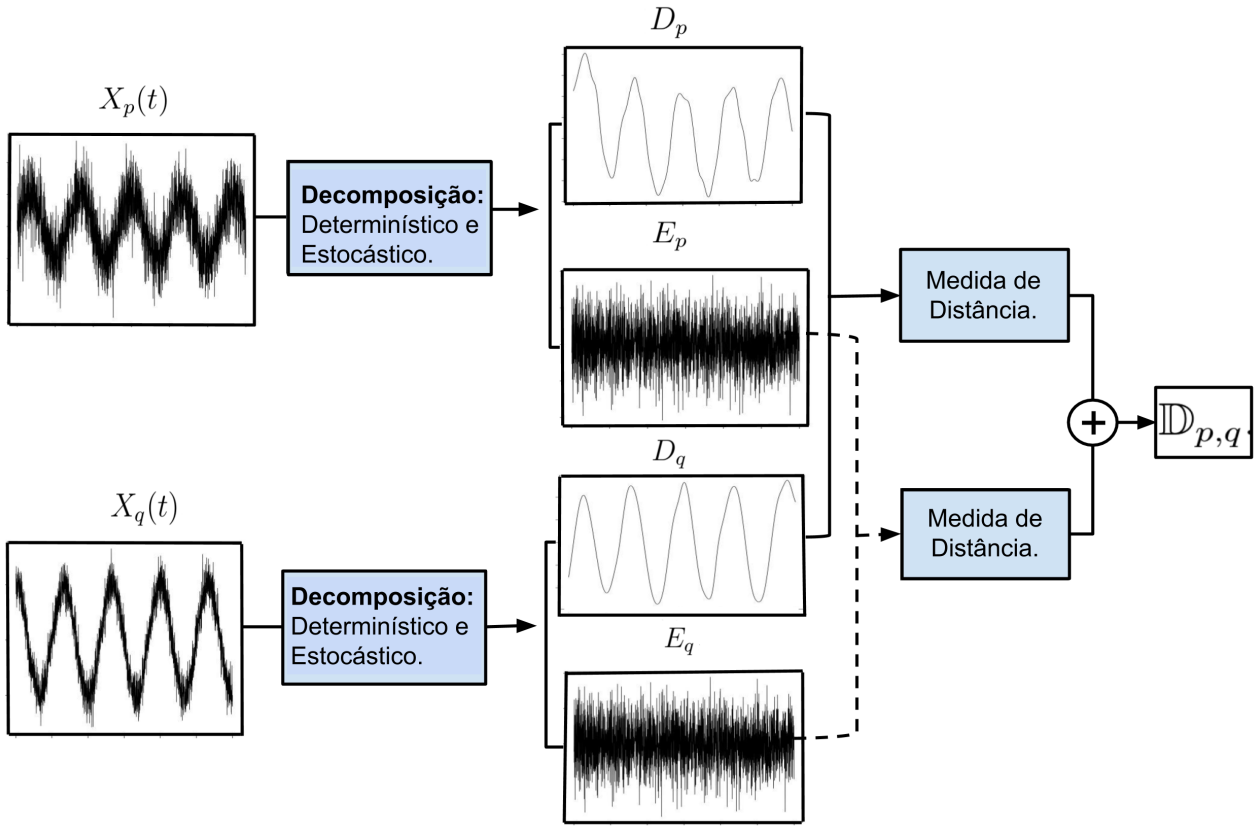


Figura 3.1: Proposta de cálculo entre séries ruidosas.

3.3 PROPOSTA DE MEDIDA PARA COMPONENTES ESTOCÁSTICOS

As medidas existentes foram desenvolvidas considerando séries temporais com comportamento determinístico, com isso, neste trabalho é proposta uma nova medida projetada para calcular a distância entre as séries temporais, cujas observações são caracterizadas apenas por processos estocásticos, ou seja, nenhum componente determinístico está presente.

A nova medida proposta para calcular a distância entre os componentes puramente estocásticos foi baseada na Transformada de Fourier (FT) (*Fourier Transform*) e na medida MDDL. A proposta considerando a FT foi baseada no trabalho apresentado por Ellis e Poliner 2007. Com a transformada de Fourier é possível converter uma série do domínio tempo para o domínio da frequência. Com isso, a nova medida é chamada de fMDDL (*Frequency-based MDDL*).

Para entender melhor essa medida, considere as séries temporais $X_p(t)$ e $X_q(t)$ compostas apenas pelo comportamento estocástico. Para cada série temporal, a medida proposta inicia o cálculo dos coeficientes de Fourier $\mathcal{F}_p = \{c_{p,1}, c_{p,2}, \ldots, c_{p,k}, \ldots, c_{p,\beta}\}$, tal que $c_{n,k}$ é calculado pela Equação 3.2. O comprimento de $\mathcal{F}_p$ é igual a série temporal original, β representa o número de coeficientes, k representa o coeficiente que está sendo

calculado.

$$c_{p,k} = \sum_{t=1}^{\beta} X_p(t) \cdot e^{i2\pi \frac{k}{\beta} t}, \forall k \in \{1, 2, \ldots, \beta\} \tag{3.2}$$

Em seguida, os coeficientes são utilizados para calcular o *power spectrum* conforme apresentado na Equação 3.3, de modo que $\mathfrak{F}(\cdot)$ e $\mathfrak{R}(\cdot)$ são as partes reais e imaginárias dos coeficientes de Fourier, respectivamente.

$$\phi_p(t) = \frac{1}{N^2}(\mathfrak{F}(\mathcal{F}_p(t))^2 + (\mathcal{R}_p(t))^2), \forall k \in \{1, 2, \ldots, N\} \tag{3.3}$$

No próximo passo, são identificados os pontos extremos e os valores máximos são selecionados ($\mathbb{M}_{max}(\cdot)$). Então, é utilizada a função de interpolação *spline* cúbica ($S^3(\cdot)$) criando um envelope superior ($\mathbb{E}_p(t)$) como mostra a Equação 3.4

$$\mathbb{E}_p(t) = S^3(\mathbb{X}_{\max}(\phi_p(t))) \tag{3.4}$$

O mesmo processo é executado na série temporal $X_q(t)$, extraindo o envelope ($\mathbb{E}_q(t)$). Finalmente, são calculadas as distâncias entre os envelopes superiores($\mathbb{E}_p(t)$) e ($\mathbb{E}_q(t)$). Neste caso, MDDL foi utilizado nesta etapa, embora ela seja recomendado para séries temporais com comportamento determinístico, uma vez que a estocasticidade presente no domínio do tempo é suavizada após a transformação das séries temporais no domínio da frequência e o cálculo do seu envelope.

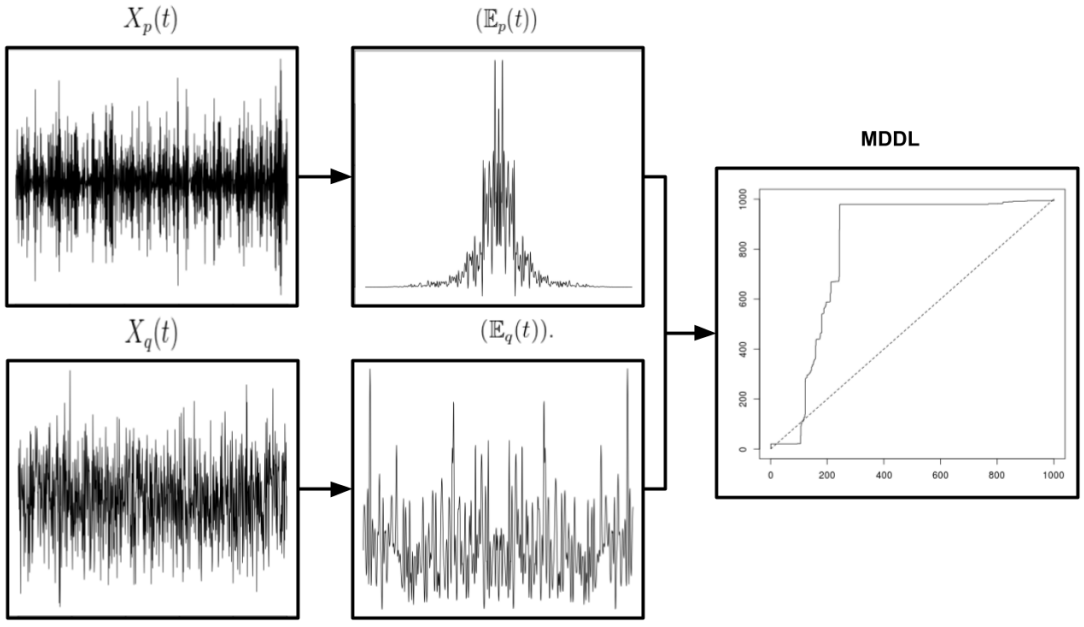


Figura 3.2: Medida fMDDL entre duas séries puramente estocásticas.

A Figura 3.2 apresenta de forma resumida a execução da fMDDL. Na figura, são calculados os coeficientes de Fourier de duas séries puramente estocásticas $X_p(t)$ e $X_q(t)$, em seguida são gerados os envelopes $(\mathbb{E}_p(t))$ e $(\mathbb{E}_q(t))$ e por fim, é calculada a MDDL entre os envelopes.

3.4 CONSIDERAÇÕES FINAIS

Esse capítulo apresentou a proposta de uma nova medida de distância para componentes estocásticos e uma nova abordagem de agrupamento que utiliza decomposição de séries temporais. O próximo capítulo apresenta os experimentos realizados com intuito de validar a hipótese deste trabalho.

Capítulo

4

EXPERIMENTOS E RESULTADOS

4.1 CONSIDERAÇÕES INICIAIS

Os experimentos apresentados neste capítulo foram realizados visando compreender a importância da decomposição como parte do processo de agrupamento de séries temporais com ruído aditivo.

Para alcançar esse objetivo, os experimentos foram organizados em quatro seções. Inicialmente, na Seção 4.2 foi avaliada a influência da decomposição nos componentes determinísticos. Os componentes estocásticos foram estudados na Seção 4.3. Na Seção 4.4, analisou-se a importância do uso combinado das medidas aplicadas nos componentes estocásticos e determinísticos. Finalmente na Seção 4.5, um processo completo de agrupamento foi realizado utilizando a abordagem proposta neste trabalho.

Com intuito de melhor compreender os experimentos, cada seção foi dividida em duas partes. A primeira consiste na configuração dos experimentos, a qual detalha como as séries foram geradas e como foi realizado o experimento [1]. A segunda parte apresenta os resultados obtidos.

4.2 ANÁLISE DOS COMPONENTES DETERMINÍSTICOS

4.2.1 Configuração dos Experimentos

Para realização dos experimentos apresentados nesta seção, foram gerados três tipos de ruído para serem adicionados às séries temporais determinísticas [Box 2015]. O primeiro ruído foi criado por um processo puramente aleatório seguindo uma distribuição de probabilidade normal com média igual a 0 e desvio padrão variando entre 0,1 e 1,0. O segundo foi criado por outro puramente aleatório seguindo uma distribuição de probabilidade uniforme com valores de intervalo entre 0,0 e 1,0. Finalmente, o último ruído foi criado através do modelo ARMA com parâmetros autorregressivos iguais a $p = 1$ e $\phi = 0{,}1$ e de médias móveis iguais a $q = 1$ e $\theta = 0{,}3$.

[1] As séries foram geradas no software R, assim como a execução dos experimentos.

Todos os parâmetros escolhidos para este experimento foram configurados com o intuito de modificar o comportamento geral do componente determinístico sem removê-lo, i.e., evitando adicionar ruídos que pudessem fazer a relação sinal-ruído [2] (Equação 4.1) se aproximar de zero, suprimindo completamente o componente determinístico. Na equação, $\sigma^2(D(t))$ e $\sigma^2(S(t))$ representam a variância dos componentes determinístico e estocástico, respectivamente.

$$SNR = \frac{\sigma^2(D(t))}{\sigma^2(S(t))} \tag{4.1}$$

As séries determinísticas foram geradas a partir das funções seno e cosseno com frequência angular igual a $\omega = \pi$. Além dessas funções, também foi utilizado o sistema Lorenz com parâmetros iguais a $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$ para produzir observações caóticas. A principal vantagem de utilizar sistemas caóticos é a complexidade de modelar suas observações, necessitando reconstruir suas observações no espaço fase [Takens 1981].

Finalmente, também foi adicionada uma tendência linear e positiva às séries temporais para entender suas influências nas medidas de distância avaliadas.

Todas as séries temporais ($x(t)$) utilizadas nos experimentos foram redimensionadas no intervalo de $[0, 1]$ conforme Equação 4.2 para evitar qualquer influência de amplitude nas observações.

$$\hat{X}(T) = \frac{X(t) - \min(X(t))}{\max(X(t)) - \min(X(t))} \tag{4.2}$$

O conjunto de dados final foi composto por séries temporais com 3000 observações, cujos os componentes estocásticos e determinísticos foram combinados conforme Tabela 4.1. A primeira coluna e a primeira linha apresentam as funções utilizadas para criar as séries temporais. As células restantes apresentam os nomes utilizados para as séries, as quais serão referenciadas no decorrer do texto. Essas séries e seus componentes estocásticos e determinísticos, obtidos após o processo de decomposição EMD, são apresentados no Apêndice A.

Tabela 4.1: Conjunto de séries compostas por componentes estocásticos e determinísticos.

	Cosseno	Cosseno+tendência	Seno	Seno+tendência	Lorenz
$\epsilon(\mu = 0, \sigma = 0,1)$	Cos1.1	Cos2.1	Sin1.1	Sin2.1	-
$\epsilon(\mu = 0, \sigma = 0,2)$	Cos1.2	Cos2.2	Sin1.2	Sin2.2	Lor1.2
$\epsilon(\mu = 0, \sigma = 0,3)$	Cos1.3	Cos2.3	Sin1.3	Sin2.3	Lor1.3
$\epsilon(\mu = 0, \sigma = 0,4)$	Cos1.4	Cos2.4	Sin1.4	Sin2.4	Lor1.4
$\epsilon(\mu = 0, \sigma = 0,5)$	Cos1.5	Cos2.5	Sin1.5	Sin2.5	Lor1.5
$\epsilon(\mu = 0, \sigma = 0,6)$	Cos1.6	Cos2.6	Sin1.6	Sin2.6	-
$\epsilon(\mu = 0, \sigma = 0,7)$	Cos1.7	Cos2.7	Sin1.7	Sin2.7	-
$\epsilon(\mu = 0, \sigma = 0,8)$	Cos1.8	Cos2.8	Sin1.8	Sin2.8	-
$\epsilon(\mu = 0, \sigma = 0,9)$	Cos1.9	Cos2.9	Sin1.9	Sin2.9	-
$\epsilon(\mu = 0, \sigma = 1,0)$	Cos1.10	Cos2.10	Sin1.10	Sin2.10	-
ARMA($p = 0.1$,$q = 0.3$)	Cos1+ARMA	Cos2+ARMA	Sin1+ARMA	Sin2+ARMA	Lor1+ARMA
$\epsilon(\min = 0, \max = 1)$	Cos1+unif	Cos2+unif	Sin1+unif	Sin2+unif	Lor1+unif

[2] Do inglês, *Signal-to-Noise Ratio*

Os resultados foram obtidos a partir de seis diferentes medidas de distância/ similaridade: i) DTW; ii) Euclidiana; iii) Manhattan; iv) Minkowski ($p = 3$); v) CID; e vi) *Cross-Correlation*; e vii) DET-CRQA. Os pares de séries temporais foram comparadas considerando duas estratégias. Na primeira, a medida é aplicada diretamente sobre as séries temporais ruidosas. Na segunda, as séries são decompostas e a medida é aplicada apenas no componente determinístico. Os resultados obtidos foram avaliados através dos testes de hipóteses *t-student* [Student 1908] e Wilcoxon-Mann-Whitney [Wilcoxon 1945] considerando um nível de significância de 5%. Para a utilização adequada do teste de hipótese foram verificadas duas premissas, a normalidade e variância dos dados. Para a normalidade foi utilizado o teste de Shapiro-Wilk [Shapiro e Wilk 1965] e para a variância, o teste de Bartlett [Bartlett 1937].

4.2.2 Resultados

Inicialmente, a medida DTW foi utilizada para calcular a distância entre duas séries puramente determinísticas criadas a partir das funções seno e cosseno com frequência angular igual a π. A distância entre as duas séries foi igual a 0,0301, apresentando um alto nível de similaridade.

Em seguida, essas séries foram analisadas após adição dos diferentes ruídos apresentados na primeira coluna da Tabela 4.1. As distâncias foram calculadas entre os pares de séries com a DTW, que são apresentados na segunda coluna da Tabela 4.2. Finalmente, a terceira coluna (Distâncias*) da Tabela 4.2 apresenta as distâncias entre os componentes determinísticos extraídos com o método de decomposição citado no Capítulo 3. É esperado que a distância entre os componentes determinísticos extraídos das séries após decomposição se aproximem do valor da distância entre as séries seno e cosseno sem adição de ruído (0,0301).

Ao analisar os resultados apresentados na tabela acima, é possível observar que quanto maior o componente estocástico, maior a distância entre as séries (segunda coluna da Tabela 4.2). Após decomposição das séries e comparação dos componentes determinísticos, as distâncias DTW estão mais próximas do resultado esperado (terceira coluna da Tabela 4.2). Com intenção de confirmar essas observações, foram aplicados os testes estatísticos, que são apresentados nas 3 últimas linhas da Tabela 4.2. Inicialmente, foi aplicado o teste Shapiro-Wilk nos resultados da segunda e terceira coluna da tabela, de acordo com os valores de p não é possível rejeitar a hipótese nula, enfatizando que há evidências de que os resultados apresentam uma distribuição normal. Em seguida, foi aplicado o teste de Bartlett, o valor de p indicou que não há evidências das variâncias serem homogêneas. Finalmente, após tais resultados, foi aplicado o teste de hipótese t-Student para variância não-homogênea, de acordo com o valor de p obtido, a hipótese nula é rejeitada, indicando que as distâncias antes e após decomposição são diferentes. Na tabela também são apresentadas as média e os valores do desvio padrão das distâncias antes e após decomposição.

No segundo experimento, foi avaliada a influência da tendência ao medir a distância entre as séries temporais. Uma vez que a EMD é capaz de detectar a tendência na série temporal (resíduo), é esperado que as medidas de similaridade/distância não sejam

Tabela 4.2: Distância DTW entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	0,0631	0,0334	(Cos2.1, Sin2.1)	0,1163	0,1420
(Cos1.2, Sin1.2)	0,0983	0,0336	(Cos2.2, Sin2.2)	0,1274	0,1341
(Cos1.3, Sin1.3)	0,1381	0,0406	(Cos2.3, Sin2.3)	0,1642	0,1063
(Cos1.4, Sin1.4)	0,1760	0,0385	(Cos2.4, Sin2.4)	0,2009	0,1483
(Cos1.5, Sin1.5)	0,2234	0,0345	(Cos2.5, Sin2.5)	0,2378	0,1268
(Cos1.6, Sin1.6)	0,2524	0,0409	(Cos2.6, Sin2.6)	0,2802	0,1272
(Cos1.7, Sin1.7)	0,2958	0,0471	(Cos2.7, Sin2.7)	0,3213	0,1066
(Cos1.8, Sin1.8)	0,3377	0,0495	(Cos2.8, Sin2.8)	0,3613	0,1697
(Cos1.9, Sin1.9)	0,3811	0,0376	(Cos2.9, Sin2.9)	0,3968	0,1042
(Cos1.10, Sin1.10)	0,4210	0,0494	(Cos2.10, Sin2.10)	0,4336	0,1403
(Cos1+ARMA, Sin1+ARMA)	0,4168	0,0559	(Cos2+ARMA, Sin2+ARMA)	0,4193	0,1228
(Cos1+unif, Sin1+unif)	0,1237	0,0314	(Cos2+unif, Sin2+unif)	0,1583	0,1434
Média	0,2439	0,04103	Média	0,2681	0,1309
Desvio Padrão	0,1265	0,0077	Desvio Padrão	0,1166	0,0195
Shapiro-Wilk	$0,4353$	$0,3332$	Shapiro-Wilk	0.2642	0.579
Bartlett	$3,01 \cdot 10^{-11}$		Bartlett	$1,179 \cdot 10^{-06}$	
t-Student	$0,0001698$		t-Student	0.001818	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

afetadas ao utilizar o processo de decomposição. Com isso, foram utilizadas as colunas 3 e 5 da Tabela 4.1, as quais apresentam as séries criadas com adição de tendência e ruído. Os resultados obtidos são apresentados nas duas últimas colunas da Tabela 4.2. Como referência, a distância entre as séries seno e cosseno puramente determinísticas com tendência é igual a $0,1495$.

Os resultados apresentados nas duas últimas colunas da Tabela 4.2 possuem um comportamento similar aos resultados do primeiro experimento. A distância aumenta à medida que a relação sinal-ruído aumenta. Por outro lado, usando a decomposição, as distâncias da DTW estão mais próximas do valor esperado (última coluna).

De maneira similar, os testes são apresentados nas 3 últimas linhas da tabela, o teste de Shapiro-Wilk foi aplicado em ambas colunas, indicando normalidade nos dados. Os resultados também apresentam variância não homogênea conforme teste de Bartlett. Por fim, o teste t-Student enfatiza diferenças entre os resultados com e sem decomposição.

Os experimentos realizados anteriormente foram repetidos alterando apenas a medida de distância utilizada. Neste novo conjunto de experimentos, a medida DTW foi substituída pela medida Euclidiana. A distância euclidiana calculada sobre as séries puramente determinísticas produzidas a partir das funções seno e cosseno é igual a 27,38. A Tabela 4.3 apresenta os resultados obtidos pelo cálculo da distância Euclidiana após a adição de diferentes ruídos nas séries temporais. Como se pode ser observado, os resultados também mostram que a abordagem de decomposição permite reduzir a influência do ruído ao calcular a distância entre duas séries. Essa conclusão é confirmada pelos testes estatísticos apresentados nas 3 últimas linhas da Tabela 4.3, no qual, o teste de

Shapiro-Wilk para a segunda e terceira coluna mostram que os resultados apresentam uma distribuição normal. O teste de Bartlett e t-Student, indicaram que a variância de ambos os resultados não são homogêneas e enfatizando a diferença após a decomposição das séries temporais.

Após adicionar tendência nas séries puramente determinísticas, a distância euclidiana não foi afetada (27,38). De forma similar, o uso da abordagem de decomposição para calcular as distâncias não foi afetado como mostram as duas últimas linhas da Tabela 4.3.

Tabela 4.3: Distância Euclidiana entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	28,56	27,24	(Cos2.1, Sin2.1)	28,54	25,14
(Cos1.2, Sin1.2)	32,04	27,75	(Cos2.2, Sin2.2)	31,48	26,55
(Cos1.3, Sin1.3)	36,24	28,23	(Cos2.3, Sin2.3)	35,58	25,72
(Cos1.4, Sin1.4)	41,03	27,51	(Cos2.4, Sin2.4)	41,61	27,21
(Cos1.5, Sin1.5)	48,14	24,87	(Cos2.5, Sin2.5)	48,24	30,85
(Cos1.6, Sin1.6)	53,48	29,18	(Cos2.6, Sin2.6)	54,74	28,60
(Cos1.7, Sin1.7)	59,69	27,01	(Cos2.7, Sin2.7)	60,60	23,41
(Cos1.8, Sin1.8)	66,93	27,65	(Cos2.8, Sin2.8)	68,74	27,41
(Cos1.9, Sin1.9)	74,61	26,84	(Cos2.9, Sin2.9)	74,60	27,56
(Cos1.10, Sin1.10)	82,13	26,09	(Cos2.10, Sin2.10)	82,33	29,18
(Cos1+ARMA, Sin1+ARMA)	87,40	27,73	(Cos2+ARMA, Sin2+ARMA)	88,49	29,38
(Cos1+unif, Sin1+unif)	34,76	27,11	(Cos2+unif, Sin2+unif)	34,88	26,80
Média	53,750	27,267	Média	54,152	27,317
Desvio Padrão	20,334	1,0721	Desvio Padrão	20,724	2,0303
Shapiro-Wilk	$0,3624$	$0,5674$	Shapiro-Wilk	$0,3736$	$0,9995$
Bartlett	$5,802 \cdot 10^{-12}$		Bartlett	$4,26 \cdot 10^{-09}$	
t-Student	$0,000881$		t-Student	$0,0009$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

A vantagem de utilizar a decomposição foi confirmada através dos testes de Shapiro-Wilk, Bartlett e *t-student*, que são apresentados na Tabela 4.3.

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 distância Manhattan não foi afetada após tendência nas séries. A Tabela 4.4 apresenta todos os resultados obtidos com a abordagem de decomposição. Na tabela, também são apresentados os resultados da distância entre as séries com tendência e os testes estatísticos.

Esses experimentos apresentaram os mesmos comportamentos, i.e., a distância aumenta conforme a influência estocástica aumenta. A vantagem de utilizar a abordagem de decomposição foi confirmada pelos testes de Shapiro-Wilk, Bartlett e *t-student*, conforme apresentado nas 3 últimas linhas da Tabela 4.4.

Os próximos resultados apresentados na Tabela 4.5, foi usada a distância de Minkowski com $p = 3$. Como apresentado no último resultado, a coluna "Distâncias*" mostra as distâncias após aplicar a abordagem de decomposição. Como referência, a distância

Tabela 4.4: Distância Manhattan entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	1380,60	1341,64	(Cos2.1, Sin2.1)	1379,57	1217,93
(Cos1.2, Sin1.2)	1497,11	1373,13	(Cos2.2, Sin2.2)	1472,45	1299,72
(Cos1.3, Sin1.3)	1641,60	1394,78	(Cos2.3, Sin2.3)	1609,70	1228,13
(Cos1.4, Sin1.4)	1817,47	1326,44	(Cos2.4, Sin2.4)	1848,52	1308,81
(Cos1.5, Sin1.5)	2115,14	1160,13	(Cos2.5, Sin2.5)	2142,42	1429,08
(Cos1.6, Sin1.6)	2366,99	1401,63	(Cos2.6, Sin2.6)	2429,22	1370,57
(Cos1.7, Sin1.7)	2602,16	1316,90	(Cos2.7, Sin2.7)	2678,07	1063,83
(Cos1.8, Sin1.8)	2943,43	1312,15	(Cos2.8, Sin2.8)	3001,11	1320,86
(Cos1.9, Sin1.9)	3256,53	1278,53	(Cos2.9, Sin2.9)	3250,85	1328,37
(Cos1.10, Sin1.10)	3629,29	1242,96	(Cos2.10, Sin2.10)	3597,55	1390,20
(Cos1+ARMA, Sin1+ARMA)	3820,359	1330,64	(Cos2+ARMA, Sin2+ARMA)	3873,68	1368,60
(Cos1+unif, Sin1+unif)	1570,136	1337,66	(Cos2+unif, Sin2+unif)	1569,51	1312,50
Média	2386,7	1318,0	Média	2404,3	1303,2
Desvio Padrão	860,02	66,913	Desvio Padrão	871,54	97,005
Shapiro-Wilk	$0,2412$	$0,2085$	Shapiro-Wilk	$0,2877$	$0,1206$
Bartlett	$3,678 \cdot 10^{-10}$		Bartlett	$1,645 \cdot 10^{-08}$	
t-Student	$0,0012$		t-Student	$0,00109$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

entre as séries puramente determinísticas é igual a $7,663852$ e não apresentou diferença após adicionar tendência.

Ao utilizar Minkowski é possível observar o mesmo comportamento dos experimentos anteriores. A vantagem de utilizar a abordagem de decomposição foi confirmada pelos testes de Shapiro-Wilk, Bartlett e *t-student*, conforme apresentado nas 3 últimas linhas da Tabela 4.5.

No próximo resultado, Tabela 4.6, utilizou-se a medida CID. A CID entre duas séries puramente determinísticas foi igual a $27,38615$. Após adicionar tendência, o valor da distância foi similar: $28,8593$. Ao considerar os resultados apresentados nas colunas "Distâncias*", observa-se a importância do uso da decomposição. Essa observação foi confirmada através da análise nos testes estatísticos apresentados nas últimas linhas desta tabela. Para o experimento sem tendência, o teste de Shapiro-Wilk rejeitou a hipótese nula e não existe evidências de que os resultados usando a abordagem de decomposição sejam de uma população normalmente distribuída. Neste caso, foi utilizado o teste de Wilcoxon-Mann-Whitney para mostrar a diferença entre os resultados com e sem decomposição.

A próxima medida utilizada no componente determinístico foi a taxa de determinismo calculada pela CRQA. Após ajustar os parâmetros para $\epsilon = 10$, $d = 2$ e $m = 1$, sendo raio, dimensão de separação e dimensão embutida, respectivamente, a taxa de determinismo entre duas séries temporais puramente determinísticas foi igual a 99,97. Neste caso, é esperado que o determinismo possa reduzir à medida que mais componente estocástico é adicionado na série temporal. Esta medida não foi afetada pela tendência, proporcio-

Tabela 4.5: Distância Minkowski entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	8,14	7,64	(Cos2.1, Sin2.1)	8,13	7,17
(Cos1.2, Sin1.2)	9,38	7,76	(Cos2.2, Sin2.2)	9,22	7,47
(Cos1.3, Sin1.3)	10,84	7,90	(Cos2.3, Sin2.3)	10,68	7,42
(Cos1.4, Sin1.4)	12,48	7,81	(Cos2.4, Sin2.4)	12,63	7,76
(Cos1.5, Sin1.5)	14,73	7,21	(Cos2.5, Sin2.5)	14,64	9,07
(Cos1.6, Sin1.6)	16,30	8,33	(Cos2.6, Sin2.6)	16,70	8,17
(Cos1.7, Sin1.7)	18,40	7,62	(Cos2.7, Sin2.7)	18,52	6,82
(Cos1.8, Sin1.8)	20,53	7,93	(Cos2.8, Sin2.8)	21,15	7,82
(Cos1.9, Sin1.9)	22,95	7,69	(Cos2.9, Sin2.9)	22,92	7,84
(Cos1.10, Sin1.10)	25,10	7,53	(Cos2.10, Sin2.10)	25,33	8,35
(Cos1+ARMA, Sin1+ARMA)	26,78	7,97	(Cos2+ARMA, Sin2+ARMA)	27,22	8,63
(Cos1+unif, Sin1+unif)	10,38	7,61	(Cos2+unif, Sin2+unif)	10,43	7,55
Média	16,334	7,75	Média	16,464	7,8391
Desvio Padrão	6,3972	0,2756	Desvio Padrão	6,5507	0,6324
Shapiro-Wilk	$0,4346$	$0,8195$	Shapiro-Wilk	$0,4355$	$0,9788$
Bartlett	$6,713 \cdot 10^{-13}$		Bartlett	$3,646 \cdot 10^{-09}$	
t-Student	$0,0007$		t-Student	$0,0008$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

Tabela 4.6: Distância CID entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	28,86	27,83	(Cos2.1, Sin2.1)	28,91	27,14
(Cos1.2, Sin1.2)	33,01	27,92	(Cos2.2, Sin2.2)	31,82	28,05
(Cos1.3, Sin1.3)	36,64	28,36	(Cos2.3, Sin2.3)	36,73	29,05
(Cos1.4, Sin1.4)	41,07	34,30	(Cos2.4, Sin2.4)	42,53	35,78
(Cos1.5, Sin1.5)	48,21	27,80	(Cos2.5, Sin2.5)	48,82	33,50
(Cos1.6, Sin1.6)	54,35	35,18	(Cos2.6, Sin2.6)	56,56	36,49
(Cos1.7, Sin1.7)	61,72	27,39	(Cos2.7, Sin2.7)	61,01	24,30
(Cos1.8, Sin1.8)	67,01	28,80	(Cos2.8, Sin2.8)	69,87	28,68
(Cos1.9, Sin1.9)	77,95	32,64	(Cos2.9, Sin2.9)	75,93	28,45
(Cos1.10, Sin1.10)	83,45	28,54	(Cos2.10, Sin2.10)	85,16	29,56
(Cos1+ARMA, Sin1+ARMA)	91,23	29,22	(Cos2+ARMA, Sin2+ARMA)	89,26	40,25
(Cos1+unif, Sin1+unif)	35,76	27,50	(Cos2+unif, Sin2+unif)	35,25	31,73
Média	54,938	29,623	Média	55,154	31,081
Desvio Padrão	21,219	2,7699	Desvio Padrão	21,141	4,5876
Shapiro-Wilk	$0,3426$	$0,0028$	Shapiro-Wilk	$0,354$	$0,5013$
Wilcoxon-Mann-Whitney	$2,219 \cdot 10^{-05}$		Bartlett	$1,669 \cdot 10^{-05}$	
			t-Student	$9,648 \cdot 10^{-09}$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

nando um resultado semelhante: 99,99. Os resultados com a série temporal ruidosa são apresentados na Tabela 4.7.

É observado que os resultados apresentados nas colunas "Distância *" mostram que a distância entre as duas séries se manteve estável após decomposição, como mostra a média. Para o experimento sem tendência, o teste de Shapiro-Wilk rejeitou a hipótese nula e não existe evidências de que os resultados usando a abordagem de decomposição sejam de uma população com distribuição normal. Com isso, utilizou-se o teste de Wilcoxon-Mann-Whitney para estudar a diferença entre os resultados com e sem decomposição.

Tabela 4.7: Distância DET-CRQA entre as séries cosseno e seno com ruído aditivo sem/-com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	37,39	99,99	(Cos2.1, Sin2.1)	79,96	99,98
(Cos1.2, Sin1.2)	10,44	99,88	(Cos2.2, Sin2.2)	45,64	99,98
(Cos1.3, Sin1.3)	11,41	99,95	(Cos2.3, Sin2.3)	93,50	99,97
(Cos1.4, Sin1.4)	8,03	99,98	(Cos2.4, Sin2.4)	39,63	99,95
(Cos1.5, Sin1.5)	6,59	99,98	(Cos2.5, Sin2.5)	43,89	99,98
(Cos1.6, Sin1.6)	9,59	99,99	(Cos2.6, Sin2.6)	21,13	99,96
(Cos1.7, Sin1.7)	6,01	99,98	(Cos2.7, Sin2.7)	28,79	99,98
(Cos1.8, Sin1.8)	10,40	99,99	(Cos2.8, Sin2.8)	31,53	99,97
(Cos1.9, Sin1.9)	6,15	99,98	(Cos2.9, Sin2.9)	15,04	99,97
(Cos1.10, Sin1.10)	5,71	99,72	(Cos2.10, Sin2.10)	24,87	99,93
(Cos1+ARMA, Sin1+ARMA)	6,80	99,81	(Cos2+ARMA, Sin2+ARMA)	23,25	99,94
(Cos1+unif, SinRF)	7,58	99,81	(Cos2+unif, Sin2+unif)	63,65	99,96
Média	10,508	99,921	Média	42,573	99,964
Desvio Padrão	8,6884	0,0933	Desvio Padrão	24,628	0,0167
Shapiro-Wilk	$3.08 \cdot 10^{-5}$	$0,002876$	Shapiro-Wilk	$0,1082$	$0,0622$
Wilcoxon-Mann-Whitney	$3,45 \cdot 10^{-5}$		Bartlett	$2,2 \cdot 10^{-16}$	
			t-Student	$5,99 \cdot 10^{-06}$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

A última medida utilizada nestes experimentos foi a similaridade *Cross-Correlation* e os experimentos foram realizados seguindo os mesmos passos apresentados anteriormente. Utilizando essa medida, a distância entre as duas séries temporais puramente determinísticas sem e com tendência foi de 0,0449 e 0,0164, respectivamente. Os resultados apresentados na Tabela 4.8 mostram que a distância entre duas séries apresentou um comportamento estável mais próximo do esperado quando o processo de decomposição foi considerado. Essa observação foi confirmada pela análise dos testes estatísticos apresentados nas últimas linhas desta tabela. De acordo com os resultados apresentados na tabela, ambos os experimentos com e sem tendência apresentaram valores de p inferiores a 0,05 para o teste de normalidade pelo teste de Shapiro-Wilk. Portanto, não seria possível usar o teste t-Student para avaliar a diferença entre os resultados com e sem decomposição. Assim, essa diferença foi enfatizada pelo uso do teste de Mann-Whitney-Wilcoxon.

Após considerar todos os experimentos apresentados nesta seção, é possível observar que a DTW e DET-CRQA apresentaram um comportamento mais estável comparando

Tabela 4.8: Distância *Cross-Correlation* entre as séries cosseno e seno com ruído aditivo sem/com tendência.

Séries (sem tendência)	Distâncias	Distâncias*	Séries (com tendência)	Distâncias	Distâncias*
(Cos1.1, Sin1.1)	0,04585	0,04504	(Cos2.1, Sin2.1)	0,01664	0,01646
(Cos1.2, Sin1.2)	0,04832	0,04520	(Cos2.2, Sin2.2)	0,01700	0,01644
(Cos1.3, Sin1.3)	0,05328	0,04519	(Cos2.3, Sin2.3)	0,01784	0,01653
(Cos1.4, Sin1.4)	0,05927	0,04563	(Cos2.4, Sin2.4)	0,01903	0,01645
(Cos1.5, Sin1.5)	0,06736	0,04555	(Cos2.5, Sin2.5)	0,02030	0,01643
(Cos1.6, Sin1.6)	0,07766	0,04571	(Cos2.6, Sin2.6)	0,02181	0,01670
(Cos1.7, Sin1.7)	0,08760	0,04625	(Cos2.7, Sin2.7)	0,02338	0,01699
(Cos1.8, Sin1.8)	0,10151	0,04537	(Cos2.8, Sin2.8)	0,02550	0,01645
(Cos1.9, Sin1.9)	0,11266	0,04424	(Cos2.9, Sin2.9)	0,02708	0,01688
(Cos1.10, Sin1.10)	0,12674	0,04817	(Cos2.10, Sin2.10)	0,02945	0,01701
(Cos1+ARMA, Sin1+ARMA)	0,14376	0,04470	(Cos2+ARMA, Sin2+ARMA)	0,03117	0,01762
(Cos+unif, Sin1+unif)	0,05207	0,04483	(Cos2+unif, Sin2+unif)	0,01782	0,01671
Média	0,0813	0,0455	Média	0,0222	0,0167
Desvio Padrão	0,0331	0,0009	Desvio Padrão	0,0050	0,0003
Shapiro-Wilk	$0,1793$	$0,03724$	Shapiro-Wilk	$0,2106$	$0,011$
Mann-Whitney-Wilcoxon	$2,95 \cdot 10^{-06}$		Mann-Whitney-Wilcoxon	$0,0002454$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

duas séries temporais. DTW tem a vantagem de procurar o melhor alinhamento entre duas séries antes de calcular suas distâncias. Por outro lado, DET-CRQA tem a vantagem de analisar a distância entre duas séries, desdobrando suas observações no espaço de fase.

Com base nesses resultados, foi analisado o comportamento dessas duas medidas em séries temporais caóticas criadas pelo sistema de Lorenz. Neste caso, foi considerada a série temporal de Lorenz apresentada na Tabela 4.1. Embora tenham os mesmos nomes na Tabela 4.9, foram produzidos diferentes ruídos para assegurar estão sendo comparadas diferentes séries temporais.

Os resultados apresentados nesta tabela enfatizam o comportamento esperado. Uma vez que o sistema de Lorenz produz observações caóticas, DTW não consegue encontrar um bom alinhamento em uma dimensão, mesmo usando um processo de decomposição. A CRQA, por sua vez, permite desdobrar as séries temporais em seu espaço de fase antes de comparar suas distâncias. Como consequência, após a decomposição das séries temporais, o componente determinístico resultante está mais próximo da série temporal de Lorenz e o nível determinístico está mais próximo do esperado.

Na seção seguinte, são apresentados os experimentos avaliando apenas o componente estocástico.

Tabela 4.9: Distância DTW e CRQA entre séries Lorenz com ruído aditivo.

Séries	DTW	DTW(Decomposição)	Séries	CRQA	CRQA(Decomposição)
(Lor1.2, Lor1.2)	0,2128	0,1078	(Lor1.2, Lor1.2)	93,53	98,13
(Lor1.3, Lor1.3)	0,3056	0,3243	(Lor1.3, Lor1.3)	90,15	98,03
(Lor1.4, Lor1.4)	0,3734	0,3024	(Lor1.4, Lor1.4)	86,41	97,79
(Lor1.5, Lor1.5)	0,4155	0,5635	(Lor1.5, Lor1.5)	81,24	97,78
(Lor1+ARMA, Lor1+ARMA)	0,6901	1,032	(Lor1+ARMA, Lor1+ARMA)	63,22	97,11
(Lor1+unif, Lor1+unif)	0,2910	1,042	(Lor1+unif, Lor1+unif)	90,10	98,12
Média	0,3814	0,562	Média	84,108	97,826
Desvio Padrão	0,1666	0,3953	Desvio Padrão	11,0545	0,3838
Shapiro-Wilk	$0,231$	$0,2438$	Shapiro-Wilk	$0,07982$	$0,07235$
Bartlett	0.0813		Bartlett	$8,42 \cdot 10^{-07}$	
t-Student	0.3268		t-Student	$0,02873$	

Distâncias: Apresenta as distâncias entre as séries ruidosas.
Distâncias*: Apresenta as distâncias entre os componentes determinísticos extraídos das séries ruidosas.

4.3 ANÁLISE DOS COMPONENTES ESTOCÁSTICOS

4.3.1 Configuração dos Experimentos

Nesta seção é apresentado um conjunto de experimentos realizados considerando séries temporais produzidas a partir de processos estocásticos. Nesse caso, devido à ausência da influência determinística, todas as medidas usadas anteriormente podem não fornecer bons resultados. No entanto, optou-se por utilizar DTW, a qual pode procurar um alinhamento ideal entre duas séries temporais.

Portanto, visando calcular a distância/semelhança entre duas séries temporais estocásticas, neste experimento foram consideradas três medidas previamente apresentadas nas Seções 2.5 e 3.3: i) DTW; ii) MDDL; e iii) fMDDL. Além disso, também foi utilizada a medida fCOR que calcula a correlação entre os espectrogramas calculados a partir de duas séries temporais analisadas. Essa última medida foi utilizada para mostrar que uma simples correlação entre duas séries temporais no domínio de frequências não é suficiente para detectar semelhanças.

O conjunto de dados para este experimento foi criado usando quatro tipos diferentes de processos estocásticos. Primeiramente, foi utilizado um processo de ruído branco com $\mu = 0$ e $\sigma = 0{,}1$ para criar duas séries temporais denominadas WN1 e WN2. Ainda usando este processo, duas séries temporais extras chamadas WN3 e WN4 foram criadas usando $\mu = 0$ e $\sigma = 1{,}0$. Em seguida, as séries temporais referidas como AR1 e AR2 foram criadas considerando um processo autorregressivo com $p = 1$ e $\phi = 0{,}2$. Também foram criadas as séries temporais MA1 e MA2 considerando um processo de média móvel com $q = 1$ e $\theta = -0{,}7$. Finalmente, foi utilizado um processo de média móvel autorregressivo para criar quatro séries temporais usando a seguinte configuração: ARMA1 e ARMA2 com $p = 1$, $q = 1$, $\phi = 0{,}1$, $\theta = 0{,}9$, e ARMA3 e ARMA4 com $p = 1$, $q = 1$, $\phi = -0{,}9$, $\theta = 0{,}1$. Todas as séries temporais foram criadas com 1000 observações. Neste experimento, nenhuma decomposição foi necessária, uma vez que o intuito é avaliar

as medidas apenas analisando o comportamento estocástico.

4.3.2 Resultados

Inicialmente, foram analisadas apenas as séries temporais produzidas pelo processo de ruído branco como mostrado na Tabela 4.10. Basicamente, as medidas foram avaliadas pela sua capacidade de discriminar WN1 e WN2 de WN3 e WN4. Nesta tabela, são apresentadas as matrizes triangulares superiores, uma vez que as medidas são simétricas.

Como é possível notar, a fMDDL forneceu os melhores resultados apresentando valores mais baixos comparando WN1 com WN2 e WN3 com WN4, e valores mais altos comparando séries criadas usando desvio padrão diferente. Esse é um resultado interessante, uma vez que todas as séries foram criadas pelo mesmo processo, tornando a discriminação entre seus parâmetros mais complicada.

Tabela 4.10: Distância entre séries geradas com ruído branco.

		WN1	WN2	WN3	WN4
DTW	WN1	0,00	0,40	0,41	0,40
	WN2		0,00	0,41	0,41
	WN3			0,00	0.41
	WN4				0,00
		WN1	WN2	WN3	WN4
MDDL	WN1	0,00	15,88	16,63	14,59
	WN2		0,00	11,15	18,63
	WN3			0,00	15,78
	WN4				0,00
		WN1	WN2	WN3	WN4
fMDDL	WN1	0,00	14,21	77,53	74,43
	WN2		0,00	24,32	37,41
	WN3			0,00	13,29
	WN4				0,00
		WN1	WN2	WN3	WN4
fCOR	WN1	1,00	0,01	0,04	-0,09
	WN2		1,00	-0,04	0,07
	WN3			1,00	-0,08
	WN4				1,00

No segundo experimento, foi analisado se as medidas são capazes de diferenciar os processos autorregressivo e de média móvel. De acordo com os resultados apresentados na Tabela 4.11, a fMDDL foi a medida que apresentou maior diferença ao comparar as distâncias entre as séries dos diferentes processos.

Com base em resultados anteriores, as medidas foram aplicadas em séries temporais produzidas pela combinação de processos autorregressivos e de média móvel. De acordo com as distâncias apresentadas na Tabela 4.12, a fMDDL também apresentou os melhores resultados para este experimento. Séries criadas com os mesmos parâmetros foram consideradas mais próximas, uma vez que quanto mais próximas as séries, menor o resultado.

Tabela 4.11: Distância entre séries geradas com processo autorregressivo e média móvel.

		AR1	AR2	MA1	MA2
DTW	AR1	0,00	0,38	0,43	0,43
	AR2		0,00	0,43	0,43
	MA1			0,00	0,43
	MA2				0,00
		AR1	AR2	MA1	MA2
MDDL	AR1	0,00	14,38	33,02	14,20
	AR2		0,00	24,6	20,78
	MA1			0,00	50,35
	MA2				0,00
		AR1	AR2	MA1	MA2
fMDDL	AR1	0,00	14,12	123,53	89,18
	AR2		0,00	78,09	104,63
	MA1			0,00	15,64
	MA2				0,00
		AR1	AR2	MA1	MA2
fCOR	AR1	1,00	0,28	-0,24	-0,35
	AR2		1,00	-0,35	-0,32
	MA1			1,00	0,38
	MA2				1,00

Tabela 4.12: Distância entre séries geradas com processo ARMA.

		ARMA1	ARMA2	ARMA3	ARMA4
DTW	ARMA1	0,00	0,34	0,48	0,49
	ARMA2		0,00	0,48	0,49
	ARMA3			0,00	0,40
	ARMA4				0,00
		ARMA1	ARMA2	ARMA3	ARMA4
MDDL	ARMA1	0,00	18,48	10,14	52,53
	ARMA2		0,00	13,00	26,86
	ARMA3			0,00	25,86
	ARMA4				0,00
		ARMA1	ARMA2	ARMA3	ARMA4
fMDDL	ARMA1	0,00	14,49	138,63	159,54
	ARMA2		0,00	147,13	147,22
	ARMA3			0,00	10,63
	ARMA4				0,00
		ARMA1	ARMA2	ARMA3	ARMA4
fCOR	ARMA1	1,00	0,64	-0,24	-0,24
	ARMA2		1,00	-0,25	-0,24
	ARMA3			1,00	0,78
	ARMA4				1,00

Finalmente, como último experimento, foram comparadas séries temporais criadas a partir de processos autorregressivo de média móvel e ruído branco. No geral, as outras medidas, além da fMDDL, deveriam apresentar bons resultados, uma vez que os processos são completamente diferentes entre si.

Entretanto, mesmo considerando a diferença significativa entre os dois processos originais, a fMDDL foi a única medida que proporcionou boa discriminação entre as séries temporais, como é mostrado na Tabela 4.13.

Tabela 4.13: Distância entre séries geradas com ruído branco e processo ARMA.

		WN1	WN2	ARMA1	ARMA2
DTW	WN1	0,00	0,40	0,39	0,40
	WN2		0,00	0,40	0,40
	ARMA1			0,00	0,37
	ARMA2				0,00
		WN1	WN2	ARMA1	ARMA2
MDDL	WN1	0,00	15,88	7,37	12,59
	WN2		0,00	11,71	17,85
	ARMA1			0,00	33,68
	ARMA2				0,00
		WN1	WN2	ARMA1	ARMA2
fMDDL	WN1	0,00	13,38	68,43	51,42
	WN2		0,00	43,60	71,72
	ARMA1			0,00	16,44
	ARMA2				0,00
		WN1	WN2	ARMA1	ARMA2
fCOR	WN1	1,00	0,04	0,07	0,06
	WN2		1,00	0,06	0,08
	ARMA1			1,00	0,23
	ARMA2				1,00

Depois de analisar as principais medidas usadas para calcular a distância/semelhança comparando individualmente os componentes determinísticos e estocásticos, na próxima seção é apresentado um conjunto de experimentos combinando as melhores medidas para cada componente para melhor discriminar séries temporais com ruído aditivo.

4.4 ANÁLISE DE SÉRIES TEMPORAIS DETERMINÍSTICAS COM RUÍDO ADITIVO

4.4.1 Configuração dos Experimentos

Como mencionado anteriormente, os experimentos realizados nesta seção visaram combinar medidas estocásticas e determinísticas para calcular a distância entre duas séries temporais com ruído aditivo. Com base nos experimentos apresentados anteriormente, foram selecionadas a DTW e DET-CRQA para comparar os componentes determinísticos e fMDDL para calcular a distância entre os estocásticos.

Ao combinar a fMDDL com a DET-CRQA, a taxa de determinismo foi normalizada calculando $\overline{CRQA} = 1 - \frac{DET}{100}$ para mantê-las na mesma escala.

Os experimentos foram realizados em séries temporais criadas após a adição de diferentes ruídos nas observações determinísticas geradas usando a função seno e o sistema de Lorenz. Para a função seno, foi utilizada a frequência angular para $\omega = \pi$, enquanto que os parâmetros para o sistema Lorenz foram ajustados para $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$

para produzir observações caóticas.

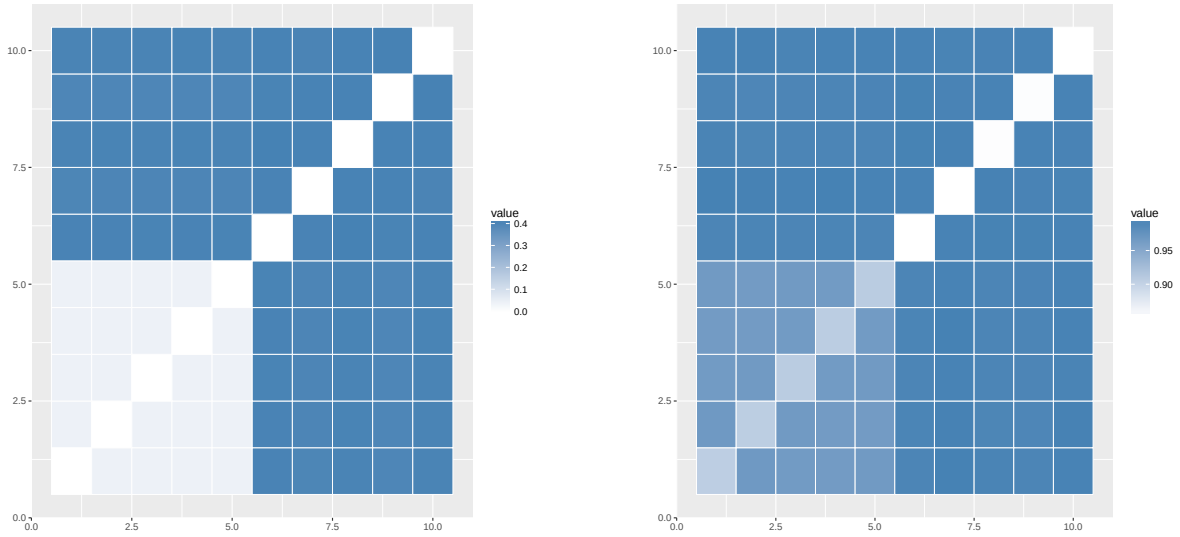
Para os componentes estocásticos, foram consideradas quatro funções diferentes: i) ruído branco (WN) com desvio padrão $\sigma = \{0{,}1; 0{,}2; 0{,}5; 1{,}0\}$; ii) processo autorregressivo (AR) com $p = 1$ e $\phi = 0,2$; iii) processo de média móvel (MA) com $q = 1$ e $\theta = -0{,}7$; e iv) processo autorregressivo de média móvel (ARMA) com $p = 1$, $q = 1$, $\phi = \{-0{,}7; 0{,}3\}$ e $\theta = \{-0{,}9; 0{,}9\}$. Foram geradas 3000 observações para cada componente determinístico e estocástico, as quais foram posteriormente utilizadas para criar as séries temporais com ruído aditivo.

Finalmente, é importante ressaltar que todas as séries temporais foram normalizadas para o intervalo de $[0; 1]$ e as distâncias calculadas entre pares de séries temporais foram representadas com mapas de calor, sendo que quanto mais próximo de 0 a distância, mais clara será a cor utilizada.

4.4.2 Resultados

Inicialmente, foram gerados 2 grupos diferentes de séries temporais ruidosas. O primeiro grupo foi criado com 5 séries senoidais com adição de 5 séries geradas a partir de ruído branco com desvio padrão igual a 0,1. Para o segundo grupo foram utilizadas as mesmas séries, porém com desvio padrão igual a 1,0.

Inicialmente, as séries temporais foram calculadas sem decomposição utilizando as distâncias CRQA e DTW. Como mostra as Figuras 4.1(a) e (b), as medidas retornaram pequenas distâncias entre as séries temporais do primeiro grupo, mas não foram capazes de discriminar o segundo, uma vez que as distâncias entre as séries temporais nesse grupo apresentaram valores elevados devido a influência do ruído ($\sigma = 1{,}0$).

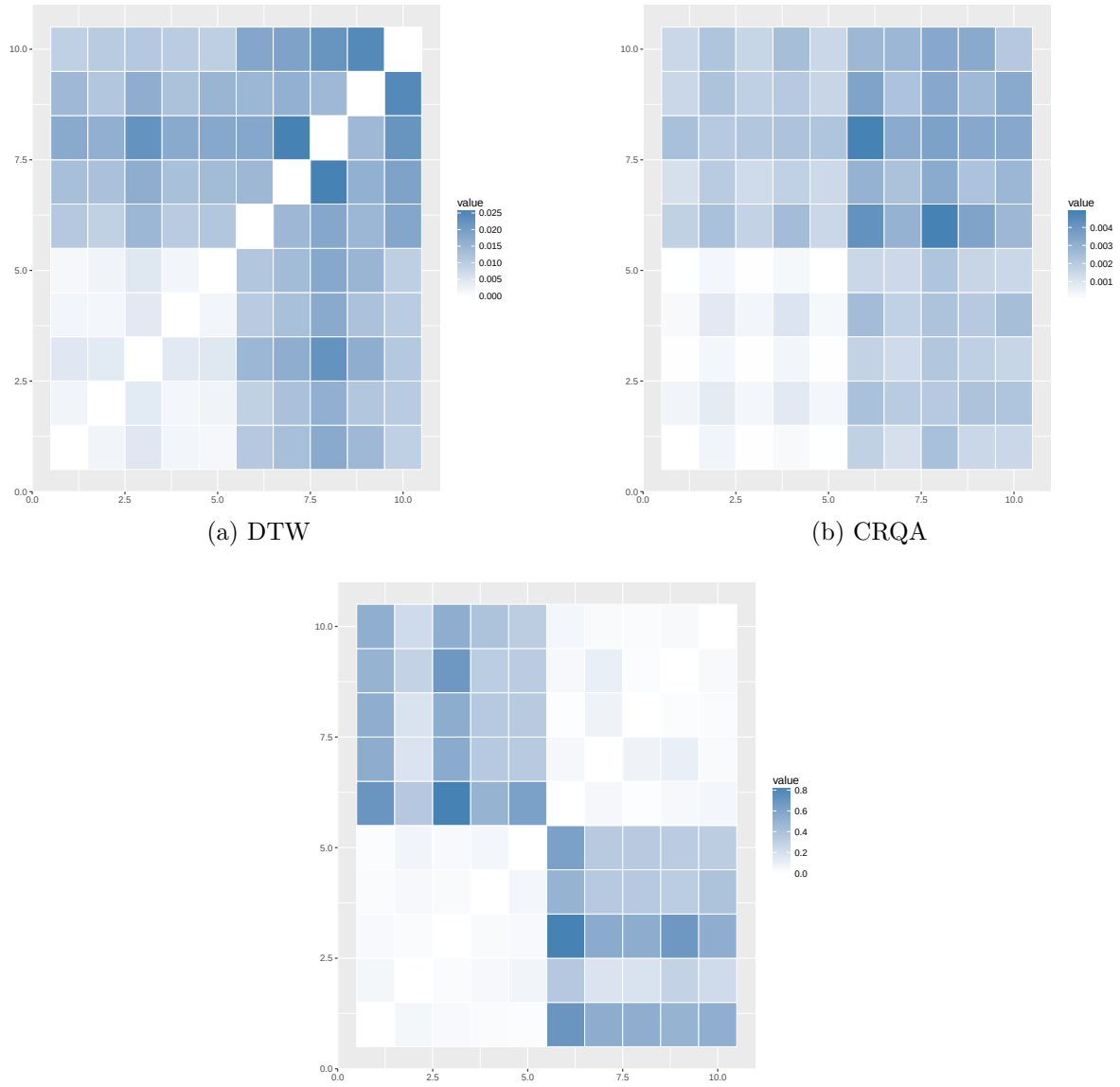


(a) DTW (b) CRQA

Figura 4.1: Distâncias DTW e CRQA entre as séries ruidosas (sin +WN).

Após a decomposição das séries temporais, as distâncias forneceram resultados con-

forme o esperado, como mostrado na Figura 4.2. Os componentes determinísticos (Figuras 4.2(a) e (b)) não apresentam discriminação relevante uma vez que foi utilizada a mesma função determinística (senoide com ($\sigma = \pi$) e todos os valores de distância foram próximos de 0. No entanto, o componente estocástico (Figura 4.2(c)), apresenta claramente pequenas distâncias de fMDDL entre as séries temporais.



(a) DTW

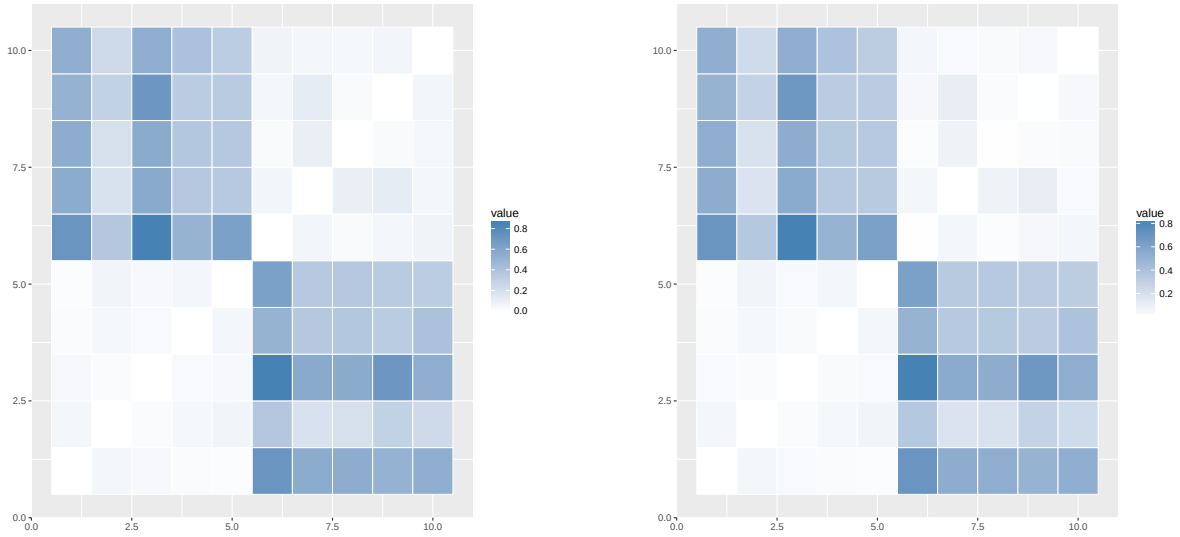
(b) CRQA

(c) fMDDL

Figura 4.2: Distância entre os componentes determinísticos((a),(b)) e estocásticos(c) após decomposição (sin +WN).

Finalmente, foram combinadas as distâncias dos componentes decompostos como mostra a Figuras 4.3. A distância considerando DWT e fMDDL apresentou resultados inferiores quando comparado ao uso da CRQA devido à etapa de decomposição, que pode alterar

levemente o alinhamento determinístico original sem modificar o atrator. Essa mudança afetou o melhor alinhamento entre os componentes determinísticos exigidos pelo DTW. Este problema não afetou a distância CRQA e fMDDL, uma vez que o CRQA calcula a distância entre os atratores dos componentes determinísticos. Assim, considerando a Figura 4 4.3, é possível notar que os dois grupos foram melhor discriminados.



(a) DTW + fMDDL (b) CRQA + fMDDL

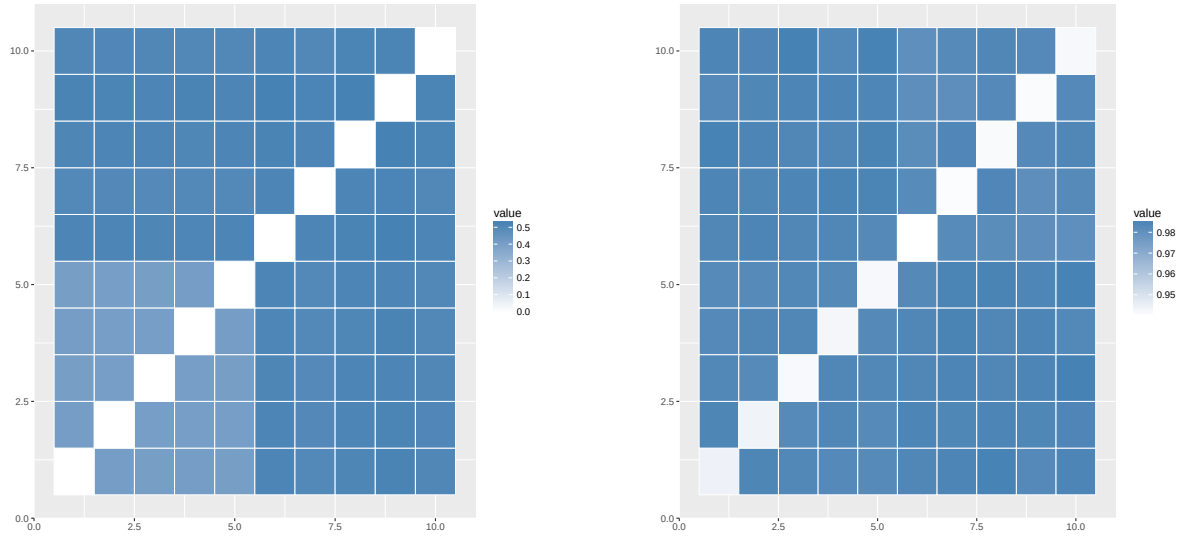
Figura 4.3: Distâncias DTW e CRQA entre as séries ruidosas após decomposição (sin +WN).

Para o segundo experimento foram criados dois grupos com a adição das observações geradas por AR($p = 1, \phi = 0{,}2$) e MA($q = 1, \theta = -0{,}7$) em séries senoidais. As distâncias DTW e CRQA entre as séries temporais ruidosas não foram úteis para encontrar diferenças entre os grupos, como mostrado na Figura 4.4, produzindo um mapa de calor homogêneo.

Após a decomposição das séries temporais, os componentes estocásticos foram discriminados pela medida de fMDDL, como mostra a Figura 4.5. As distâncias DTW e CRQA não foram relevantes, uma vez que foram utilizados os mesmos componentes determinísticos para criar os dois grupos. A CRQA proporcionou bons resultados, uma vez que o mapa de calor apresentou uma cor homogênea, mostrando todas as distâncias mais próximas de zero como esperado.

Finalmente, as distâncias dos componentes estocásticos e determinísticos foram combinadas conforme apresentadas na Figura 4.6. Similar ao resultado anterior, a fMDDL combinada com a CRQA permitiu identificar dois grupos diferentes.

Para o próximo experimento, foram adicionados dois processos ARMA ($p = 1, q = 1, \phi = -0{,}7, \theta = -0{,}9$) e ($p = 1, q = 1, \phi = 0{,}3, \theta = 0{,}9$) em séries senoidais. Como observado anteriormente, DTW e CRQA aplicadas diretamente nas séries ruidosas não foram capazes de encontrar os dois grupos esperados, como mostra as Figuras 4.7(a) e 4.7(b).



(a) DTW (b) CRQA

Figura 4.4: Distâncias DTW e CRQA entre as séries ruidosas (sin +AR e sin +MA) .

Os resultados obtidos utilizando a decomposição foram semelhantes aos anteriores, enfatizando as diferenças entre os componentes estocásticos (Figura 4.8(c)) e alta similaridade entre os determinísticos (Figuras 4.8(a) e (b)).

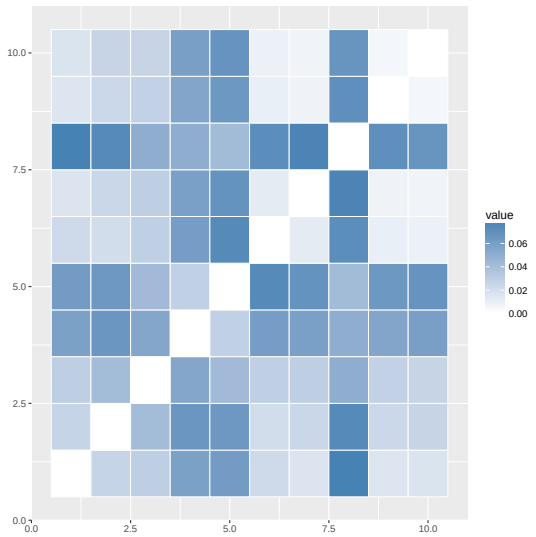
Considerando a combinação entre as medidas estocásticas e determinísticas (Figura 4.9), é possível observar que a fMDDL combinada com CRQA apresentou melhores resultados, embora a discriminação entre as séries temporais do primeiro grupo não tenha sido, visualmente, tão relevante quanto a segunda.

Para o próximo experimento, foram criados dois grupos adicionando o ruído branco ($\sigma = 0{,}5$) e ARMA ($p = 1, q = 1, \phi = -0{,}7, \theta = -0{,}9$), os quais foram adicionados as séries temporais senoidais. Como esperado, a DTW e a CRQA não foram capazes de discriminar corretamente os grupos, como mostra a Figura 4.10.

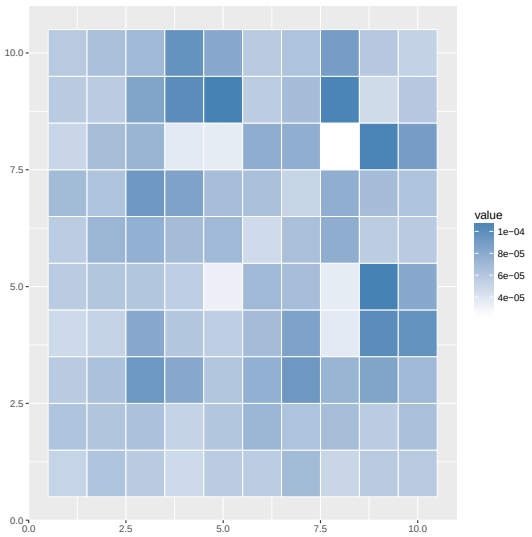
Após a decomposição das séries e análise individual de cada componente, é possível notar que a fMDDL apresentou bons resultados, discriminando os diferentes tipos de ruído (Figura 4.11(c)). A DTW e CRQA também apresentaram bons resultado, uma vez que a maioria dos componentes determinísticos apresentou distâncias próximas a 0 (Figuras 4.11(a) e 4.11(b), respectivamente).

A Figura 4.12 apresenta os resultados após a combinação das medidas estocásticas e determinísticas. Neste caso, vale ressaltar que a CRQA e a fMDDL apresentaram os bons resultados, discriminando corretamente os dois grupos (Figura 4.12 (b)).

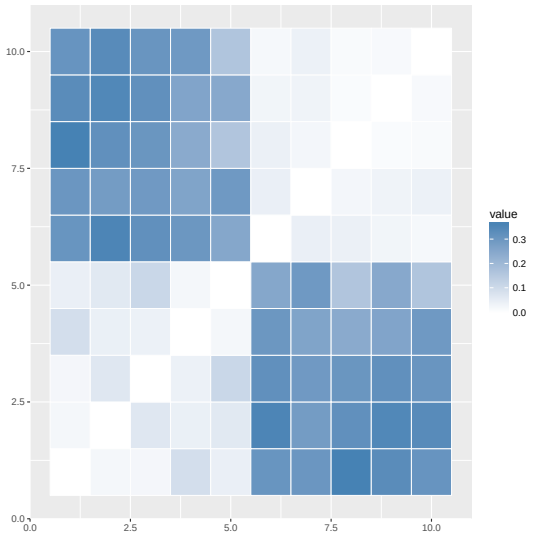
Finalmente, no último experimento, foi avaliada a relevância da decomposição calculando a distância entre as séries temporais mais complexas criadas pela combinação de ruído uniforme ($min = 0$, $max = 0{,}3$) e ruído branco ($\sigma = 0{,}05$) com séries temporais caóticas produzidas pelo sistema de Lorenz. Da mesma forma que os experimentos apresentados anteriormente, as distâncias DTW e CRQA nas séries ruidosas não foram úteis para discriminar os dois grupos, como mostrado na Figura 4.13.



(a) DTW



(b) CRQA

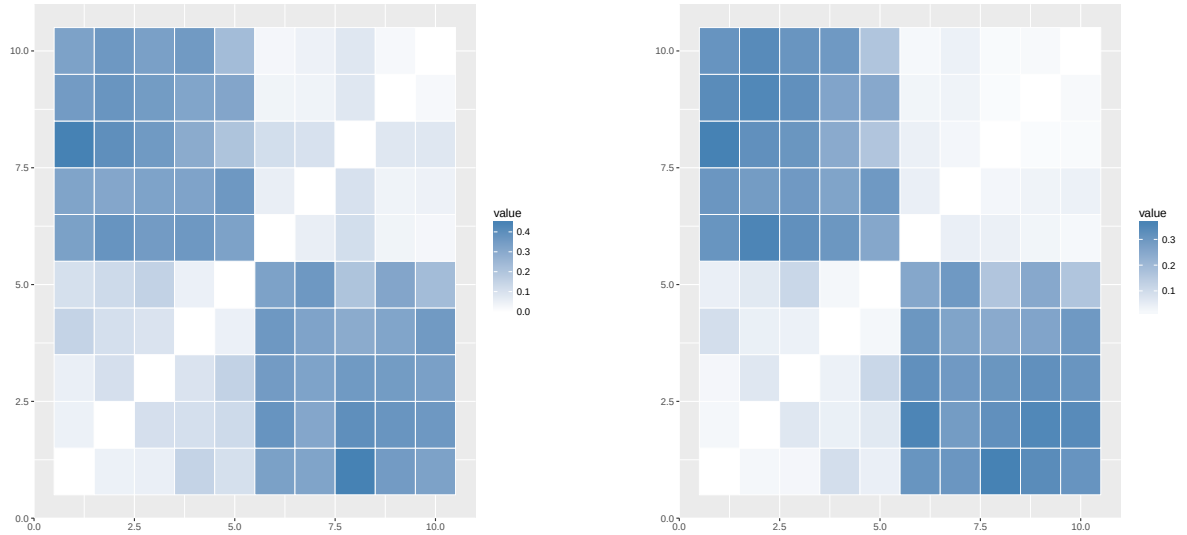


(c) fMDDL

Figura 4.5: Distâncias DTW, CRQA entre os componentes determinísticos((a),(b)) e distância fMDDL entre os componentes estocásticos(c) (sin +AR e sin +MA).

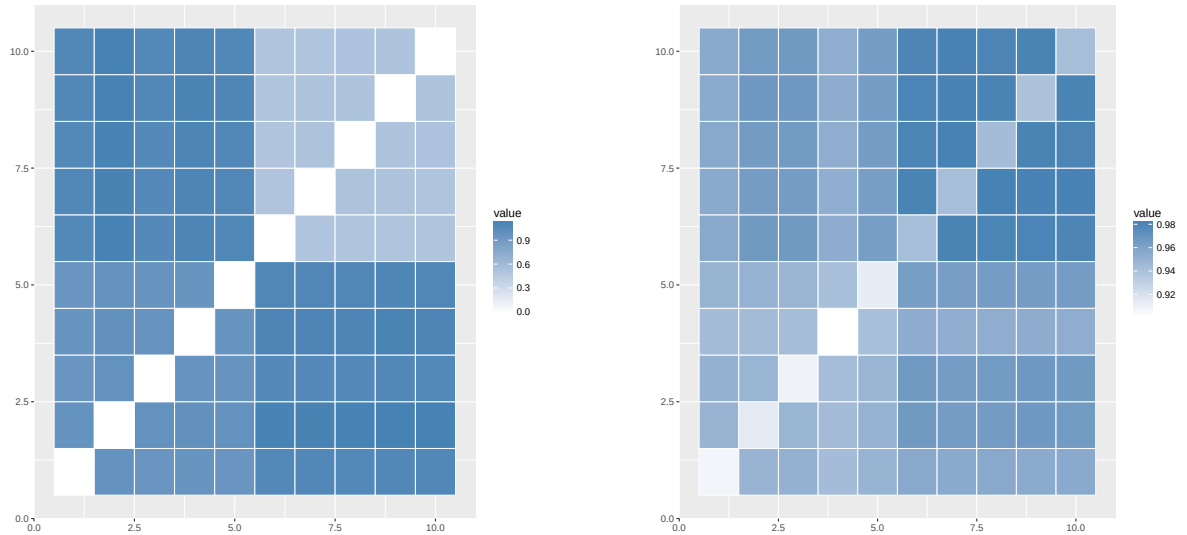
Por outro lado, analisando individualmente os componentes decompostos, é observado que a fMDDL discriminou os dois processos estocásticos (Figura 4.14(c)), enquanto DTW e CRQA mostraram distâncias mais próximas de zero para todos os componentes determinísticos, uma vez que foram criados usando o mesmo processo, Sistema Lorenz (Figura 4.14 (a), Figura 4.14(b)).

Finalmente, ao combinar as distâncias estocásticas e determinísticas, é possível notar que a CRQA e a fMDDL apresentaram bons resultados (Figura 4.15), especialmente para analisar uma série temporal complexa criada após a combinação de uma série temporal



(a) DTW + fMDDL (b) CRQA + fMDDL

Figura 4.6: Distâncias DTW e CRQA entre as séries ruidosas após decomposição (sin +AR e sin +MA).

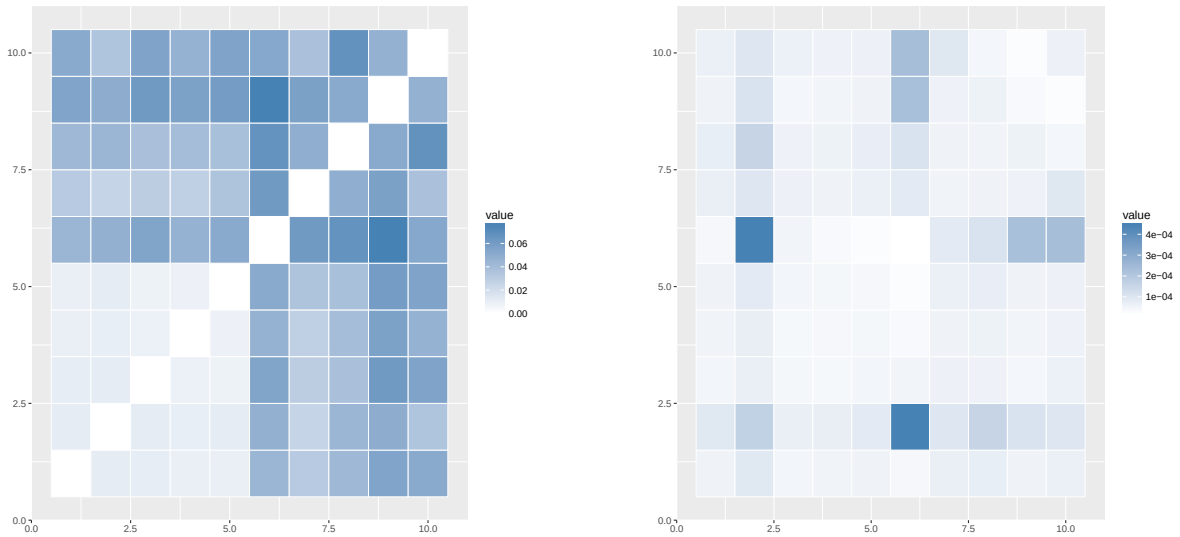


(a) DTW (b) CRQA

Figura 4.7: Distâncias DTW e CRQA entre as séries temporais (sin +ARMA).

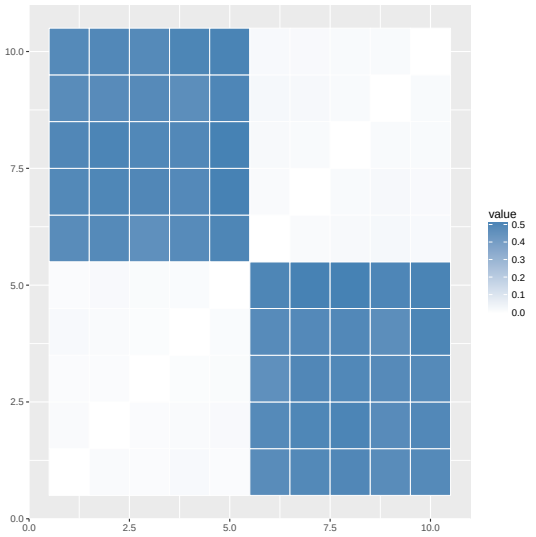
caótica a diferentes tipos ruidosos.

Considerando os resultados apresentados nesta seção, é possível notar a importância da decomposição em séries temporais com ruído aditivo, antes de calcular a distância entre elas.



(a) DTW

(b) CRQA



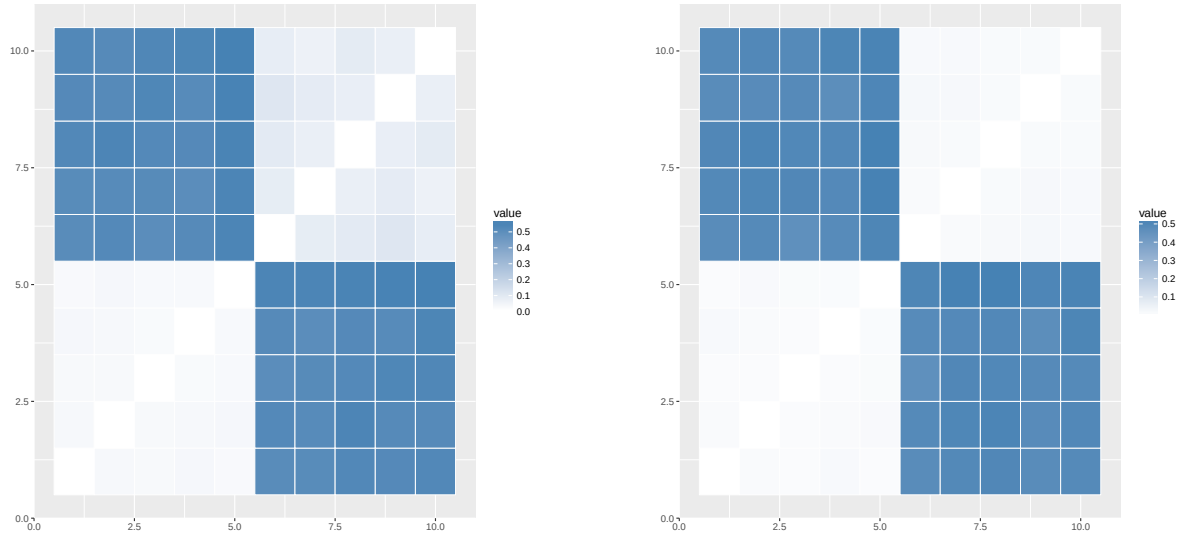
(c) fMDDL

Figura 4.8: Distâncias DTW, CRQA entre os componentes determinísticos((a),(b)) e distância fMDDL entre os componentes estocásticos(c) (sin +ARMA).

4.5 ANÁLISE DE AGRUPAMENTO

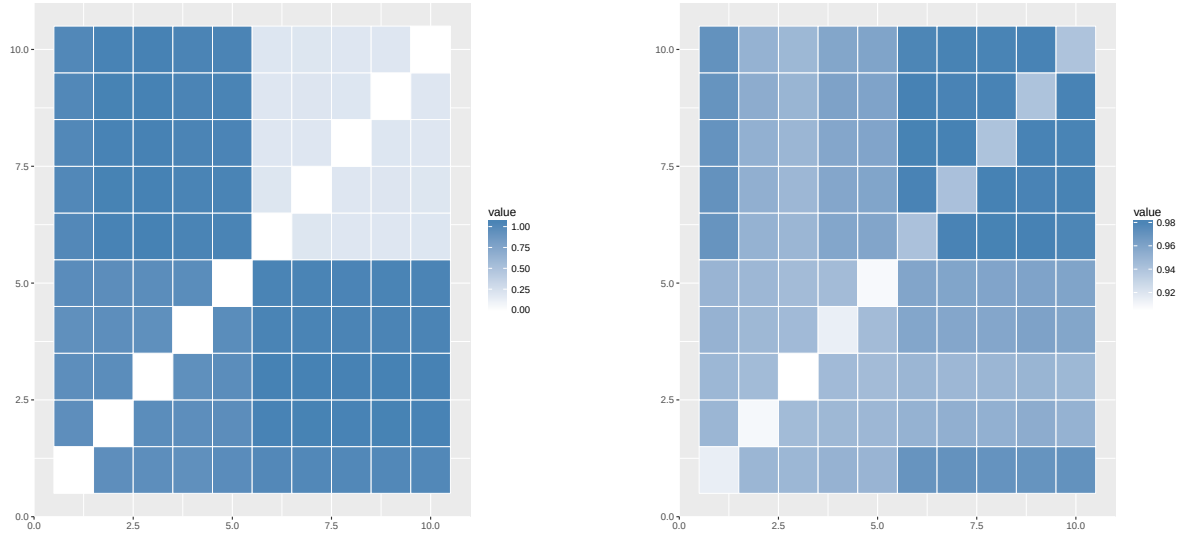
4.5.1 Configuração dos Experimentos

O último experimento foi realizado com o intuito de analisar a importância da aplicação de medidas de distância em componentes decompostos para agrupamento de séries temporais. Para isso, foram geradas 200 séries temporais, cada série contendo uma combinação de 1500 observações estocásticas e determinísticas, a quais foram organizadas em quatro grupos: i) 50 séries temporais combinando seno ou cosseno a um ruído branco com média



(a) DTW + fMDDL (b) CRQA + fMDDL

Figura 4.9: Distâncias DTW e CRQA entre as séries ruidosas após decomposição (sin +ARMA).

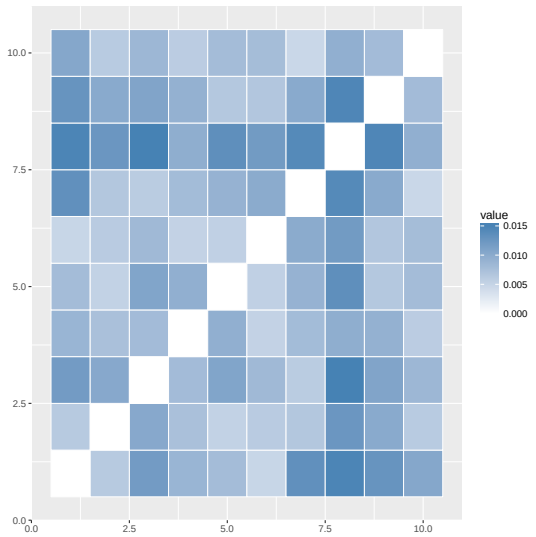


(a) DTW (b) CRQA

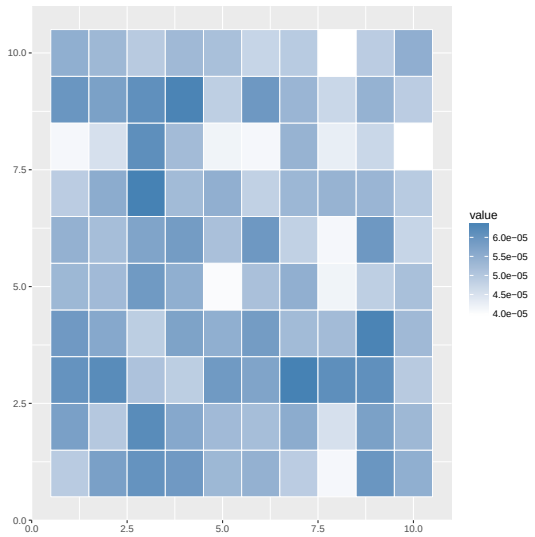
Figura 4.10: Distâncias DTW e CRQA entre as séries ruidosas (sin +ARMA e sin +WM).

igual a 0 e desvio padrão igual a 0,5; ii) 50 séries temporais seno ou cosseno combinadas a um processo ARMA com $p = q = 1$, $\phi = -0{,}7$ e $\theta = -0{,}9$; iii) 50 séries temporais combinando Lorenz com ruído branco com média igual a 0 e desvio padrão igual a 0,05; iv) 50 séries temporais combinando Lorenz com ruído uniforme no intervalo $[0, 3]$.

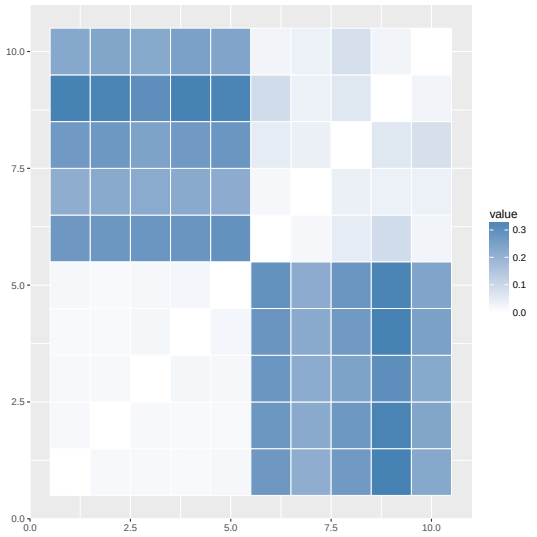
Foi utilizado o sistema de Lorenz com parâmetros iguais a $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$ para produzir 20000 observações caóticas. Em seguida, foram selecionadas aleatoria-



(a) DTW



(b) CRQA

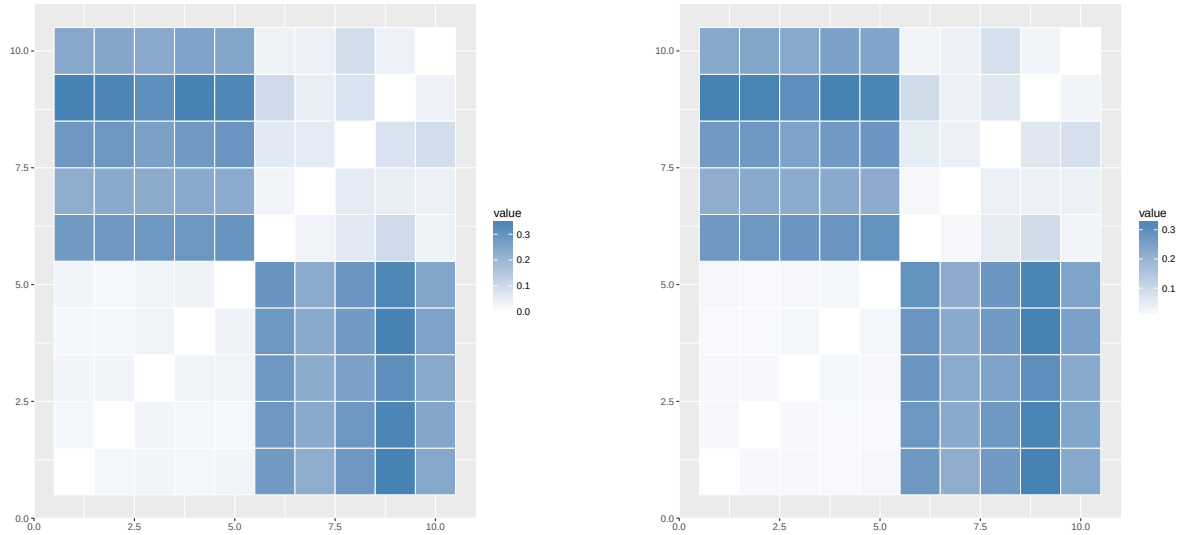


(c) fMDDL

Figura 4.11: Distâncias DTW, CRQA entre os componentes determinísticos ((a),(b)) e distância fMDDL entre os componentes estocásticos (c) (sin +ARMA e sin +WM) .

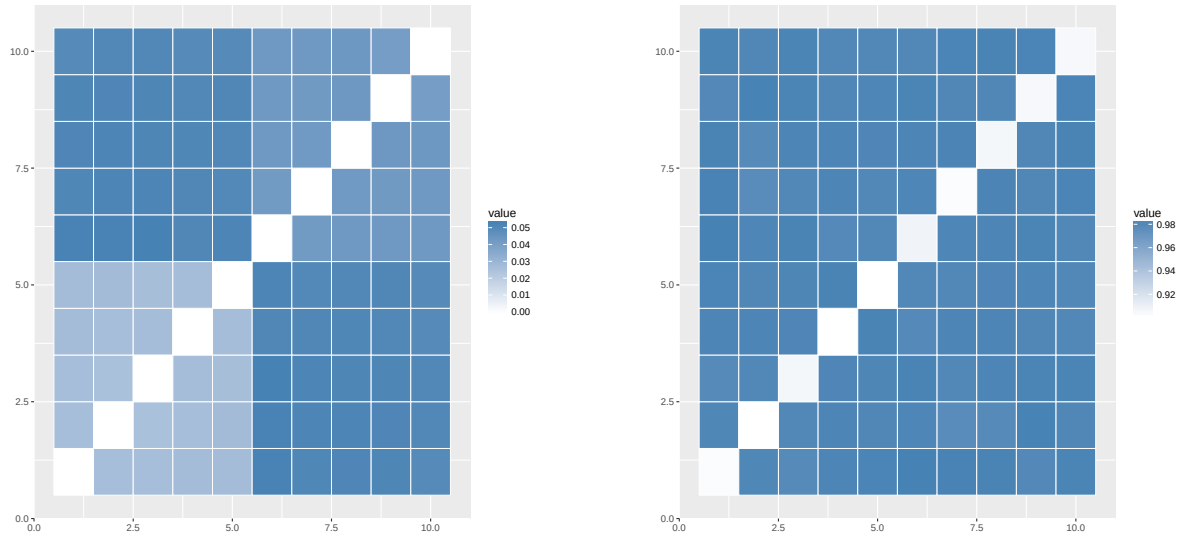
mente 100 janelas de 1500 observações para a criação dos grupos iii) e iv) apresentados anteriormente.

Para validar a hipótese deste trabalho, foi utilizado o algoritmo particional K-means [MacQueen 1967] para agrupar a série temporal com o parâmetro $k = 4$. Inicialmente, esse algoritmo funciona selecionando k centroides. Então, os centroides são iterativamente atualizados de acordo com sua distância em relação à série temporal. Na última etapa, quando nenhuma atualização é realizada, as séries temporais são organizadas em k grupos, de acordo com os k centroides. As estruturas obtidas com K-means foram avaliadas



(a) DTW + fMDDL (b) fMDDL

Figura 4.12: Distâncias DTW e CRQA entre as séries ruidosas após decomposição (sin +ARMA e sin +WM) .

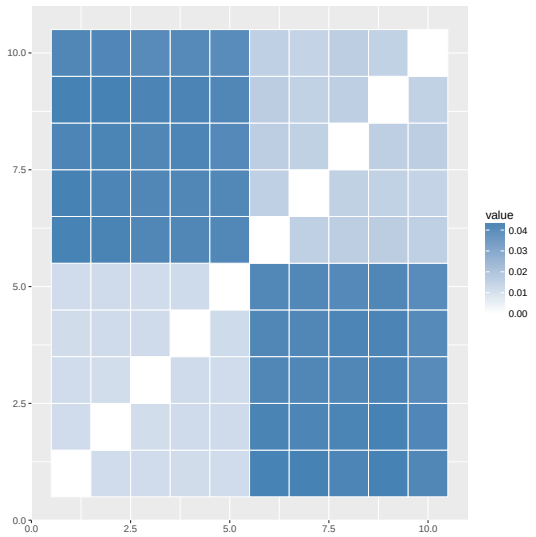


(a) DTW (b) CRQA

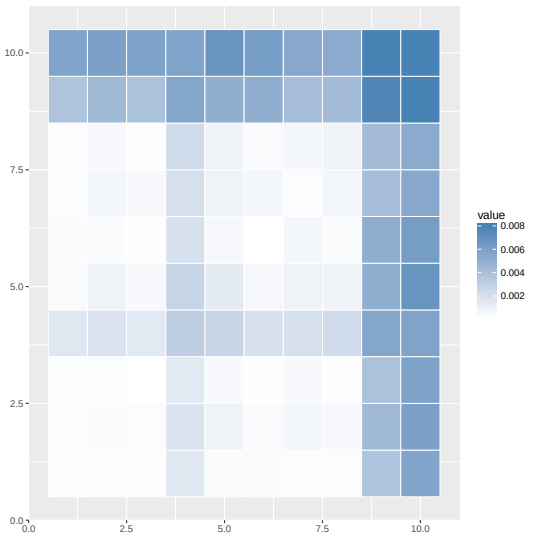
Figura 4.13: Distâncias DTW e CRQA entre as séries ruidosas (Lor + WN e Lor + unif).

utilizando três índices externos amplamente adotados em pesquisas de agrupamento e que foram discutidos na Seção 2.6: i) Fowlkes-Mallows; ii) Jaccard; e iii) Rand.

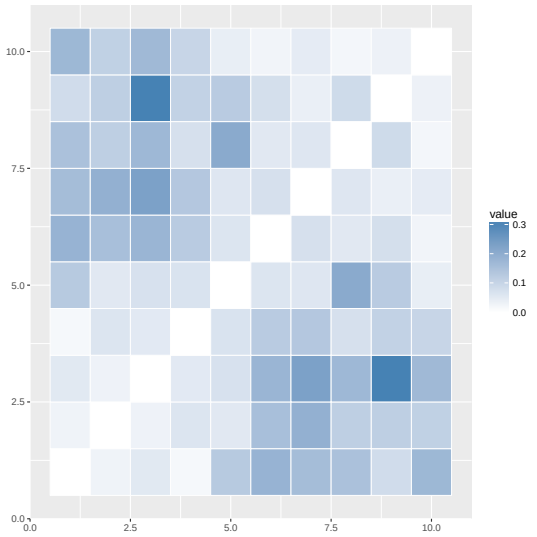
Todos os índices variam entre 0 e 1, de tal forma que quanto maior o valor, melhor é o agrupamento. Na próxima seção, são apresentados os resultados obtidos.



(a) DTW



(b) CRQA

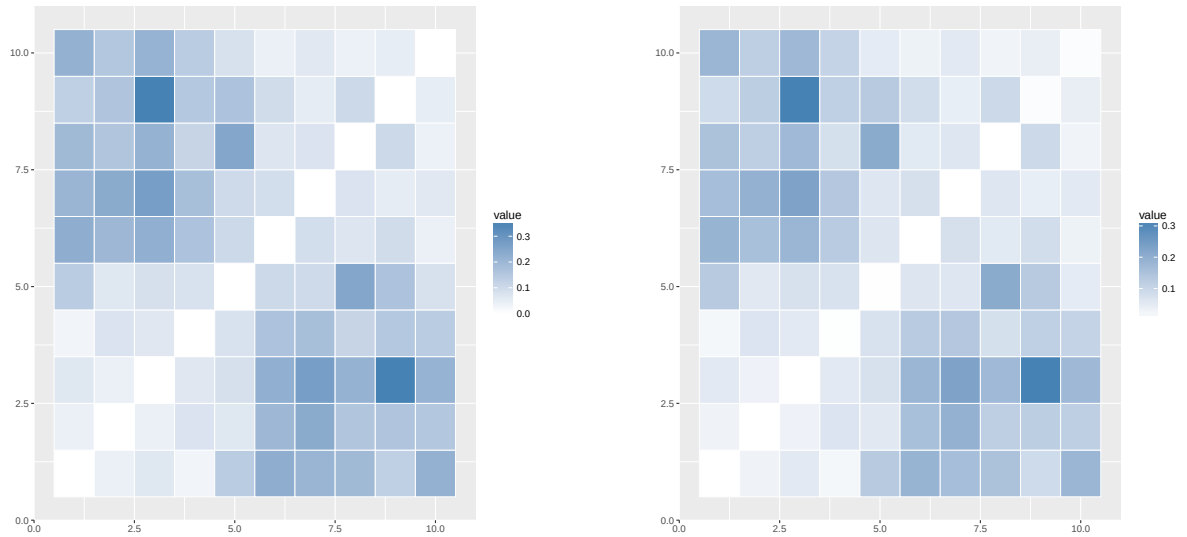


(c) fMDDL

Figura 4.14: Distâncias DTW, CRQA entre os componentes determinísticos ((a),(b)) e distância fMDDL entre os componentes estocásticos (c) (Lor + WN e Lor + unif)

4.5.2 Resultados

Os resultados obtidos são mostrados na Figura 4.16. Como é possível notar, resultados dos índices melhoraram a partir da decomposição das séries temporais e análise individual de seus componentes.



(a) DTW + fMDDL (b) CRQA + fMDDL

Figura 4.15: Distâncias DTW e CRQA entre as séries ruidosas após decomposição (Lor + WN e Lor + unif) .

4.6 CONSIDERAÇÕES FINAIS

Nesta seção, foram apresentados os experimentos executados neste trabalho, os quais buscaram verificar a importância da utilização da decomposição de séries temporais no processo de agrupamento de séries temporais. Com base nos resultados obtidos, foi possível validar a hipótese definida na Seção 1.

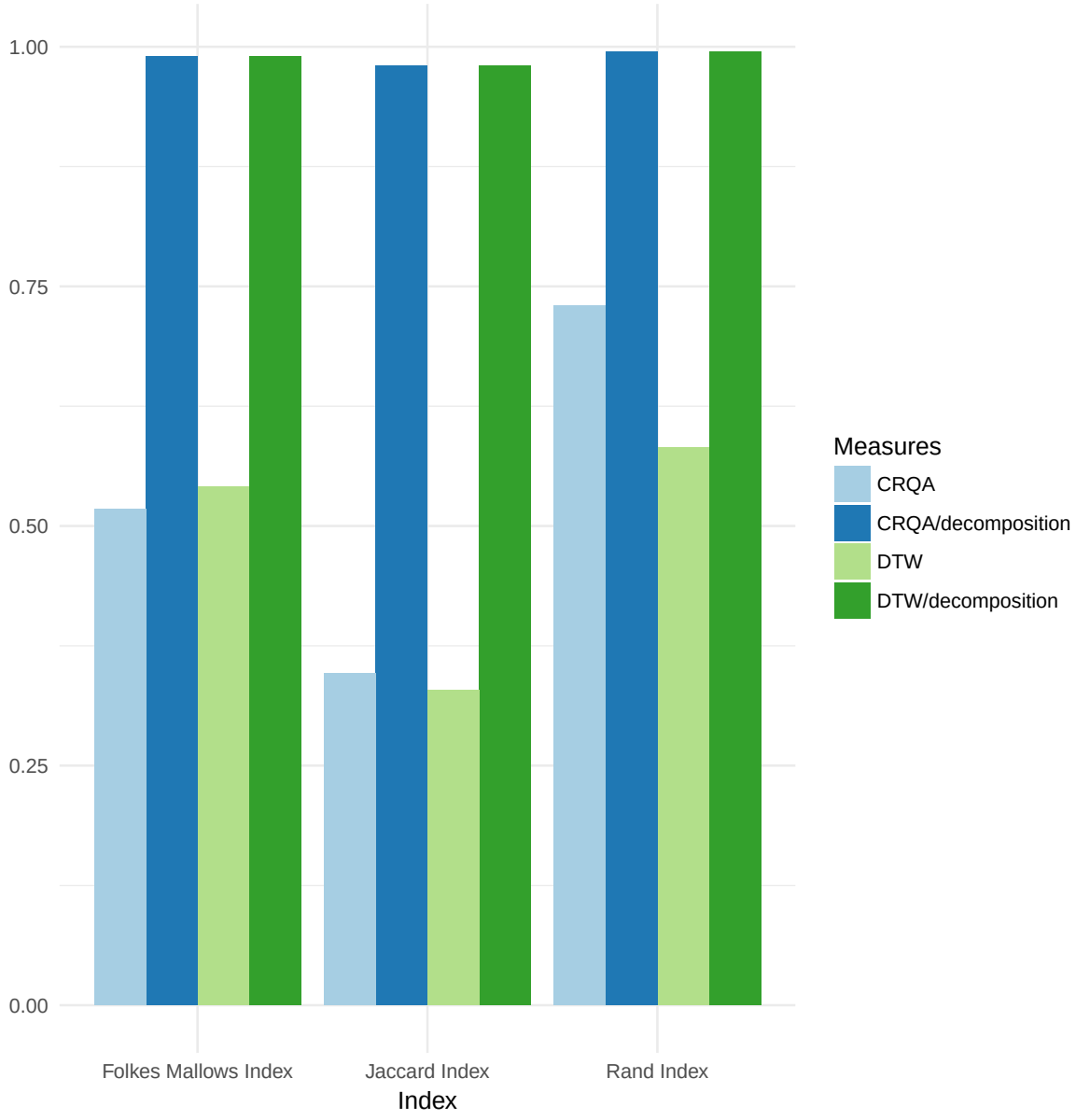


Figura 4.16: Índices de validação externa dos grupos gerados com K-means antes e após decomposição.

Capítulo 5

CONCLUSÕES

Neste trabalho de mestrado, uma nova abordagem de agrupamento de séries temporais foi desenvolvida para reduzir a influência de ruído aditivo. Essa nova abordagem atua, basicamente, no cálculo das medidas de similaridades e foi desenvolvida com base na hipótese que o agrupamento de séries temporais apresenta maior acurácia quando medidas de similaridade (ou distância) são, individualmente, calculadas sobre comportamentos estocásticos e determinísticos.

Durante o desenvolvimento deste trabalho, observou-se que as medidas usadas para calcular a similaridade/distância entre séries temporais foram desenvolvidas supondo a existência de padrões determinísticos. Contudo, quando as séries são compostas apenas por observações estocásticas, tais medidas não conseguem identificar diferenças existentes entre suas regras geradoras, afetando a qualidade na extração correta de grupos.

Para resolver este problema, uma nova medida chamada fMDDL foi proposta neste trabalho, a qual permite distinguir séries temporais estocásticas analisando seus componentes no domínio das frequências. A combinação da fMDDL com medidas determinísticas permitiu identificar grupos de séries criadas com diferentes estruturas determinísticas somadas com diferentes tipos de ruído.

A comprovação da hipótese deste trabalho foi realizada em 4 etapas. Inicialmente, séries ruidosas foram decompostas e medidas determinísticas foram analisadas apenas sobre os componentes determinísticos, permitindo a escolha das medidas DTW e DET-CRQA. Em seguida, a medida fMDDL foi avaliada sobre os componentes estocásticos. Na terceira etapa, avaliou-se a importância de usar fMDDL combinada com DTW e DET-CRQA. Por fim, essa combinação de medidas foi utilizada em um processo completo de agrupamento. Resultados enfatizaram a importância de analisar individualmente as influências estocásticas e determinísticas, comprovando a hipótese deste trabalho.

A diferença pouco significativa entre DTW e DET-CRQA no resultado obtido com o agrupamento, pode ser explicado pelo viés da técnica de decomposição. À medida que a influencia do ruído aumenta nas séries estudadas, a técnica de decomposição tende a extrair componentes determinísticos com formato similar a senoides. Isso faz com que

o resultado da DTW seja semelhante ao resultado obtido com a DET-CRQA. Contudo, se houvesse uma técnica de decomposição capaz de extrair o componente estocástico preservando os atratores esperados das séries caóticas, a técnica DET-CRQA tenderia a apresentar melhores resultados.

Assim, como trabalho futuro, espera-se estudar outras técnicas de decomposição para melhor caracterizar as diferenças entre DET-CRQA e DTW. Além disso, espera-se ainda utilizar a abordagem proposta em diferentes tarefas de aprendizado de máquina. Neste sentido, a aplicação da abordagem para classificar séries temporais poderia ser realizada sem grandes alterações quando comparada à atividade de agrupamento, visto que a única modificação necessária seria a troca de algoritmos de agrupamento por algoritmos de classificação.

REFERÊNCIAS BIBLIOGRÁFICAS

AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering – a decade review. *Information Systems*, v. 53, p. 16 – 38, 2015.

AGRAWAL, R. *Automatic subspace clustering of high dimensional data for data mining applications*. [S.l.]: ACM, 1998.

ALLIGOOD, K.; SAUER, T.; YORKE, J. *Chaos: An Introduction to Dynamical Systems*. [S.l.]: Springer New York, 1997. (Textbooks in Mathematical Sciences).

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.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, The Royal Society, v. 160, n. 901, p. 268–282, 1937.

BATISTA, G. E.; WANG, X.; KEOGH, E. J. A complexity-invariant distance measure for time series. In: SIAM. *Proceedings of the 2011 SIAM international conference on data mining*. [S.l.], 2011. p. 699–710.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: SEATTLE, WA. *KDD workshop*. [S.l.], 1994. v. 10, n. 16, p. 359–370.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738.

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. *Time Series Analysis: Forecasting and Control*. [S.l.]: Wiley, 2015. (Wiley Series in Probability and Statistics).

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

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).

DANIELSSON, P.-E. Euclidean distance mapping. *Computer Graphics and image processing*, Elsevier, v. 14, n. 3, p. 227–248, 1980.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, Taylor & Francis, v. 4, n. 1, p. 95–104, 1974.

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.

ELLIS, D. P. W.; POLINER, G. E. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In: *ICASSP*. [S.l.: s.n.], 2007. v. 4, p. IV–1429–IV–1432. ISSN 1520-6149.

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.

ESTER, M. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FACELI, K. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.: s.n.], 2011. 192 p.

FEI, W. 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.

FISHER, D. H. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, v. 2, n. 2, p. 139–172, Sep 1987. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1007/BF00114265⟩.

FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, Taylor Francis, v. 78, n. 383, p. 553–569, 1983.

GROENEN, P. J. F.; MATHAR, R.; HEISER, W. J. The majorization approach to multidimensional scaling for minkowski distances. *Journal of Classification*, v. 12, n. 1, p. 3–19, Mar 1995. Disponível em: ⟨https://doi.org/10.1007/BF01202265⟩.

GUHA, S.; RASTOGI, R.; SHIM, K. Cure: An efficient clustering algorithm for large databases. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 27, n. 2, p. 73–84, jun. 1998. ISSN 0163-5808. Disponível em: ⟨http://doi.acm.org/10.1145/276305.276312⟩.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. *Journal of Intelligent Information Systems*, v. 17, n. 2, p. 107–145, Dec 2001. ISSN 1573-7675. Disponível em: ⟨https://doi.org/10.1023/A:1012801612483⟩.

HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton University Press Princeton, 1994.

HINNEBURG, A.; KEIM, D. A. An efficient approach to clustering in large multimedia databases with noise. In: *KDD*. [S.l.: s.n.], 1998. v. 98, p. 58–65.

HÖPPNER, F.; KLAWONN, F. Compensation of translational displacement in time series clustering using cross correlation. In: ADAMS, N. M. (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. 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.

JACCARD, P. Nouvelles researches sur la distribution florale. *Bull Soc Vaud Sci Nat*, v. 44, p. 223–270, 1908.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988.

KAUFMAN, L.; ROUSSEEUW, P. J. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, Wiley Online Library, p. 68–125, 1990.

KOHONEN, T. Self-organizing maps, ser. *Information Sciences. Berlin: Springer*, v. 30, 2001.

LIAO, T. W. Clustering of time series data—a survey. *Pattern Recognition*, v. 38, n. 11, p. 1857 – 1874, 2005.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MARWAN, N.; KURTHS, J. Cross recurrence plots and their applications. *Mathematical Physics Research at the Cutting Edge*, Nova Science Hauppauge, p. 101–139, 2004.

MARWAN, N. Recurrence plots for the analysis of complex systems. *Physics Reports*, Elsevier, v. 438, n. 5, p. 237–329, 2007.

MEESRIKAMOLKUL, W.; NIENNATTRAKUL, V.; RATANAMAHATANA, C. A. Shape-based clustering for time series data. In: TAN, P.-N. (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.

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.

NG, R. T.; HAN, J. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, IEEE, v. 14, n. 5, p. 1003–1016, 2002.

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, Dec 2015. ISSN 0219-3116. Disponível em: ⟨https://doi.org/10.1007/s10115-014-0808-1⟩.

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.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, Taylor & Francis Group, v. 66, n. 336, p. 846–850, 1971.

RENDÓN, E. Internal versus external cluster validation indexes. *International Journal of computers and communications*, v. 5, n. 1, p. 27–34, 2011.

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.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

STUDENT. The probable error of a mean. *Biometrika*, JSTOR, p. 1–25, 1908.

TAKENS, F. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, Springer, v. 898, n. 1, p. 366–381, 1981.

THEODORIDIS, S.; KOUTROUBAS, K. Feature generation ii. *Pattern Recognition*, v. 2, p. 269–320, 1999.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001.

TORMENE, P. 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.

WANG, W. Sting: A statistical information grid approach to spatial data mining. In: *VLDB*. [S.l.: s.n.], 1997. v. 97, p. 186–195.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin*, JSTOR, v. 1, n. 6, p. 80–83, 1945.

XU, R.; WUNSCH, D. C. *Clustering. Hoboken*. [S.l.]: NJ: Wiley, 2009.

ZENG, Y. An adaptive meta-clustering approach: combining the information from different clustering results. In: IEEE. *Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society*. [S.l.], 2002. p. 276–287.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 25, n. 2, p. 103–114, jun. 1996. ISSN 0163-5808. Disponível em: ⟨http://doi.acm.org/10.1145/235968.233324⟩.

ZHANG, X. 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

Este apêndice consta as séries temporais utilizadas nos experimentos para análise dos componentes determinísticos, que são apresentados no Capitulo 4 (Seção 4.2). As configurações dessas séries foram apresentadas na Tabela 4.1. Além das séries temporais ruidosas também são apresentados seus componentes (estocástico e determinístico), os quais foram extraídos pelo processo de decomposição EMD.

A.1 SÉRIES COSSENO+RUÍDO

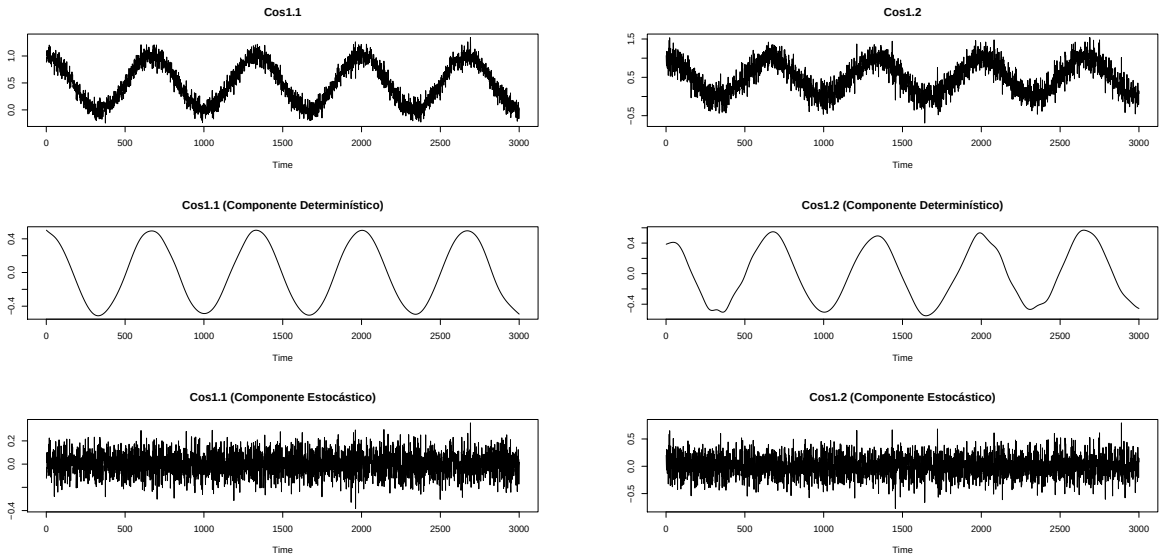


Figura A.1: Cos1.1 e Cos1.2

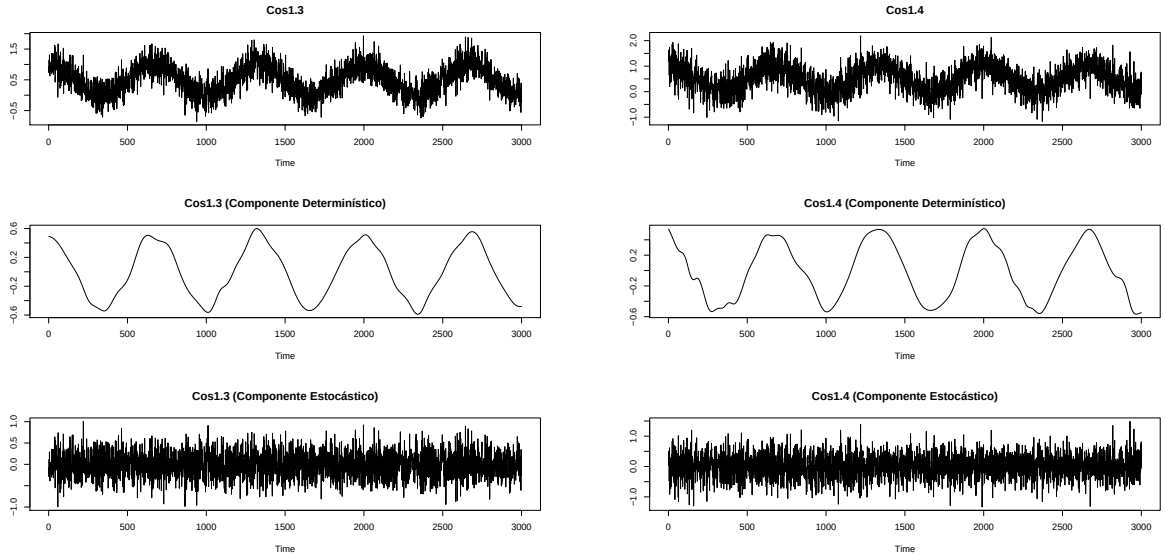


Figura A.2: Cos1.3 e Cos1.4

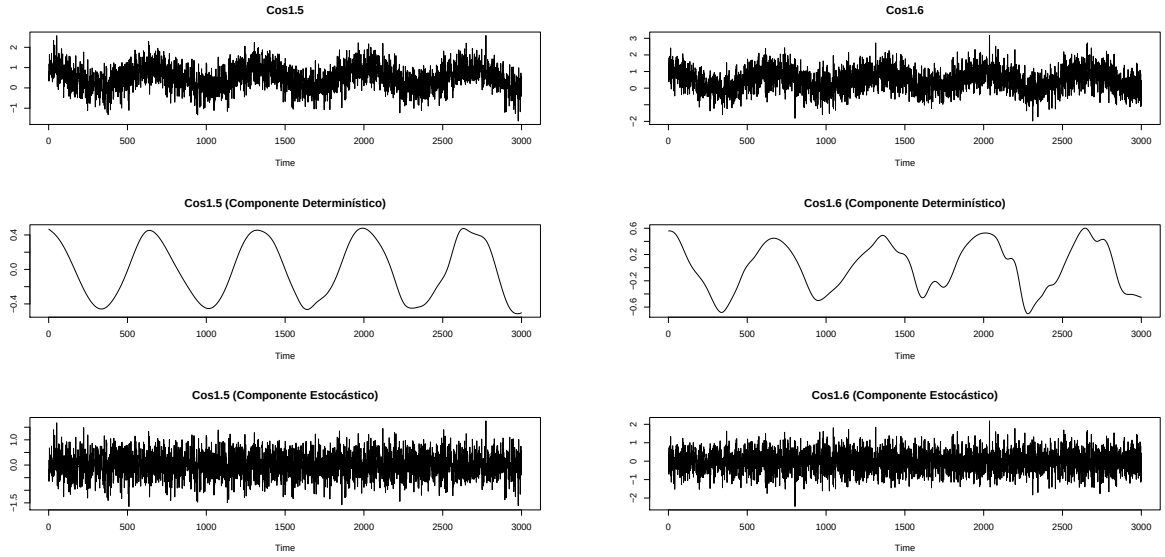


Figura A.3: Cos1.5 e Cos1.6

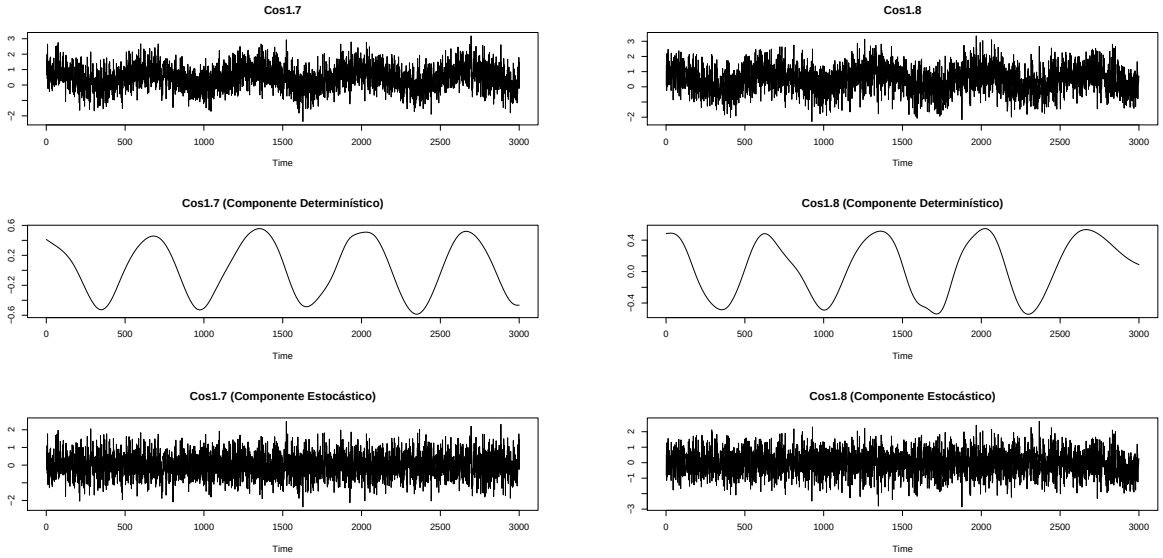


Figura A.4: Cos1.7 e Cos1.8

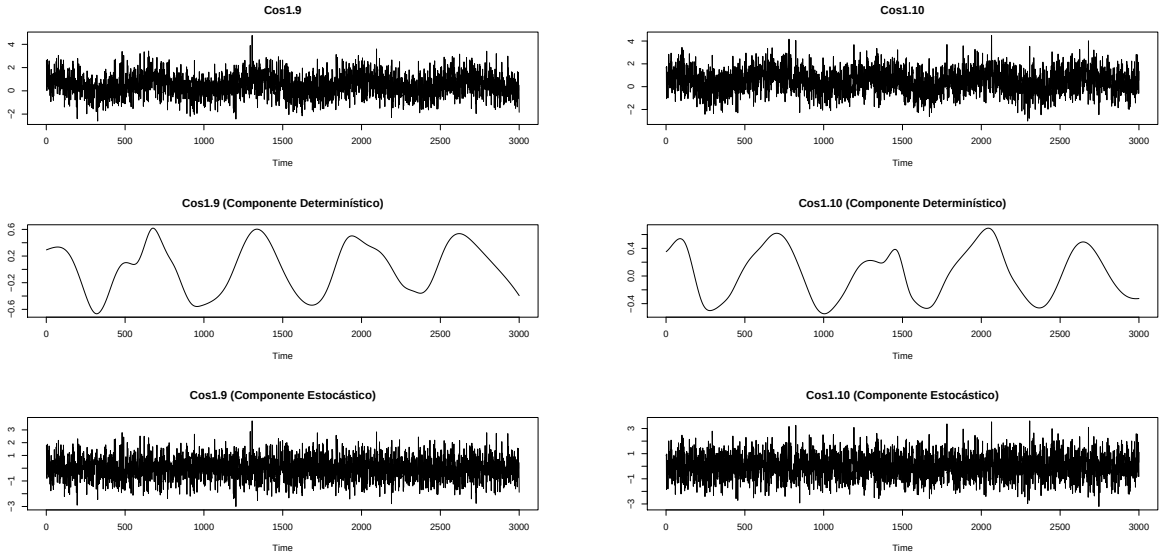


Figura A.5: Cos1.9 e Cos1.10

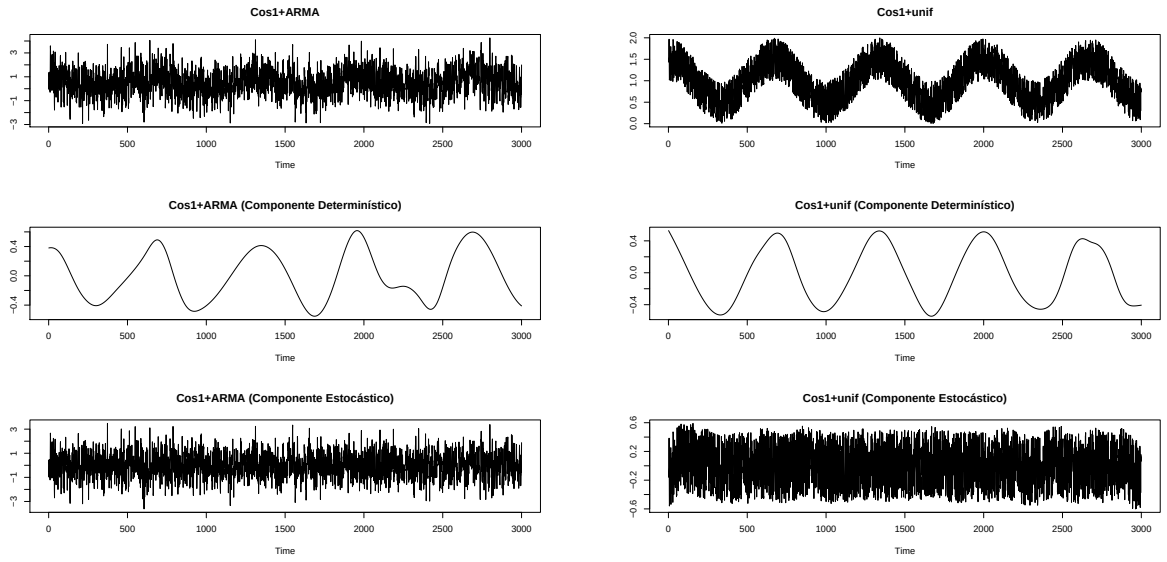


Figura A.6: Cos1+ARMA e Cos1+unif

A.2 SÉRIES COSSENO+RUÍDO+TENDÊNCIA

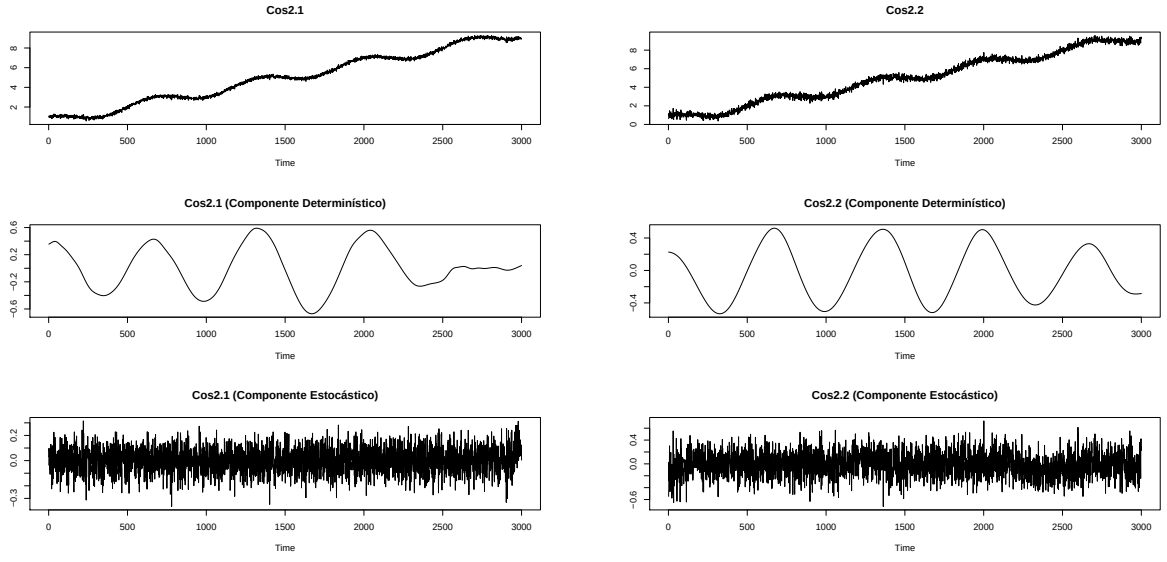


Figura A.7: Cos2.1 e Cos2.2

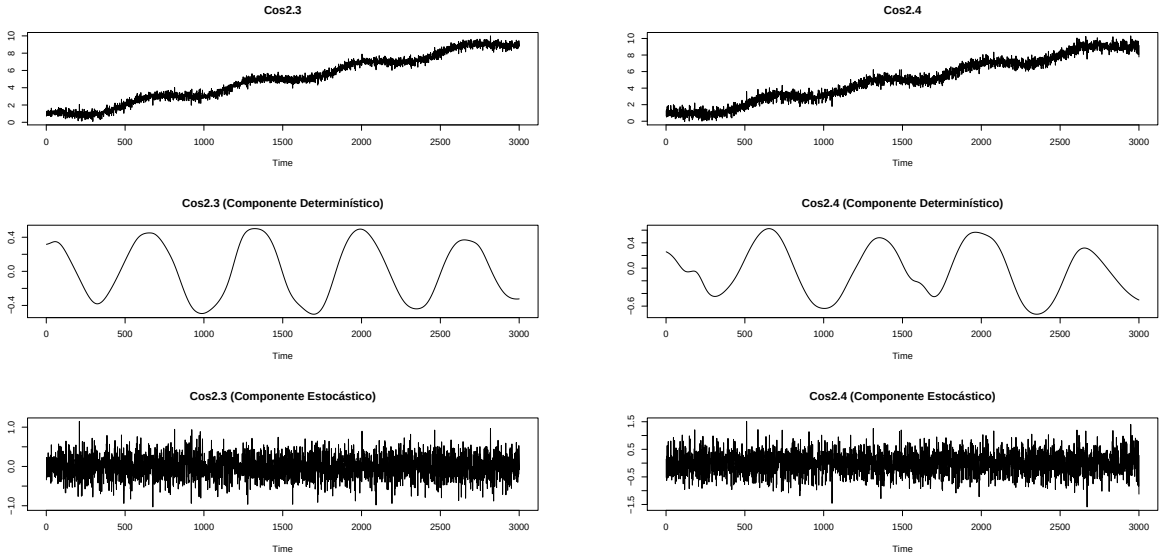


Figura A.8: Cos2.3 e Cos2.4

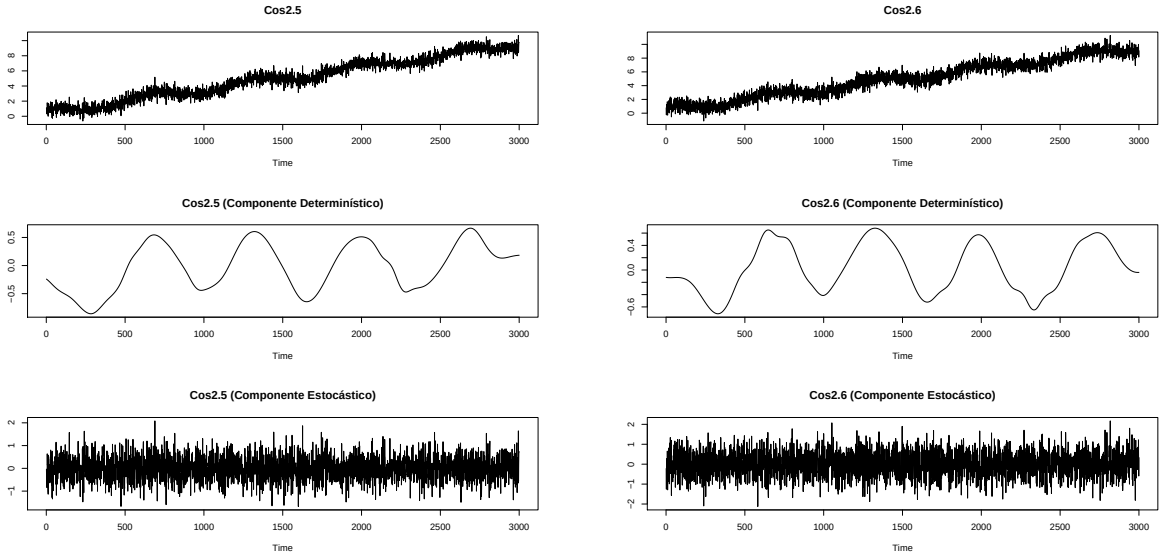


Figura A.9: Cos2.5 e Cos2.6

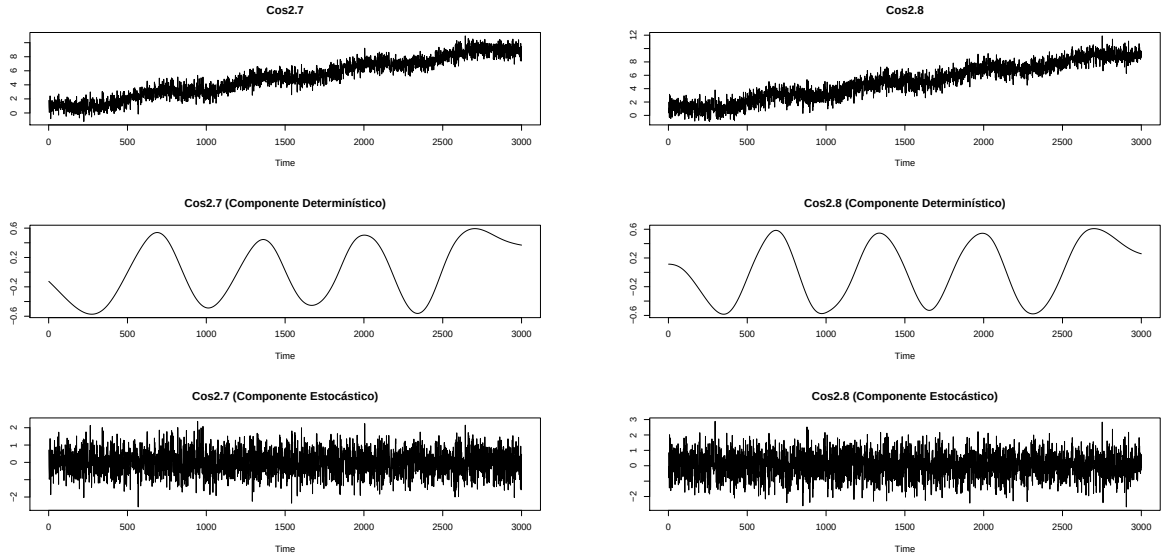


Figura A.10: Cos2.7 e Cos2.8

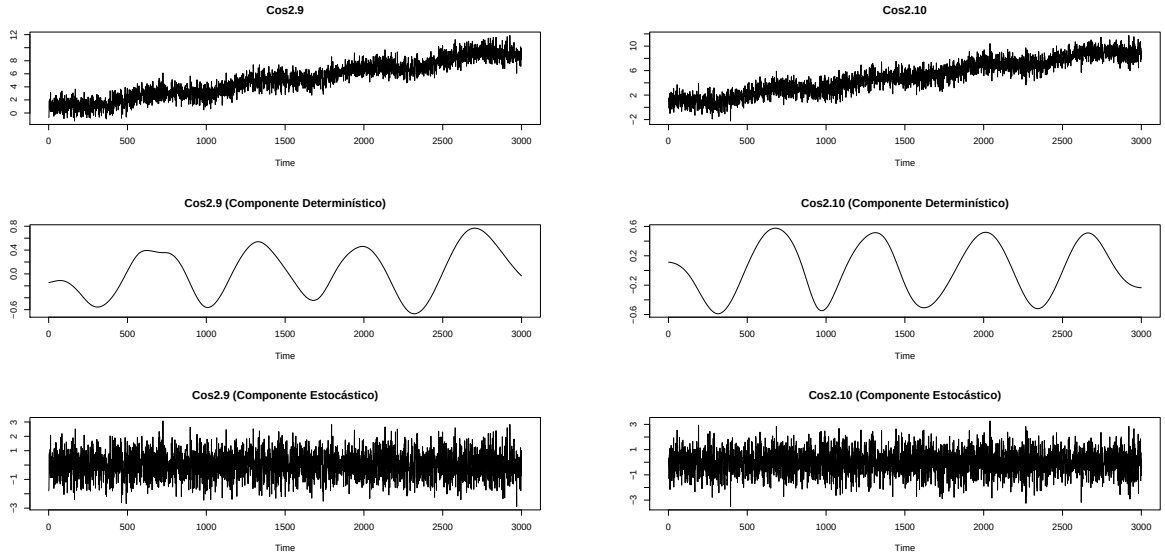


Figura A.11: Cos2.9 e Cos2.10

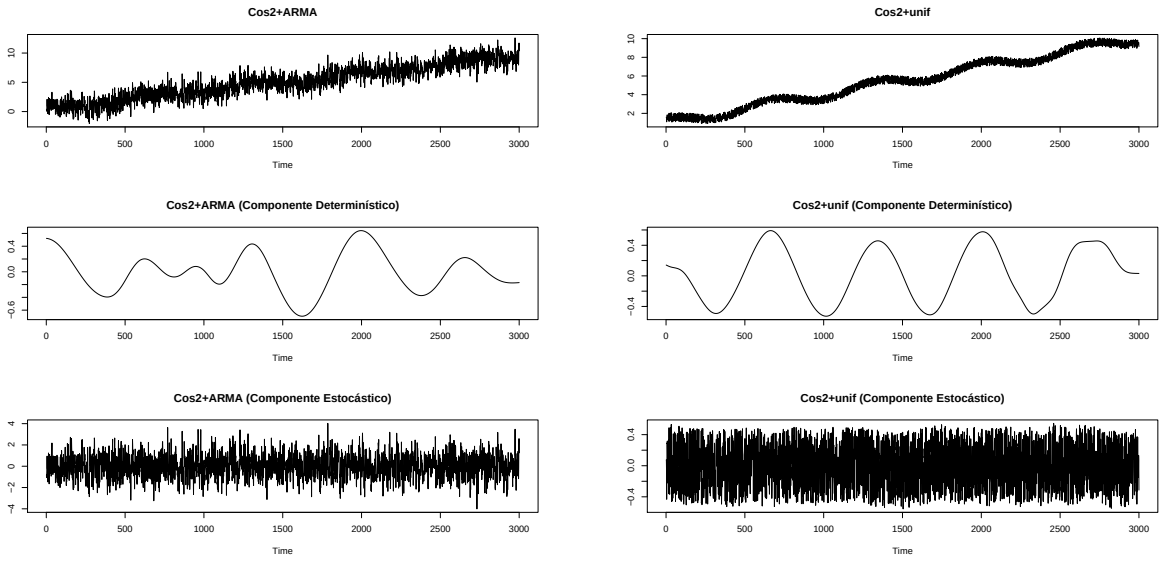


Figura A.12: Cos2+ARMA e Cos2+unif

A.3 SÉRIES SENO+RUÍDO

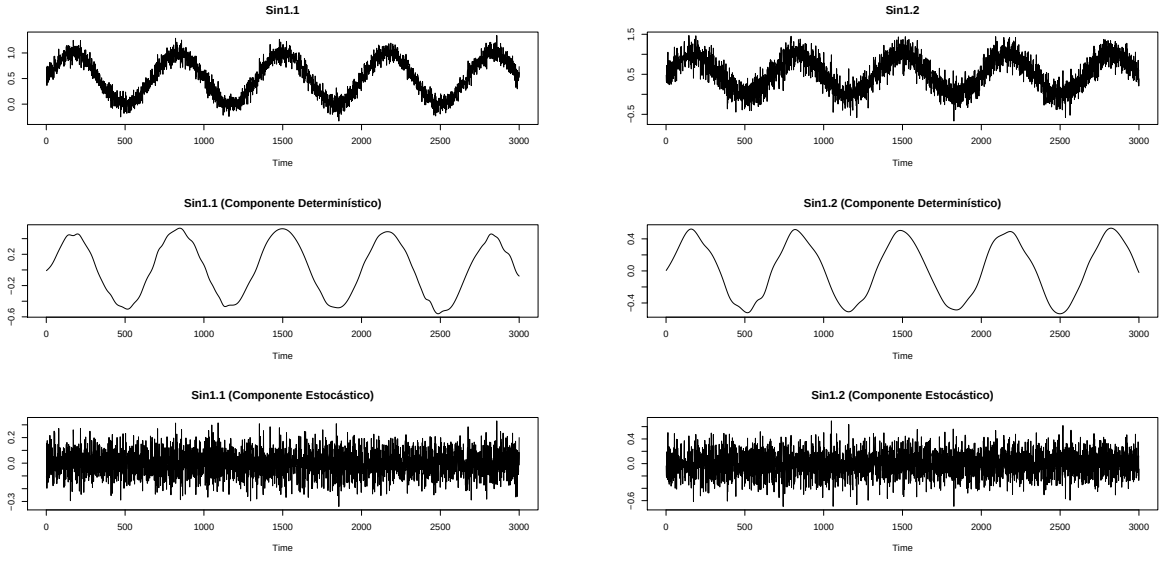


Figura A.13: Sin1.1 e Sin1.2

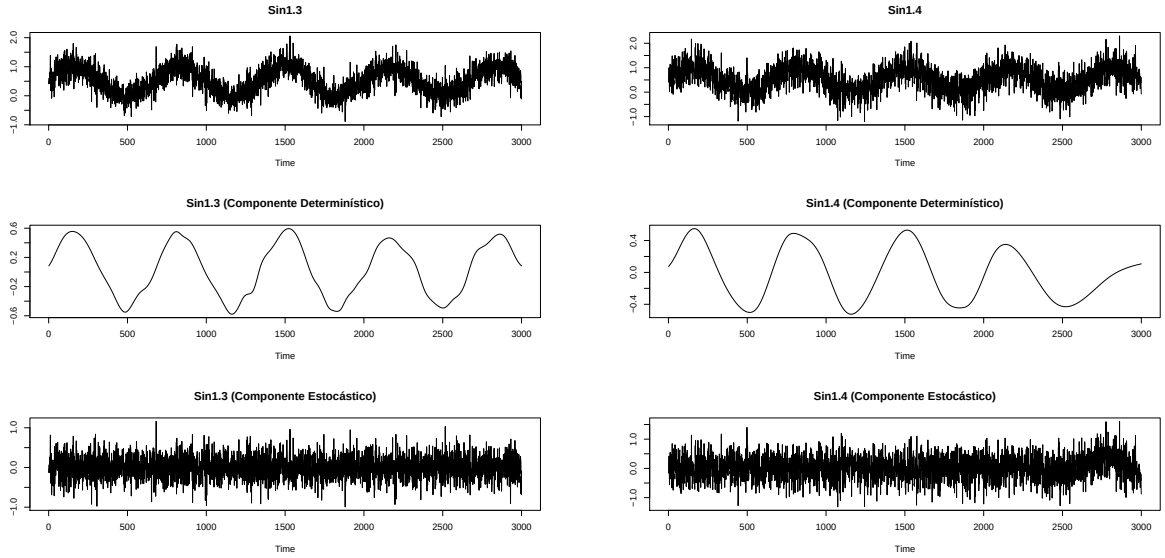


Figura A.14: Sin1.3 e Sin1.4

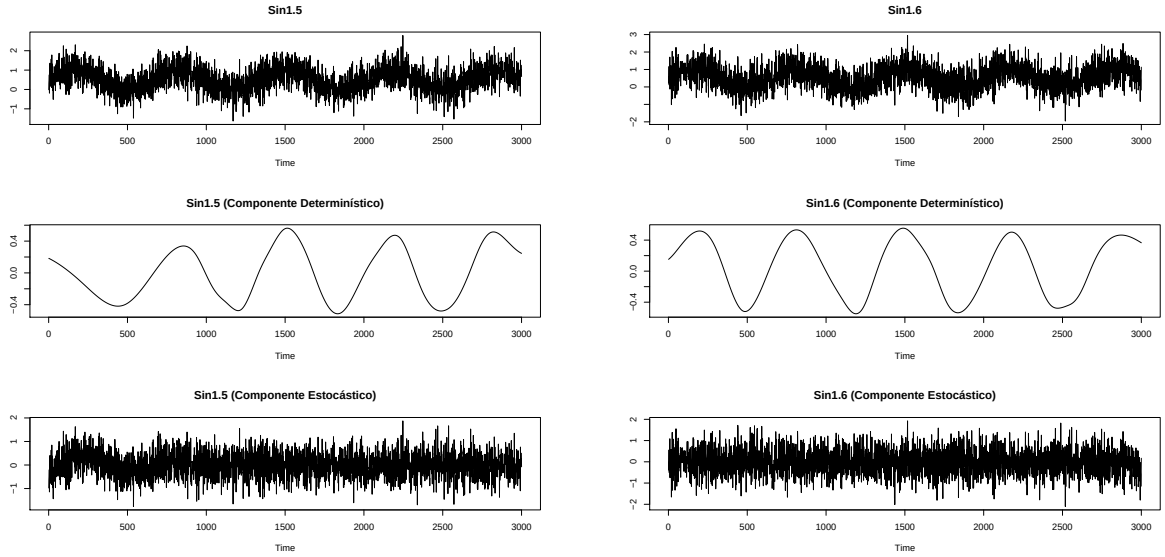


Figura A.15: Sin1.5 e Sin1.6

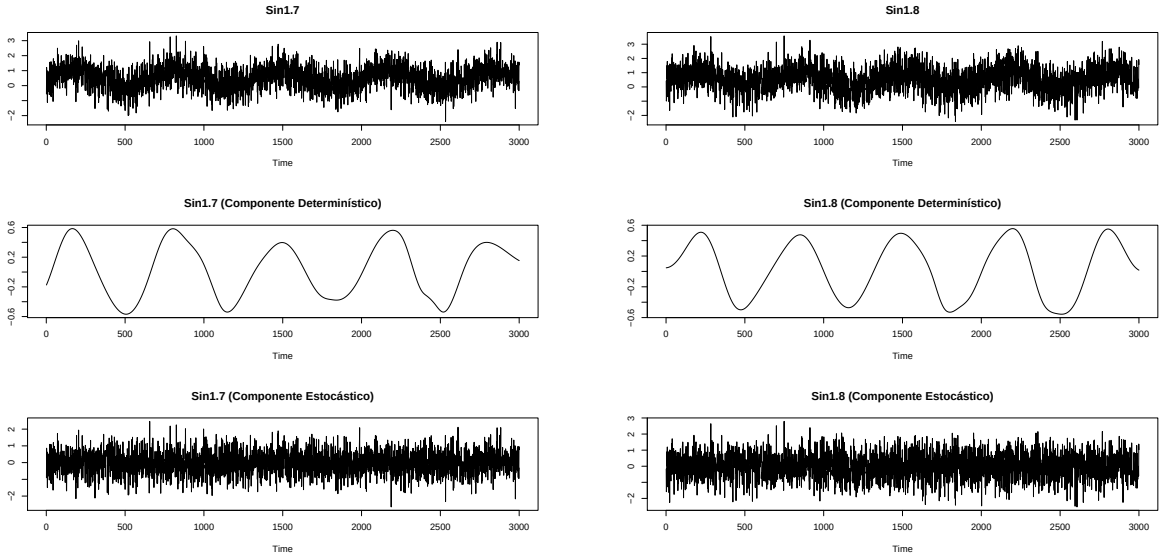


Figura A.16: Sin1.7 e Sin1.8

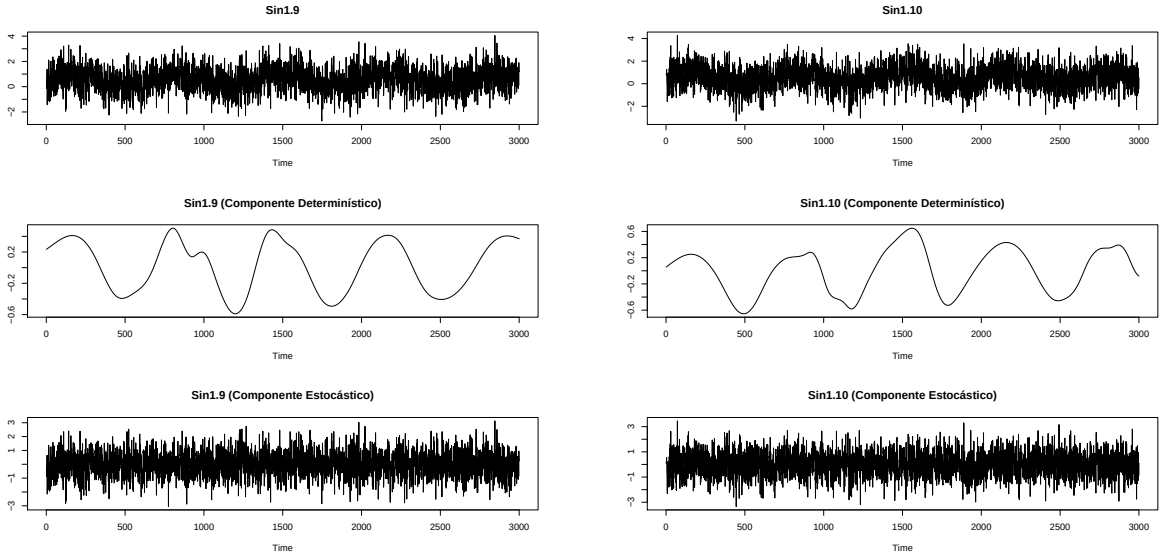


Figura A.17: Sin1.9 e Sin1.10

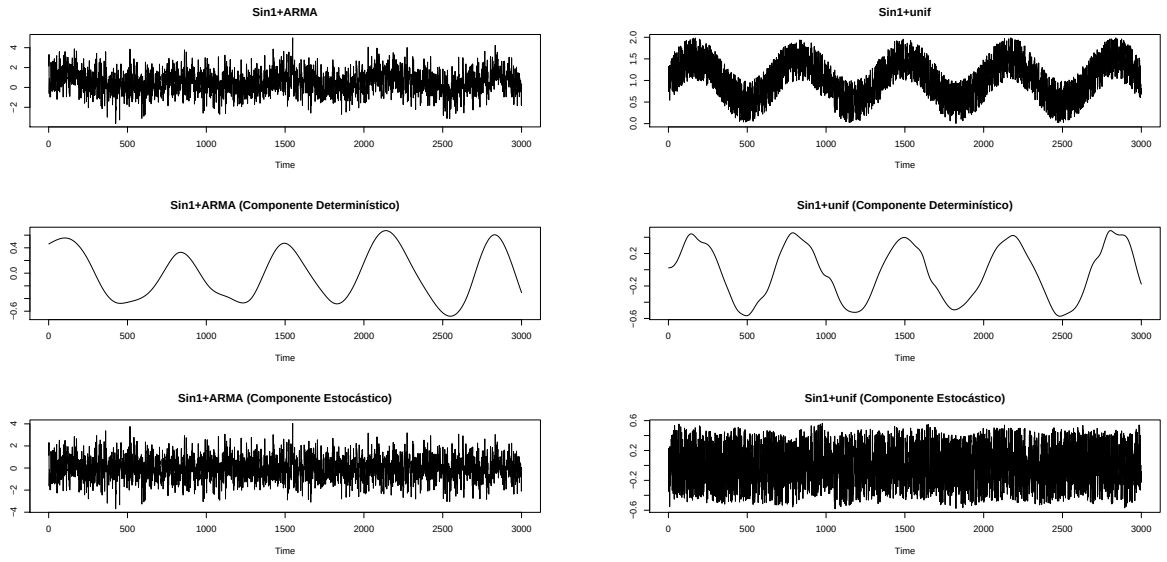


Figura A.18: Sin1+ARMA e Sin1+unif

A.4 SÉRIES SENO+RUÍDO+TENDÊNCIA

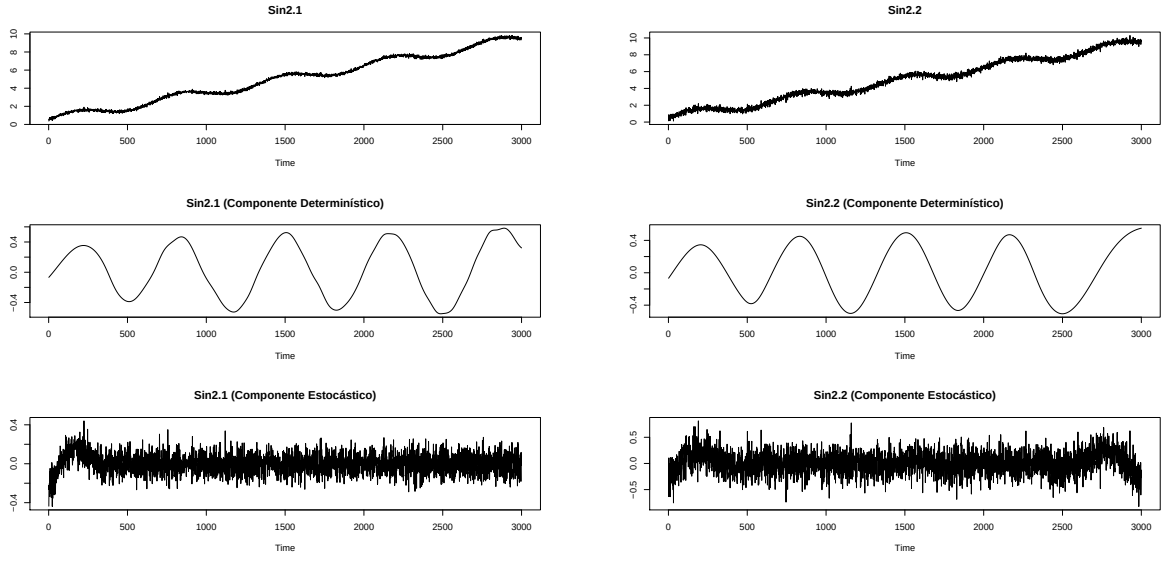


Figura A.19: Sin2.1 e Sin2.2

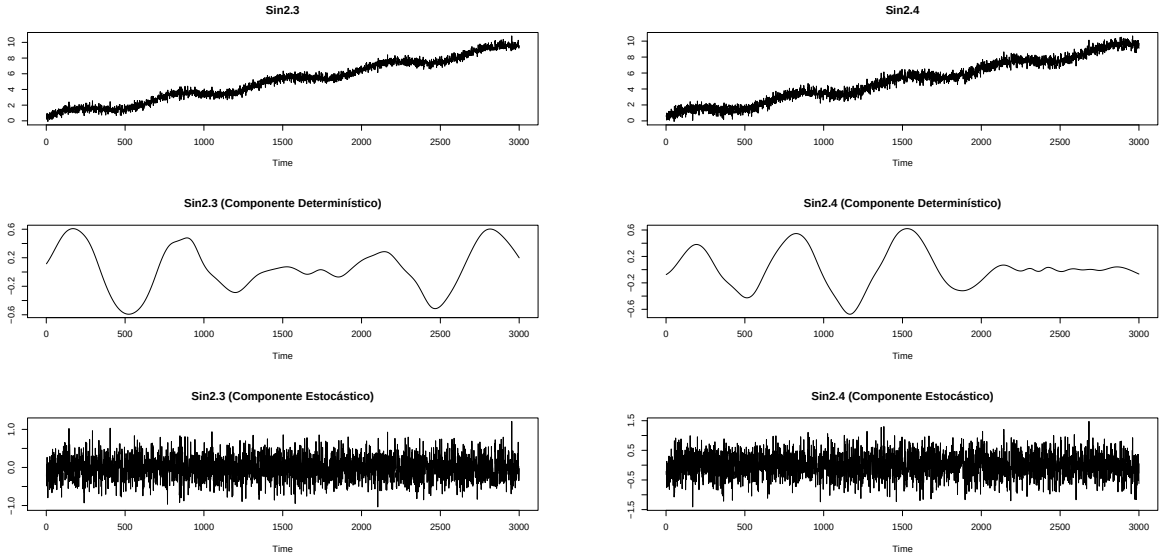


Figura A.20: Sin2.3 e Sin2.4

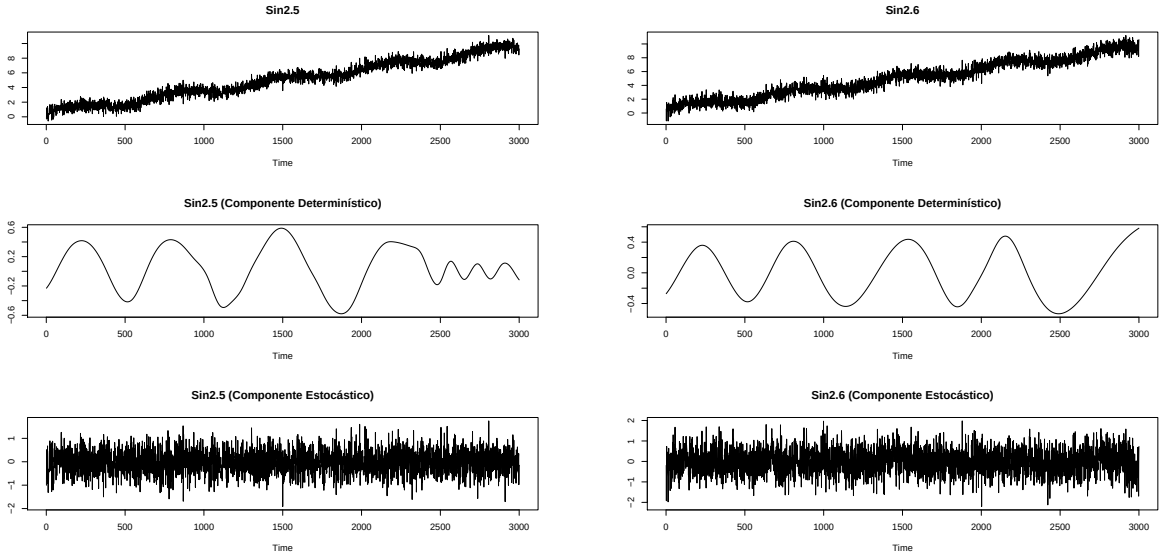


Figura A.21: Sin2.5 e Sin2.6

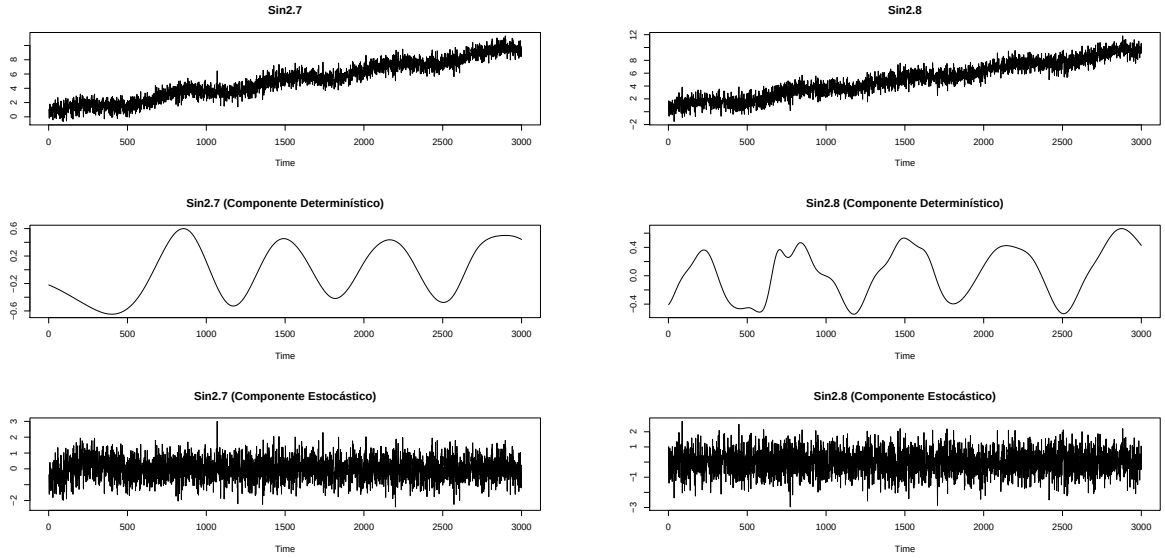


Figura A.22: Sin2.7 e Sin2.7

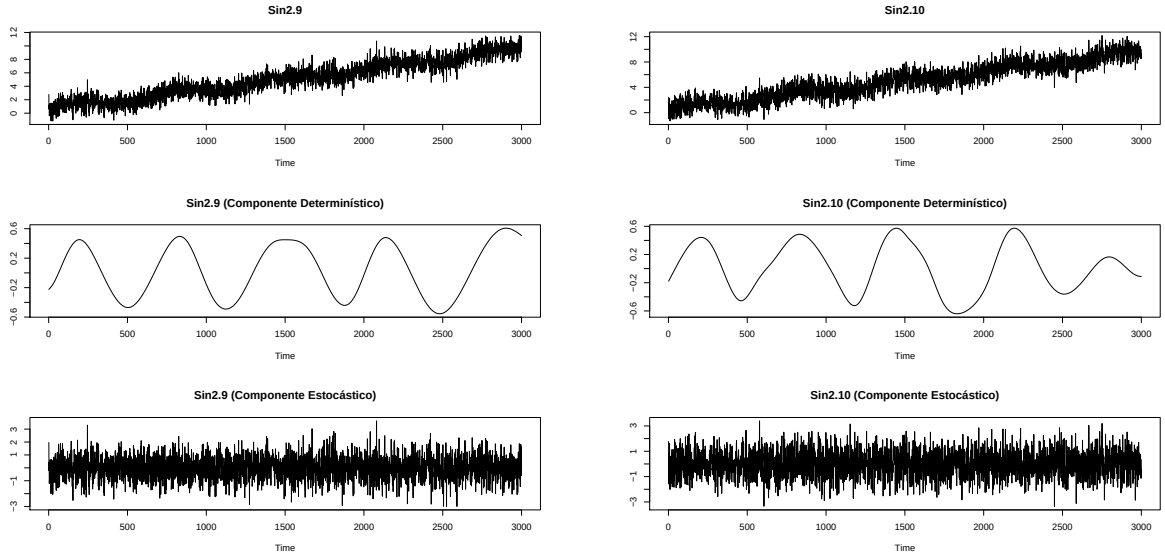


Figura A.23: Sin2.9 e Sin2.10

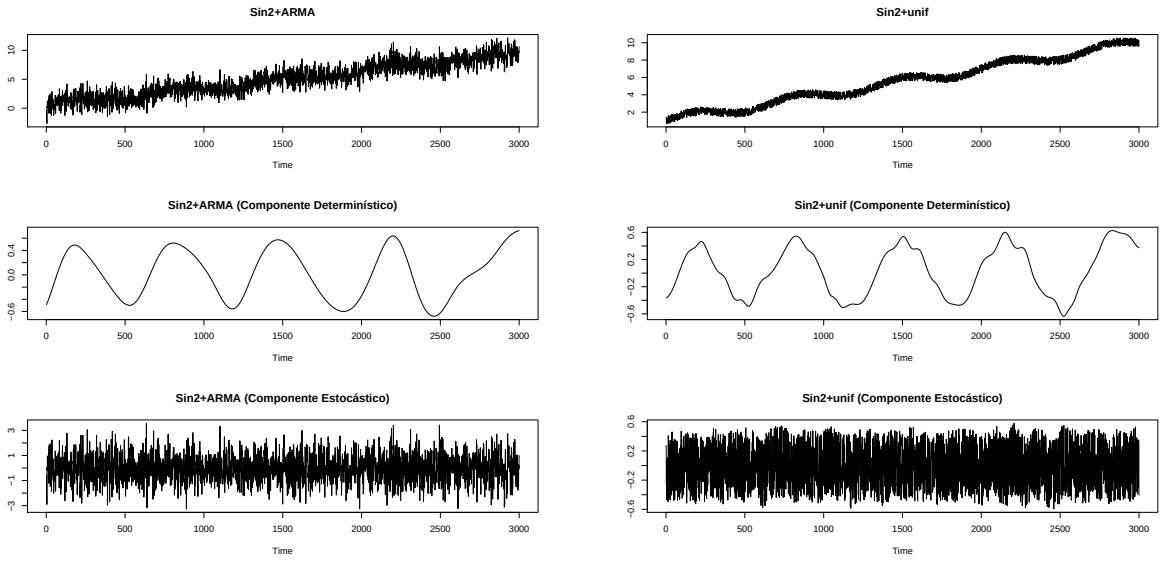


Figura A.24: Sin2+ARMA e Sin2+unif

A.5 SÉRIES LORENZ+RUÍDO

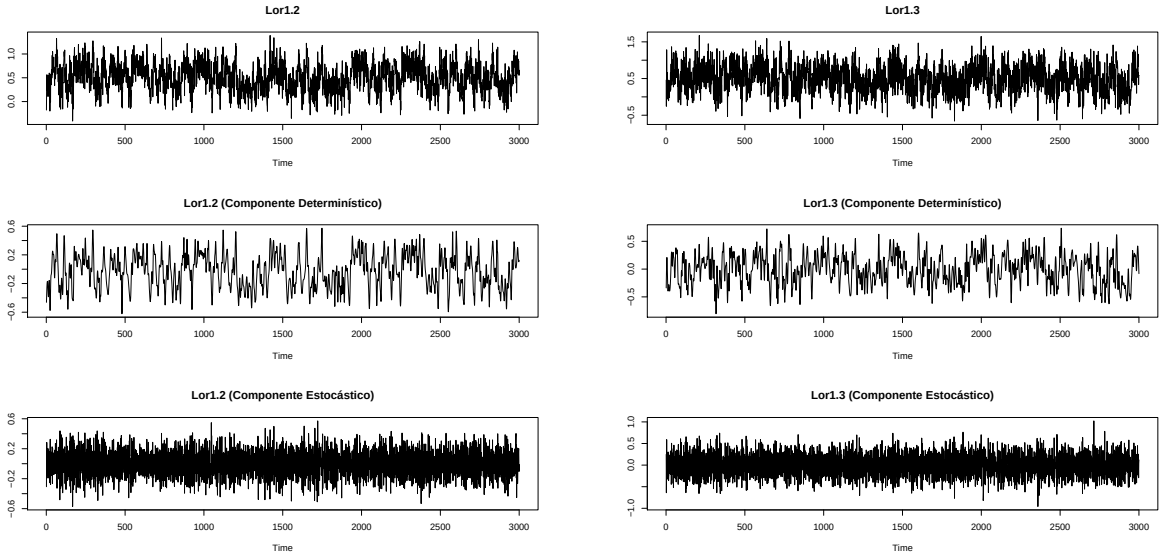


Figura A.25: Lor5.2 e Lor5.3

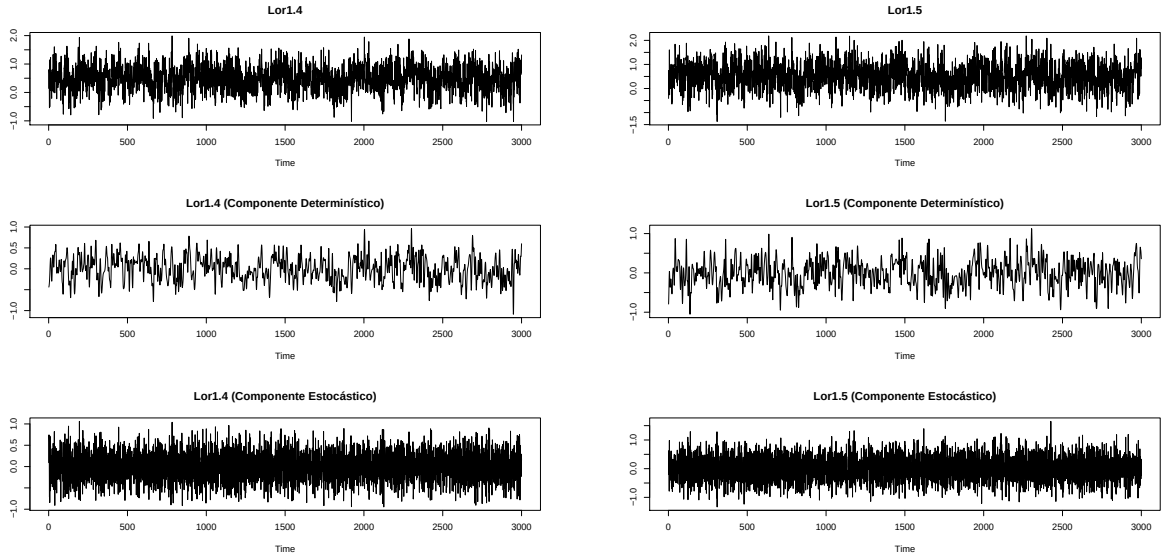


Figura A.26: Lor5.4 e Lor5.5

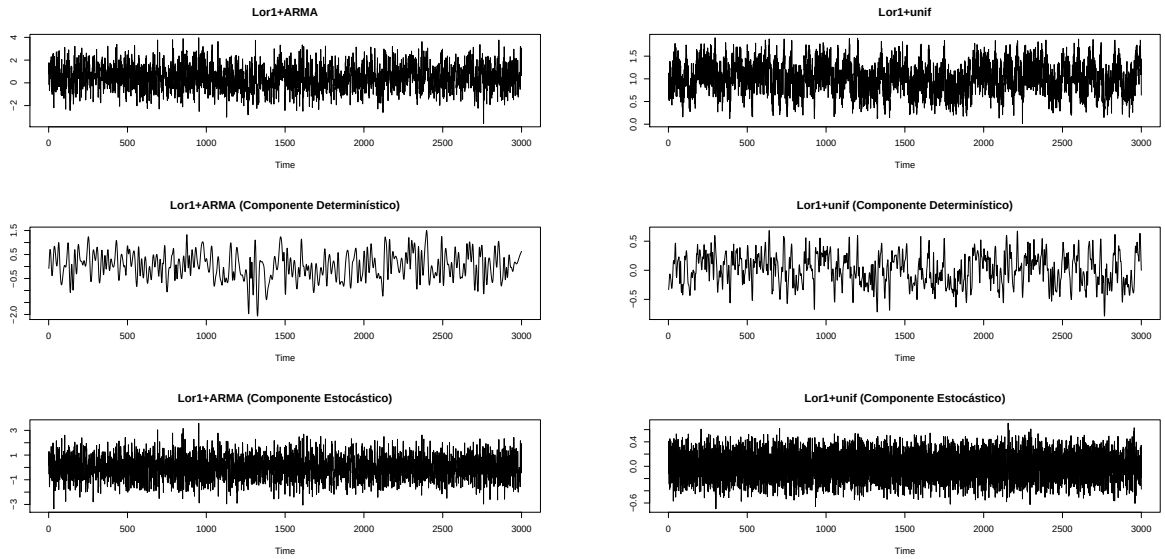


Figura A.27: Lor+ARMA e Lor+unif