UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

CARLO CORRÊA SOLCI

BOOTSTRAP LOCAL PARA SÉRIES ESTACIONÁRIAS INCOMPLETAS NA PRESENÇA DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO A PROBLEMAS NA ÁREA DA QUALIDADE DO AR

VITÓRIA
2022

CARLO CORRÊA SOLCI

BOOTSTRAP LOCAL PARA SÉRIES ESTACIONÁRIAS INCOMPLETAS NA PRESENÇA DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO A PROBLEMAS NA ÁREA DA QUALIDADE DO AR

Tese apresentada ao Programa de Pós-graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Doutor em Engenharia Ambiental, na área de concentração Poluição do Ar.
Orientador: Prof. Dr. Valdério Anselmo Reisen.
Coorientador: Prof. Dr. Paulo Jorge Canas Rodrigues.

VITÓRIA
2022

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Solci, Carlo Corrêa, 1990-
S684b Bootstrap local para séries estacionárias incompletas na presença de observações atípicas : uma aplicação a problemas na área da qualidade do ar / Carlo Corrêa Solci. - 2022.
108 f. : il.

Orientador: Valdério Anselmo Reisen.
Coorientador: Paulo Jorge Canas Rodrigues.
Tese (Doutorado em Engenharia Ambiental) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Ar - Poluição. 2. Análise de séries temporais. 3. Estatística robusta. I. Reisen, Valdério Anselmo. II. Rodrigues, Paulo Jorge Canas. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 628

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

BOOTSTRAP LOCAL PARA SÉRIES ESTACIONÁRIAS INCOMPLETAS NA PRESENÇA DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO A PROBLEMAS NA ÁREA DA QUALIDADE DO AR

Carlo Correa Solci

Banca Examinadora:

_____
Prof. Dr. Valdério Anselmo Reisen
Orientador - PPGEA/CT/UFES

_____
Prof. Dr. Paulo Jorge Canas Rodrigues
Coorientador Externo – UFBA

_____
Prof.ª Dr.ª Elisa Valentim Goulart
Examinador Interna – PPGEA/CT/UFES

_____
Prof. Dr. Neyval Costa Reis Jr.
Examinador Interno – PPGEA/CT/UFES

_____
Prof. Dr. Pascal Bondon
Examinador Externo – CentraleSupélec/França

_____
Prof.ª Dr.ª Glaura da Conceição Franco
Examinadora Externa – UFMG

_____
Prof. Dr. Felipe Elorrieta López
Examinador Externo – USACH/Chile

Elisa Valentim Goulart
Coordenadora do Programa de Pós-Graduação em Engenharia Ambiental
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Vitória/ES, 23 de setembro de 2022

Datas e horários baseados em Brasília, Brasil
Sincronizado com o NTP.br e Observatório Nacional (ON) em
29/09/2022 às 20:27:26

Carlo Solci PhD documents

Data e Hora de Criação: 27/09/2022 às 12:46:53

Documentos que originaram esse envelope:
- 63 - Documentos de defesa (PhD documents) - Carlo Correa Solci.docx (Documento Microsoft Word) - 5 página(s)
- documents Carlo Solci-PhD.docx (Documento Microsoft Word) - 1 página(s)

Hashs únicas referente à esse envelope de documentos

[SHA256]: 0daa42ea64cf6303052369cc92b00773d1a30fc6cfb9ac578e0fbf7db8d816ca
[SHA512]: 9af252248ceb8eaef4a162d9d50c6d519d06ce0bc9d184f8b517f44c08242aaca24919796e992a32f4636022df983868df31be32b669e51446af796b43493de8

Lista de assinaturas solicitadas e associadas à esse envelope

ASSINADO - Valderio Anselmo Reisen (valderioanselmoreisen@gmail.com)
Data/Hora: 27/09/2022 - 13:41:12, IP: 187.36.171.78, Geolocalização: [-20.278594, -40.297797]
[SHA256]: 18f82958792c342d7e1740cfc9161b898f6f7f62d744f1ec9566b36994c073e0

valderio anselmo reisen

ASSINADO - Pascal Bondon (pascal.bondon@centralesupelec.fr)
Data/Hora: 28/09/2022 - 09:58:17, IP: 78.192.100.135, Geolocalização: [48.8472576, 2.3560192]
[SHA256]: 573a71544a1c9b20f9e478f01a006536d1bb581b57e87d7614dbd33b6825db5e

ASSINADO - Felipe Elorrieta (felipe.elorrieta@usach.cl)
Data/Hora: 28/09/2022 - 10:14:29, IP: 152.230.59.54
[SHA256]: c9751b46b7348aa8135b8f7435b3eda9294dfcd37c3d49bc02436f97ae9f1d38

ASSINADO - Glaura Franco (glaura@est.ufmg.br)
Data/Hora: 28/09/2022 - 11:22:21, IP: 191.185.77.20
[SHA256]: f4bec5d2bd9d9964ba676b2664735a65984b563a732c38a50aa637b1a8406412

ASSINADO - Neyval Reis (neyval@gmail.com)
Data/Hora: 29/09/2022 - 13:05:14, IP: 177.133.92.73, Geolocalização: [-20.331984, -40.28204]
[SHA256]: 6a34113187c8d162a781d1b6f869ea2a88080517f9c7e82ba1b29bbde0f773a6

ASSINADO - Paulo Canas (paulocanas@gmail.com)
Data/Hora: 29/09/2022 - 20:24:56, IP: 186.222.179.25
[SHA256]: 3d83399a72db87284318ec13a5c284b331c2a62668d11aa0cc9c4d7866a8808e

ASSINADO - Elisa Valentin (elisavalentim@gmail.com)
Data/Hora: 29/09/2022 - 20:27:26, IP: 179.217.61.69
[SHA256]: e199d5161f670285d340f91686c0db76c306b69ad26d725178c7442fa7ccc151

Histórico de eventos registrados neste envelope

29/09/2022 20:27:26 - Envelope finalizado por elisavalentim@gmail.com, IP 179.217.61.69
29/09/2022 20:27:26 - Assinatura realizada por elisavalentim@gmail.com, IP 179.217.61.69
29/09/2022 20:24:56 - Assinatura realizada por paulocanas@gmail.com, IP 186.222.179.25
29/09/2022 13:05:14 - Assinatura realizada por neyval@gmail.com, IP 177.133.92.73
29/09/2022 13:05:10 - Envelope visualizado por neyval@gmail.com, IP 177.133.92.73
28/09/2022 11:22:21 - Assinatura realizada por glaura@est.ufmg.br, IP 191.185.77.20
28/09/2022 11:22:01 - Envelope visualizado por glaura@est.ufmg.br, IP 191.185.77.20
28/09/2022 10:14:29 - Assinatura realizada por felipe.elorrieta@usach.cl, IP 152.230.59.54
28/09/2022 10:05:47 - Envelope visualizado por felipe.elorrieta@usach.cl, IP 152.230.59.54
28/09/2022 09:58:18 - Assinatura realizada por pascal.bondon@centralesupelec.fr, IP 78.192.100.135
27/09/2022 13:41:12 - Assinatura realizada por valderioanselmoreisen@gmail.com, IP 187.36.171.78

Agradecimentos

"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."
(Albert Einstein)

Lista de Figuras

Lista de Tabelas

Resumo

Os estudos de poluição atmosférica geralmente envolvem medições e análises de dados de concentrações de poluentes, como é o caso do MP_{10} (material particulado de diâmetro inferior a 10 µm), do SO_2 (dióxido de enxofre) e de outros poluentes. Esses dados normalmente possuem características estatísticas importantes como autocorrelação, sazonalidade, observações faltantes e a presença de picos na série que apesar de não serem observações atípicas (outliers) pela alta frequência com a qual ocorrem, podem ser modelados como tais pelo efeito que têm na série. Todas essas características exigem atenção especial durante a análise dos dados e dificultam a obtenção de intervalos de confiança para os parâmetros de modelos de séries temporais estacionárias por meio de teoria assintótica. Com essa motivação, este estudo propôs metodologias de bootstrap no domínio da frequência para séries temporais fracamente estacionárias na presença de observações faltantes e/ou de contaminação por observações atípicas aditivas. As metodologias sugeridas são baseadas no bootstrap local de Paparoditis & Politis (1999), com a robustez sendo atingida por meio de substituição do periodograma clássico pelo M-periodograma de Reisen, Lévy-Leduc & Taqqu (2017) e quando há presença de observações faltantes se substitui a série temporal original pela sua versão de amplitude modulada proposta por Parzen (1963). Nesse contexto, a eficiência das metodologias de bootstrap propostas em estimar intervalos de confiança de parâmetros de modelos para séries temporais fracamente estacionárias foi verificada por meio de estudos de Monte Carlo em diferentes cenários, incluindo: contaminação por observações atípicas aditivas e presença de observações faltantes. Para efeito de comparação, em alguns casos também foi considerada a metodologia de bootstrap de Paparoditis & Politis (1999), bem como as estimativas dos parâmeteros sem o bootstrap pelas versões clássica e robusta das metodologias de Whittle (1953) e de Dunsmuir & Robinson (1981). O interesse prático em poluição do ar é avaliar se os intervalos de confiança dos parâmetros obtidos pelas metodologias robustas apresentam uma redução do efeito de deslocamento para a esquerda que os intervalos clássicos possuem devido à perda de memória causada pelas observações atípicas aditivas, além da possibilidade de calcular esses intervalos sem a utilização de técnicas de imputação para obter uma série temporal completa. As metodologias de bootstrap propostas foram aplicadas para calcular intervalos de confiança de parâmetros de ajuste do modelo autorregressivo (AR), e em alguns casos também do modelo sazonal autorregressivo (SAR), à dados de MP_{10} de estações da rede de monitoramento da qualidade do ar da Região da Grande Vitória - ES.

Palavras-chave: poluição do ar, análise de séries temporais, análise espectral, bootstrap, periodograma, observações faltantes, modulação de amplitude, observações atípicas, robustez e MP_{10}.

Abstract

Studies about air pollution typically involve measurements and analysis of pollutants, such as PM_{10} (particulate matter with diameter lower than 10 µm), SO_2 (sulfur dioxide) and others. These data typically have important features like serial correlation, seasonality, missing observations and the presence of peaks that despite not being atypical observations (outliers) because of their high frequency of occurrence, can be modeled as such owing to the effect that they have on the series. All these features demand special attention during data analysis and complicate the obtainment of confidence intervals for the parameters of stationary time series models through asymptotic theory. With this motivation, this study proposed bootstrap methodologies in the frequency domain for weakly stationary time series in the presence of missing observations and/or of contamination by additive outliers. The suggested methodologies are based on the local bootstrap of Paparoditis & Politis (1999), with the robustness being achieved by the substitution of the classical periodogram with the M-periodogram of Reisen, Lévy-Leduc & Taqqu (2017) and when there is presence of missing observations the original time series is replaced by its amplitude modulated version proposed by Parzen (1963). In this context, the efficiency of the proposed bootstrap methodologies in estimating confidence intervals of parameters of models for weakly stationary time series was verified through Monte Carlo studies under different scenarios, including: additive outliers contamination and presence of missing observations. For comparison purposes, in some cases it was also considered the bootstrap methodology of Paparoditis & Politis (1999), as well as the parameter estimates without the bootstrap via the classical and robust versions of the methodologies of Whittle (1953) and of Dunsmuir & Robinson (1981). The practical purpose in air pollution is to evaluate if the confidence intervals of the parameters obtained by the robust methodologies present a reduction in the effect of left shift that the classical intervals have due to the memory loss caused by the additive outliers, in addition to the possibility of calculating these intervals without using imputation techniques to obtain a complete time series. The proposed bootstrap methodologies were applied to calculate confidence intervals of parameters of adjustment of the autoregressive (AR) model, and in some cases also of the seasonal autoregressive (SAR) model, to MP_{10} data of stations of the air quality monitoring network of the Greater Vitória Region - ES.

Keywords: air pollution, time series analysis, spectral analysis, bootstrap, periodogram, missing observations, amplitude modulation, outliers, robustness and PM_{10}.

Sumário

1 Introdução

A qualidade do ar de uma região é determinada pela proximidade de fontes poluidoras e pelos níveis de emissão dos poluentes das mesmas, pela capacidade com que a atmosfera da região consegue absorver, dispersar e remover esses contaminantes. Parâmetros meteorológicos como temperatura, pressão atmosférica, umidade relativa do ar, velocidade e direção predominante do vento podem ampliar ou reduzir a capacidade do transporte e da dispersão atmosférica dos poluentes.

As emissões de poluentes atmosféricos podem ser classificadas em antropogênicas e naturais (GODISH, 1997). As emissões antropogênicas são aquelas provocadas pela ação do homem, geralmente nas indústrias, nos transportes e em processos de geração de energia. As emissões naturais são causadas por processos naturais, tais como queimadas naturais, partículas do solo ressuspensas pelo vento, etc. Quanto à origem, os poluentes são classificados em níveis primários ou secundários. Os poluentes primários são aqueles lançados diretamente na atmosfera, como resultado dos processos industriais, dos gases de exaustão dos motores de combustão interna, construção civil e outros. Os poluentes secundários são aqueles formados a partir de reações químicas que ocorrem na atmosfera entre os poluentes primários.

Um dos poluentes atmosféricos mais relevantes é o material particulado (MP), devido ao efeito danoso à saúde das pessoas, dos animais e da vegetação, pela interferência nas mudanças climáticas regionais e globais, e pelo incômodo de sua deposição nas superfícies dos materiais e edificações (WHO, 2005; JACOBSON, 2002). A gravidade dos danos causados pelo MP levou à uma maior abordagem sobre o assunto na literatura e ao crescimento do monitoramento de fontes de emissões e de dados da qualidade do ar em várias regiões do mundo.

O material particulado é composto de partículas capazes de permanecer em suspensão na atmosfera devido às suas pequenas dimensões. Como exemplos podem ser citados a poeira, a fuligem e as partículas de óleo (BRAGA et al., 2005). A forma e a composição química do MP podem ser bastante diversificadas. Normalmente a classificação é feita de acordo com o tamanho da partícula: nos casos de diâmetros aerodinâmicos inferiores a 2,5 µm e 10 µm são denominados $MP_{2,5}$ e MP_{10}, respectivamente. Na literatura, o MP_{10} também é definido como partículas inaláveis. Segundo Baird (2002), as partículas são classificadas como grossas (diâmetro maior que 2,5 µm) e finas (diâmetro menor que 2,5 µm). A importância do tamanho das partículas está relacionada aos danos que elas podem causar à saúde. Holgate et al. (1999) afirmam que as partículas finas são as principais responsáveis por esses danos, uma vez que podem atingir e prejudicar o sistema respiratório inferior.

Tendo em vista os efeitos do material particulado na saúde e no meio ambiente, com o passar do tempo, têm surgido legislações ambientais para regulamentar os níveis de emissões e de qualidade do ar. A legislação brasileira, através da resolução CONAMA nº 491 de 2018 (CONAMA, 2018), estabeleceu, respectivamente, os seguintes padrões intermediários (PI-1, que estão em vigor) de concentração de partículas inaláveis para longa e curta exposição: (1) a concentração média aritmética anual deve ser de, no máximo, $40\,\mu g/m^3$; e (2) a concentração média de 24 horas deve ser de, no máximo, $120\,\mu g/m^3$. A Organização Mundial de Saúde (OMS) estabeleceu, no ano de 2005, as diretrizes de $50\,\mu g/m^3$ para a concentração média de 24 horas e de $20\,\mu g/m^3$ para a média aritmética anual de MP_{10} (WHO, 2005).

Para auxiliar a manutenção dos níveis de concentrações dentro dos padrões da legislação, é primordial a realização do monitoramento da qualidade do ar. Na região metropolitana da Grande Vitória (RMGV), a Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), é responsável por desempenhar esse papel. A rede foi inaugurada em julho de 2000 pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA) e fornece dados horários meteorológicos e de diversos poluentes, sendo que o poluente MP_{10} é monitorado por todas as estações que compõem a RAMQAr.

Esta disponibilidade de dados monitorados sequencialmente no tempo, aliada à necessidade de avaliar a qualidade do ar e de fazer previsões de concentrações desses poluentes, justificam a abordagem de séries temporais para os dados de poluição atmosférica coletados pela RAMQAr. A escolha da RMGV se justifica pelos seguintes fatos sobre a mesma: possui diversas fontes de emissão de poluentes atmosféricos devido à suas indústrias e ao seu crescente desenvolvimento urbano; no que se refere ao efeito dos poluentes na saúde da população, houve um aumento do número de atendimentos hospitalares por doenças respiratórias e cardiovasculares em função de seu crescimento (SOUZA et al., 2014).

Na prática, são utilizados modelos para se entender a dinâmica e prever valores futuros de séries temporais. Nesse contexto, é essencial ter conhecimento das incertezas que estão associadas à estimação dos parâmetros desses modelos, o que pode ser atingido por meio de técnicas de reamostragem. No caso das metodologias clássicas de séries temporais, é necessário que se tenha uma série completa para ajustar modelos, hipótese essa que não é aplicável aos dados de concentração de poluentes atmosféricos da RMGV uma vez que os mesmos têm presença de observações faltantes.

Dentre as diferentes técnicas de reamostragem existentes destaca-se o bootstrap que foi criado originalmente por Efron (1979) para estudar observações que são independentes. Posteriormente foram criadas diversas técnicas de bootstrap para séries temporais que podem ser divididas nas metodologias nos domínios do tempo e da frequência.

Metodologias de bootstrap no domínio do tempo:

- Bootstrap em blocos.

- Boootstrap markoviano.
- Bootstrap não-paramétrico residual.
- Bootstrap AR-sieve.
- Bootstrap residual para modelos paramétricos.

Metodologias de bootstrap no domínio da frequência:

- Bootstrap multiplicativo residual.
- Boootstrap local.
- Bootstrap híbrido.

O fato das séries temporais poderem ser decompostas em diversos ciclos de frequências e amplitudes diferentes através de uma série de Fourier é uma oportunidade para analisá-las no domínio da frequência. O uso da decomposição de séries temporais via análise espectral, surgiu como alternativa para identificação da amplitude de cada ciclo correspondente a uma determinada frequência. A função densidade espectral, ou simplesmente espectro, mostra a decomposição da variância de uma amostra de dados através de diferentes frequências. Assim, o espectro descreve as propriedades cíclicas de uma determinada série temporal. Supõe-se que as flutuações do processo subjacente são produzidas por um grande número de ciclos elementares de diferentes frequências e que a contribuição de cada ciclo é constante em toda a amostra. O espectro fornece então a contribuição relativa de cada um desses ciclos elementares para a variância do processo global. O periodograma é uma ferramenta muito útil na decomposição de uma série temporal em componentes cíclicas com diferentes frequências e amplitudes. Isso é particularmente útil quando se deseja estimar o período de uma série, pois o mesmo é dado pela frequência que possui a maior amplitude (frequência dominante). No entanto, a versão clássica do periodograma possui a desvantagem de ser muito sensível à presença de outliers, e portanto, ela se torna sem utilidade em situações nas quais os dados reais são contaminados por observações atípicas. Outra limitação que a versão clássica do periodograma possui é de não ser adequada para utilização quando a série possui observações faltantes.

A metodologia do bootstrap proposta por Efron (1979) que envolve amostragens com reposição, pode ser aplicada às ordenadas do periodograma devido ao fato das mesmas possuírem independência assintótica. Só que no caso do periodograma é mais interessante fazer uma reamostragem "local" ao invés da reamostragem "global" do bootstrap comum. Isso se deve ao fato da distribuição das ordenadas do periodograma não ser constante mas sim mudar de forma lenta com o seu índice. Assim sendo, Paparoditis & Politis (1999) propuseram o bootstrap local para o periodograma que pode ser utilizado na estimação de intervalos de confiança de estatísticas baseadas no periodograma tais como os parâmetros

de um modelo para uma série temporal fracamente estacionária. No entanto, ainda falta a generalização do bootstrap local para estimação de parâmetros de modelos de séries temporais quando as mesmas têm presença de observações atípicas aditivas e de dados faltantes.

Nesse contexto, um problema típico das séries temporais de poluentes atmosféricos na RMGV é a presença de picos na série que apesar de não poderem ser considerados observações atípicas por causa da alta frequência com a qual ocorrem na série, podem ser modelados como observações atípicas aditivas já que provocam um efeito na série similar ao causado por tal tipo de observação, isto é, têm efeito de observações que fogem do padrão da série e que não têm nenhum efeito nas observações subsequentes (FOX, 1972). Esses picos fazem com que haja perda de memória nas estimativas dos parâmetros de modelos de séries temporais fracamente estacionárias e, assim, faz com que os intervalos de confiança associados à esses parâmetros sejam deslocados para a esquerda. Além disso, as séries temporais de poluentes atmosféricos na RMGV possuem observações faltantes que fazem com que seja necessário o uso de técnicas de imputação para tornar essas séries completas.

Neste estudo, serão utilizadas médias diárias do MP_{10} para garantir similaridade com o disposto na resolução CONAMA nº 491 de 2018. Na prática, a ocorrência de observações faltantes em médias diárias de MP_{10}, pode ser justificada, na maior parte dos casos, por falhas dos equipamentos de monitoramento. Já a presença de picos de concentração se deve a fugas diárias dos padrões de emissões de fonte e/ou meteorológicos. Os picos causam um problema que é recorrente na área de poluição do ar: deteriorações da qualidade do ar de uma região nos intervalos de tempo em que os mesmos ocorrem. Adicionalmente, esses picos podem prejudicar a eficiência do bootstrap local em modelos para séries temporais fracamente estacionárias, como mencionado previamente.

Para superar essas limitações e, visando propor uma técnicas que ofereçam maior eficiência nas aplicações empíricas, nesta tese sugerem-se três metodologias de bootstrap no domínio da frequência para estimar intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias. A primeira metodologia, que é abordada no primeiro artigo desta tese, propõe a substituição do periodograma clássico pelo M-periodograma robusto de Reisen, Lévy-Leduc & Taqqu (2017) no bootstrap local de Paparoditis & Politis (1999) com o intuito de reduzir o efeito de deslocamento para a esquerda dos intervalos de confiança dos parâmetros de modelos de séries temporais fracamente estacionárias quando há contaminação por observações atípicas aditivas. Um estudo empírico será conduzido para comparar o desempenho da metodologia robusta proposta com o da metodologia clássica de Paparoditis & Politis (1999) na estimação de intervalos de confiança de parâmetros de modelos AR e SAR em cenários com e sem contaminção por observações atípicas aditivas.

A segunda metodologia proposta neste trabalho, que é abordada no segundo artigo

desta tese, propõe a substituição da série temporal original pela sua versão de amplitude modulada proposta por Parzen (1963) no bootstrap local de Paparoditis & Politis (1999). Enquanto que a terceira metodologia, que também é abordada no segundo artigo desta tese, propõe a substituição da série temporal original pela sua versão de amplitude modulada proposta por Parzen (1963) e a substituição do periodograma clássico pelo M-periodograma robusto de Reisen, Lévy-Leduc & Taqqu (2017) no bootstrap local de Paparoditis & Politis (1999). Um estudo empírico será conduzido para comparar o desempenho das metodologias propostas na estimação de intervalos de confiança de parâmetros de modelos AR em cenários com: séries completas e sem contaminação por observações atípicas aditivas; dados incompletos e sem contaminação por outliers aditivos; e séries incompletas e com contaminação por observações atípicas aditivas.

O restante desta tese está dividido da seguinte forma: no Capítulo 2 são apresentados o objetivo geral e os específicos; no Capítulo 3 é feita uma revisão da literatura referente à área poluição atmosférica e ao bootstrap aplicado à essa área; no Capítulo 4 são exibidos os materiais e métodos utilizados; no Capítulo 5 são mostrados os principais resultados e discussões, compilados no formato de dois artigos científicos; no Capítulo 6 são elaboradas as conclusões; enquanto dois estudos adicionais (um está compilado no formato de artigo científico e o outro é um artigo em compilação) são apresentados no Apêndice A.

2 Objetivos

2.1 Objetivo Geral

O objetivo geral desta tese é propor metodologias de bootstrap para obtenção de intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias na presença de dados faltantes e/ou de observações atípicas, para aplicação em dados de material particulado inalável da Região Metropolitana da Grande Vitória.

2.2 Objetivos Específicos

Os objetivos específicos desta tese são os seguintes:

- Obter uma metodologia robusta para estimar intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias por meio do bootstrap local de Paparoditis & Politis (1999) com a substituição do periodograma clássico pelo M-periodograma de Reisen, Lévy-Leduc & Taqqu (2017);

- Desenvolver uma metodologia clássica para estimar intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias na presença de observações faltantes por meio da aplicação do bootstrap local de Paparoditis & Politis (1999) no periodograma da versão de amplitude modulada da série;

- Obter uma metodologia robusta para estimar intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias na presença de observações faltantes por meio da aplicação do bootstrap local de Paparoditis & Politis (1999) no M-periodograma de Reisen, Lévy-Leduc & Taqqu (2017) da versão de amplitude modulada da série;

- Aplicar os estimadores de intervalos de confiança de parâmetros de modelos às séries temporais de material particulado inalável da rede de monitoramento da qualidade do ar da Região da Metropolitana da Grande Vitória.

3 Revisão Bibliográfica

3.1 Poluição Atmosférica

A Resolução nº 491 de novembro de 2018 do Conselho Nacional do Meio Ambiente (CONAMA, 2018), caracteriza como poluente atmosférico qualquer forma de matéria em quantidade, concentração, tempo ou outras características, que tornem ou possam tornar o ar: i) impróprio ou nocivo à saúde; ii) inconveniente ao bem-estar público; iii) danoso aos materiais, à fauna e à flora; ou, iv) prejudicial à segurança, ao uso e ao gozo da propriedade e às atividades normais da comunidade.

No que tange às emissões de poluentes, essas podem ser classificadas em antropogênicas e naturais. Quanto às antropogênicas, as mesmas são decorrentes das ações do homem (indústria, transporte, geração de energia e outras). Já as naturais originam-se de processos naturais, como emissões vulcânicas e processos microbiológicos. Além disso, os poluentes podem ser classificados, segundo a sua origem, em primários e secundários. Os primários são lançados diretamente na atmosfera pelas fontes de emissão, como por exemplo: SO_2, CO, NO_X e hidrocarbonetos (HC). Já os secundários formam-se na atmosfera por meio da reação química entre poluentes primários ou desses com constituintes naturais da atmosfera. Podem-se citar como exemplos: O_3 e peróxido de hidrogênio (H_2O_2).

A seguir encontra-se uma breve descrição dos principais poluentes atmosféricos.

- Dióxido de enxofre (SO_2): gás tóxico e incolor, pode ser emitido por fontes naturais ou por fontes antropogênicas e pode reagir com outros compostos na atmosfera, formando material particulado de diâmetro reduzido;

- Dióxido de nitrogênio (NO_2): gás poluente altamente oxidante, sendo que sua presença na atmosfera é fator proponderante na formação do ozônio troposférico;

- Hidrocarbonetos (HC): compostos formados de carbono e hidrogênio e que podem se apresentar na forma de gases, partículas finas ou gotas;

- Material particulado (MP): mistura complexa de sólidos com diâmetro reduzido. Seus componentes apresentam características físicas e químicas variadas. Geralmente o material particulado é classificado de acordo com o diâmetro das partículas em função da relação existente entre diâmetro e possibilidade de penetração no trato respiratório. É importante lembrar que nos casos de diâmetros aerodinâmicos inferiores a 2,5 µm e 10 µm o material particulado é denominado $MP_{2,5}$ ou partículas respiráveis e MP_{10} ou partículas inaláveis, respectivamente;

- Monóxido de carbono (CO): gás inodoro e incolor, é formado no processo de queima de combustíveis;
- Ozônio (O_3): poluente secundário, é formado a partir de outros poluentes atmosféricos e é altamente oxidante na troposfera (camada inferior da atmosfera).

Vale frisar que em relação aos padrões de qualidade do ar existe uma diferença entre os valores estabelecidos pela Resolução CONAMA nº 491 de 2018, em vigor (PI-1), e as diretrizes da OMS (revisadas em 2005). Além disso, o Governo do Estado do Espírito Santo, por meio do Decreto no 3463-R de 2013, estabeleceu padrões de qualidade do ar para o Espírito Santo que também apresentam diferenças quando à Resolução CONAMA nº 491 de 2018. Na Tabela 3.1 são apresentados os padrões nacionais e estaduais de qualidade do ar e as diretrizes da OMS. Pode-se verificar a grande discrepância que há entre os padrões nacionais em vigor e as diretrizes da OMS.

3.2 Estado da Arte em Estimação de Intervalos de Confiança de Parâmetros de Modelos de Séries Temporais de Poluição do Ar

Para o gerenciamento da qualidade do ar é necessário conhecer as concentrações de poluentes e gerar previsões satisfatórias delas. A utilização de modelos de previsão é uma ferramenta importante para conhecer o comportamento e características de determinados poluentes, podendo, desta forma, prever possíveis picos de concentração. Para isto, pode-se fazer uso de duas classes de modelos, os experimentais e os matemáticos. Nesta última, têm-se os modelos determinísticos e os modelos estocásticos. O presente estudo se concentrará na classe de modelagem estocástica. Uma vez que o objetivo desta pesquisa é propor metodologias de bootstrap para obtenção de intervalos de confiança de parâmetros de modelos de séries temporais fracamente estacionárias com dados faltantes e/ou observações atípicas, para aplicação em dados de poluentes atmosféricos, esta seção visa descrever alguns estudos que aplicaram técnicas estatísticas para lidar com a presença de dados faltantes, analisar séries temporais no domínio da frequência (análise espectral), tratar observações atípicas aditivas (estatística robusta) e estimar intervalos de confiança via bootstrap.

Dunsmuir & Robinson (1981) propuseram uma metodologia para estimação de modelos para séries temporais discretas na presença de dados faltantes. Algumas justificativas foram dadas para a utilização dessa metodologia sobre alternativas à ela; a escolha do estimador geralmente é governada pelo padrão dos dados faltantes, pela natureza do modelo de séries temporais, e por considerações computacionais. A performance da metodologia na estimação de modelos simples foi estudada pelas simulações, e foi aplicada a séries temporais de médias diárias em partes por milhão do poluente atmosférico CO medidas

Tabela 3.1 – Padrões nacionais e estaduais de qualidade do ar e diretrizes da OMS

			$MP_{2.5}$ [$\mu g/m^3$]	MP_{10} [$\mu g/m^3$]	PTS [$\mu g/m^3$]	PS [g/m^2. 30 dias]	SO_2 [$\mu g/m^3$]	NO_2 [$\mu g/m^3$]	O_3 [$\mu g/m^3$]	CO [ppm]	FUMAÇA [$\mu g/m^3$]	Pb[a] [$\mu g/m^3$]
Padrões de Qualidade do Ar (CONAMA nº 491/2018)	Curta exposição	PI-1	60 (24h)	120 (24h)	-	-	125 (24h)	260 (1h)[b]	140 (8h)[c]	-	120 (24h)	-
		PI-2	50 (24h)	100 (24h)	-	-	50 (24h)	240 (1h)[b]	130 (8h)[c]	-	100 (24h)	-
		PI-3	37 (24h)	75 (24h)	-	-	30 (24h)	220 (1h)[b]	120 (8h)[c]	-	75 (24h)	-
		PF	25 (24h)	50 (24h)	240 (24h)	-	20 (24h)	200 (1h)[b]	100 (8h)[c]	09 (8h)[c]	50 (24h)	-
	Longa exposição	PI-1	20 (ano)[d]	40 (ano)[d]	-	-	40 (ano)[d]	60 (ano)[d]	-	-	40 (ano)[d]	-
		PI-2	17 (ano)[d]	35 (ano)[d]	-	-	30 (ano)[d]	50 (ano)[d]	-	-	35 (ano)[d]	-
		PI-3	15 (ano)[d]	30 (ano)[d]	-	-	20 (ano)[b]	45 (ano)[d]	-	-	30 (ano)[d]	-
		PF	10 (ano)[d]	20 (ano)[d]	80 (ano)[e]	-	-	40 (ano)[d]	-	-	20 (ano)[d]	0,5 (ano)[d]
Metas e padrão estadual (Decreto nº 3463-R 2013)	Curta exposição	MI1-ES	-	120 (24h)	180 (24h)	14	60 (24h)	240 (1h)	140 (8h)	-	-	-
		MI2-ES	50 (24h)	80 (24h)	170 (24h)	-	40 (24h)	220 (1h)	120 (8h)	-	-	-
		MI3-ES	37 (24h)	60 (24h)	160 (24h)	-	30 (24h)	210 (1h)	110 (8h)	-	-	-
		PF-ES	25 (24h)	50 (24h)	150 (24h)	-	20 (24h)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-	-
	Longa exposição	MI1-ES	-	45 (ano)[d]	65 (ano)[e]	-	40 (ano)[d]	50 (ano)[d]	-	-	-	-
		MI2-ES	20 (ano)[d]	33 (ano)[d]	63 (ano)[e]	-	30 (ano)[d]	45 (ano)[d]	-	-	-	-
		MI3-ES	15 (ano)[d]	25 (ano)[d]	62 (ano)[e]	-	20 (ano)[d]	42 (ano)[d]	-	-	-	-
		PF-Es	10 (ano)[d]	20 (ano)[d]	60 (ano)[e]	-	-	40 (ano)[d]	-	-	-	
Diretriz OMS	Curta exposição		25 (24h)	50 (24h)	-	-	20 (24h) 500 (10min)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-	-
	Longa exposição		10 (ano)[d]	20 (ano)[d]	-	-	-	40 (ano)[d]	-	-	-	-

Nota: 1) O tempo de média considerado para o cálculo da concentração do poluente está indicado entre parênteses; e, 2) [a] Medido nas partículas totais em suspensão; [b] Média horária; [c] Máxima média móvel obtida no dia; [d] Média aritmética anual; [e] Média geométrica anual.
Fonte: Iema (2014), OMS (2005), CONAMA (2018).

na cidade de Boston, nos Estados Unidos, pelo período de 114 semanas começando no dia 03/06/1969.

Iglesias, Jorquera & Palma (2006) propuseram uma metodologia estatística para manipular regressões com longa dependência nos erros e dados faltantes. A estratégia de

estimação foi desenvolvida através das abordagens Bayesiana e clássica. O estudo foi ilustrado com aplicação em um conjunto de dados de concentrações de poluentes atmosféricos da cidade de Santiago, no Chile, para o período de 01/01/1989 a 31/12/1996, com um número muito alto de observações faltantes: 531 dias sem observação. Para correção dos dados faltantes, os autores utilizaram o filtro de Kalman. A fim de explicar a variação nas concentrações de MP_{10}, os autores utilizaram como variáveis explanatórias da regressão a velocidade do vento, a precipitação e as concentrações de CO_2 e de SO_2. Essas variáveis mostraram-se correlacionadas com a concentração de MP_{10}. De acordo com os resultados apresentados, a aplicação da metodologia aos dados reais, com a abordagem clássica, mostrou que a inferência pode ser distorcida se a longa dependência nos erros não for considerada.

Gómez-Carracedo et al. (2014) utilizaram um conjunto de dados de qualidade do ar (NO, NO_2, NO_X, CO, O_3, MP_{10}, $MP_{2,5}$ e $MP_{1,0}$) de uma estação de imersão automática situada numa zona costeira suburbana próxima da cidade de Corunha, na Espanha, com taxas de dados faltantes variando de 4% a 24%, para verificar se grandes diferenças ocorrem quando métodos de imputação distintos (Média Incondicional, Mediana Modificada, Método Baseado em Componentes Principais, *Expectation Maximization* (EM) e Imputação Múltipla), são aplicados para o preenchimento dos dados faltantes. Todos os métodos foram executados de forma semelhante, embora a Imputação Múltipla tenha gerado valores imputados mais dispersos. De acordo com os resultados apresentados, as principais diferenças ocorreram quando uma variável com valores ausentes teve correlação fraca com as outras características e quando uma variável apresentou carregamentos relevantes em vários fatores não rotacionados, o que algumas vezes alterou a ordem dos fatores rotacionados. Os melhores resultados foram obtidos com o algoritmo EM.

Junger & Leon (2015) discutiram questões teóricas e metodológicas para imputação de dados faltantes em séries temporais multivariadas de poluentes atmosféricos. Os autores apresentaram um método baseado em imputação que usa o algoritmo EM sob a suposição de distribuição normal. Diferentes abordagens foram consideradas para a filtragem da componente temporal. Foi realizado um estudo de simulação para avaliar a validade e o desempenho do método proposto em comparação com alguns métodos frequentemente utilizados. Cada método de imputação e configuração foi comparado com um valor de referência, que foi estimado utilizando a análise do conjunto de dados completo que compreendeu as 366 observações do ano de 2004 de concentrações diárias de MP_{10} medidas em dez estações de monitoramento da cidade de São Paulo, no Brasil. As simulações mostraram que, quando a quantidade de dados faltantes era de apenas 5%, a análise completa dos dados produziu resultados satisfatórios independentemente do mecanismo gerador dos dados ausentes, enquanto a validade começou a degenerar quando a proporção de valores faltantes excedia 10%. Segundo os resultados, o método de imputação proposto pelos autores apresentou boa acurácia e precisão em diferentes contextos com relação aos

padrões de observações ausentes. A maioria das imputações obteve resultados válidos, mesmo sob ausência não aleatória. Os métodos propostos no trabalho foram implementados como um pacote denominado *multivariate time series data imputation* (mtsdi) no software estatístico R.

Miller et al. (2018) avaliaram quatro métodos de imputação de dados faltantes para abordar os dados de concentração de BTEX (benzeno, tolueno, etilbenzeno e xilenos), medidos em Windsor e Sarnia, Ontário, Canadá, no outono de 2005. Os métodos de imputação avaliados foram: imputação do valor médio, ponderação inversa da distância, proporções interespécies e regressão. As concentrações e relações entre espécies geralmente foram similares entre as duas cidades. Utilizando essas cidades industrializadas como estudos de caso, os autores demonstraram que a utilização das técnicas de proporções interespécies ou de regressão dos dados para os quais há informação completa, junto com uma concentração medida (i.e. benzeno) para prever concentrações perdidas (i.e. TEX) resulta em uma boa concordância entre os valores previstos e os medidos. Os autores apontam que, na ausência de quaisquer concentrações conhecidas, o método ponderação inversa da distância pode fornecer uma concordância razoável entre as concentrações observadas e estimadas para as espécies BTEX, e foi superior à imputação de valor médio que não foi capaz de preservar a tendência espacial.

Hies et al. (2000) apresentaram um método eficaz para analisar diferentes fontes de poluição do ar em uma série temporal de carbono elementar. Como uma segunda função, essa técnica permite uma classificação rápida e eficiente dos locais de monitoramento. Séries temporais diárias de medições de carbono elementar em várias localidades urbanas de Berlim na Alemanha de 01/04/1994 a 31/03/1995 foram avaliadas com suas correspondentes densidades espectrais estimadas por meio de uma versão suavizada do periodograma. Periodicidades típicas e bem conhecidas causadas por influências antropogênicas e meteorológicas foram identificadas pelos espectros de coerência e de fase. Chegou-se a conclusão de que como as amplitudes relativas das várias influências variam dependendo de onde os locais de monitoramento se situam na área urbana, o uso da densidade espectral estimada ajuda a achar a influência do tráfego, do aquecimento doméstico por carvão e do transporte de longo alcance na concentração de carbono elementar.

Sebald et al. (2000) utilizaram análise espectral para investigar os processos de formação e de decomposição do ozônio troposférico. Apesar do ozônio ser um poluente atmosférico reativo e secundário, uma abordagem similar à utilizada por Hies et al. (2000) para o carbono elementar pode ser aplicada à série temporal de ozônio. As séries temporais horárias de ozônio em várias localidades da Alemanha dos meses de abril a setembro dos anos de 1993 a 1995 foram divididas em componentes sazonais de baixa e alta frequências para detectar a razão para concentrações extremamente altas de ozônio no verão. A componente de alta frequência foi avaliada utilizando a correspondente densidade espectral estimada por meio do periodograma suavizado. Foi demonstrado que as flutuações me-

teorológicas de escalas grande e sinótica afetam as concentrações de ozônio em todos os locais de monitoramento. Os resultados obtidos afirmaram o domínio de flutuações espaciais homogêneas e indicaram uma camada residual relativamente uniforme sobre uma grande região.

Choi et al. (2008) investigaram a variação semanal antropogênica dependente da região em poluentes atmosféricos e sua relação com as condições meteorológicas da China nos verões de 2001 até 2005. Análise espectral foi aplicada à observações de concentrações diárias locais de MP_{10} e precipitação de 31 estações terrestres, estimativas de reanálise de variáveis atmosféricas regionais e dados de nuvens obtidos via satélite. Os resultados confirmaram a presença de interação entre MP_{10} e as condições meteorológicas na camada limite e sugeriram uma possível conexão entre a formação de nuvens e MP_{10} em uma escala semanal.

Sarnaglia et al. (2016) propuseram uma metodologia de estimação baseada no estimador de Whittle para ajustar modelos PARMA quando o processo é contaminado por outliers aditivos e/ou tem ruído de cauda pesada. Ela foi derivada pela troca da transformada de Fourier comum pelo estimador não linear de M-regressão na equação de regressão harmônica que leva ao periodograma clássico. Um experimento de Monte Carlo foi conduzido para estudar o comportamento de amostra finita do estimador proposto em cenários de séries contaminadas e não contaminadas. O método de estimação proposto foi aplicado para ajustar um modelo PARMA às médias diárias das concentrações de SO_2 da estação da RAMQAr de Jardim Camburi de 01/01/2005 até 29/12/2006.

Fajardo et al. (2018) propuseram o M-periodograma para utilização em séries temporais com memória longa, estabeleceram suas propriedades assintóticas e investigaram suas propriedades empíricas para amostras finitas sob diferentes cenários. Além de ser interessante para utilização em séries temporais com memória longa, esse periodograma é resistente a outliers aditivos, resistência essa que foi testada por meio de simulações. A metodologia proposta foi aplicada à série temporal das médias diárias das concentrações de MP_{10} da estação da RAMQAr do Centro de Vila Velha de 01/03/2008 até 31/12/2009.

Reisen et al. (2018) propuseram um estimador robusto semi-paramétrico para os parâmetros fracionários do modelo sazonal autorregressivo fracionalmente integrado de médias móveis (SARFIMA), por meio da utilização de um periodograma robusto tanto nas frequências muito baixas quanto nas frequências sazonais. Foi demonstrado pelas simulações que a metodologia robusta se comporta como a clássica na estimação dos parâmetros de memória longa se não há outliers (sem contaminação). Por outro lado, no cenário com contaminação (presença de outliers), a metodologia clássica levou a resultados enganosos enquanto a metodologia robusta não foi afetada. A metodologia proposta foi aplicada para modelar e prever concentrações médias diárias do poluente atmosférico SO_2 da estação da RAMQAr de Cariacica de 01/01/2005 até 31/12/2009, pois essa série possui características de memória longa sazonal e grandes picos ocasionais de concentrações de

poluente.

Rao et al. (1985) utilizaram estatística de valores extremos e a técnica de bootstrap para observações independentes e identicamente distribuídas proposta por Efron (1979) para revelar a performance de modelos de qualidade do ar em simular a distribuição acumulada das concentrações medidas. Essas técnicas foram aplicadas à predições do modelo de qualidade do ar RAM da U.S. Environmental Protection Agency e à medições de uma base de dados de concentrações horárias de SO_2 de Saint Louis, Estados Unidos.

Hanna (1989) utilizou os procedimentos de reamostragem de bootstrap e de jackknife para estimar as incertezas ou os intervalos de confiança de medidas de performance relacionadas a concentrações de qualidade do ar, pois as distribuições das mesmas, em geral, não podem ser facilmente transformadas em formato gaussiano. Esses procedimentos de reamostragem foram aplicados à predições de sete modelos de qualidade do ar para o experimento de dispersão costeira de Carpinteria, Estados Unidos. Intervalos de confiança do vício médio fracionário e do erro quadrático médio normalizado foram calculados para cada modelo e para diferenças entre modelos. Concluiu-se que essas incertezas são em algumas ocasiões tão grandes para conjuntos de dados de aproximadamente 20 elementos que não se pode afirmar com 95% de confiança que a medida de performance do "melhor" modelo é significativamente diferente da obtida para outro modelo.

Martin & Roberts (2006) utilizaram o bootstrap estacionário proposto por Politis & Romano (1994) como forma de abordar os problemas de seleção de modelos em estudos de séries temporais de mortalidade relacionada ao material particulado. A metodologia proposta, denominada bootstrap model averaging (BOOT), foi aplicada aos dados diários de mortalidade, clima e MP_{10} de Cook Country, nos Estados Unidos, no período de 1987 a 2000. Já Roberts & Martin (2010) propuseram o double BOOT como uma extensão do BOOT, essa metodologia é similar ao BOOT com a diferença que utiliza uma segunda camada de bootstrap. A metodologia foi aplicada às séries temporais diárias de mortalidade, temperatura, temperatura de ponto de orvalho e concentração de partículas respiráveis ($MP_{2,5}$) de cinco cidades (Birmingham, Orlando, Seattle, Saint Louis e Tampa) dos Estados Unidos nos anos de 1999 e 2000 e mostrou ser uma alternativa viável ao BOOT.

Barbosa (2009) utilizou Modelos Aditivos Generalizados (MAG) e a técnica de bootstrap para explicar a associação entre as concentrações diárias dos poluentes MP_{10}, O_3 e NO_2 e o número de atendimentos hospitalares por causas respiratórias em crianças de 0 a 6 anos de idade na Região Metropolitana da Grande Vitória de janeiro de 2001 a dezembro de 2004. Os resultados mostraram que os procedimentos e os intervalos de confiança de bootstrap condicional apresentaram um desempenho satisfatório quando utilizados na classe MAG, que por sua vez encontrou efeitos maléficos dos poluentes investigados na saúde das crianças que apresentaram problemas respiratórios no período de estudo.

Jhun et al. (2015) avaliaram o impacto de mudanças climáticas de longo prazo na qualidade do ar e na saúde nos Estados Unidos de 1994 até 2012. Foram quantificados

aumentos relacionados ao clima passado, ou "penalidade do clima", no O_3 e no $MP_{2,5}$, e posteriormente estimado o excesso de mortes associado. Utilizando modelos aditivos generalizados, foi obtida a penalidade do clima como o aumento adicional na poluição do ar relativo a tendências assumindo condições climáticas constantes (isto é, tendências ajustadas de acordo com o clima). Já os desvios padrão das penalidades do clima foram estimados utilizando uma metodologia de bootstrap em blocos.

Barakat et al. (2015) realizaram um estudo das concentrações horárias de SO_2 e de MP_{10} nas cidades de Décimo de Ramadan e de Zagazig no Egito em 2009. Esse estudo utilizou o bloco de máximas e o pico acima de um limiar para avaliar as medições dos poluentes atmosféricos considerados. Uma técnica de bootstrap foi utilizada para melhorar as estimativas de parâmetros no modelo de valores extremos e sua validade foi checada pelo teste de Kolmogorov-Smirnov. Foi sugerido um novo método de modelar valores extremos que pode converter qualquer dado ordenado em um bloco de dados ampliado utilizando m dos n bootstraps. Também foi investigada a inconsistência e a consistência fraca do bootstrap de estatísticas de ordem centrais e intermediárias para uma escolha apropriada de tamanho de reamostragem.

Shang (2018) considerou métodos de bootstrap para a estimação da covariância de longo prazo de séries temporais funcionais estacionárias. Foi introduzido um método de bootstrap versátil que se baseia em análise de componentes principais funcional, onde pode ser feito o bootstrap dos *scores* das componentes principais por entropia máxima. Dois outros métodos de bootstrap reamostraram funções de erro, depois da estrutura de dependência ter sido modelada linearmente por um modelo funcional autoregressivo ou não linearmente por regressão kernel funcional. Por uma série de simulações de Monte Carlo, foram avaliadas e comparadas as performances de amostra finita desses três métodos de bootstrap na estimação da covariância de longo prazo de uma série temporal funcional. Os métodos de bootstrap propostos foram aplicados aos dados de MP_{10} medidos de meia em meia hora na cidade de Graz na Áustria no período que vai de 01/10/2010 até 31/03/2011 para construir a distribuição da sua covariância de longo prazo estimada.

As diversas aplicações que a literatura apresenta para a junção das técnicas estatísticas relacionadas à metodologias para o tratamento de dados faltantes, análise de séries temporais no domínio da frequência, estatística robusta e estimação de intervalos de confiança via bootstrap fazem com que essa área de estudo seja umas das mais dinâmicas na análise de séries temporais de concentrações de poluentes atmosféricos. Dessa forma, justifica-se a escolha desse tema como objeto de estudo desta tese. Outra motivação para a escolha desse tema foi o fato dele ser relativamente recente dentro da área de poluição do ar, o que faz com que sua exploração tenda a trazer novos resultados.

4 Materiais e Métodos

4.1 Região de Estudo

A Região Metropolitana da Grande Vitória (RMGV) é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra, Fundão, Guarapari e Viana. A RMGV é localizada na região sudeste do estado do Espírito Santo. Sua área é de 2331,03 km^2 e possui uma população de aproximadamente 1,884 milhão de habitantes, o que representa cerca de 48% da população total do estado (IBGE, 2011).

A RMGV é o principal polo industrial e econômico do estado, com aproximadamente 63,13% do Produto Interno Bruto (PIB) do Espírito Santo. Nessa região encontram-se atividades de siderurgia, pelotização, mineração (pedreiras), cimenteira, indústria alimentícia, usina de asfalto, etc. No ano de 2007 a RMGV possuía 49,18% da frota veicular total do estado (IJSN, 2008).

No que se refere ao relevo, a RMGV é caracterizada por: cadeias montanhosas nas porções Noroeste (Mestre Álvaro) e Oeste (Região Serrana); planícies (Aeroporto e manguezais) e planaltos (Planalto Serrano) na porção Norte; e planícies (Barra do Jucu) na porção Sul. Todas porções são intercaladas por maciços rochosos de pequeno e médio porte. As condições de relevo no geral são favoráveis, em grande parte da região, à circulação de ventos para dispersão de poluentes (IEMA, 2006).

4.2 Rede Automática de Monitoramento da Qualidade do Ar

O início do funcionamento da Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr) foi no ano de 2000. A rede é de propriedade e responsabilidade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). A RAMQAr é composta de 9 estações de monitoramento distribuídas nos municípios da RMGV da seguinte forma: 3 estações no município de Serra (Laranjeiras, Carapina e Cidade Continental); 3 estações no município de Vitória (Jardim Camburi, Enseada do Suá e Centro); 2 estações no município de Vila Velha (Ibes e Centro) e 1 estação no município de Cariacica (Vila Capixaba). A localização espacial das estações de monitoramento da RAMQAr encontra-se na Figura 4.1.

Os poluentes monitorados nas estações da RAMQAr são: dióxido de enxofre (SO_2), partículas totais em suspensão (PTS), partículas inaláveis (MP_{10}), partículas respiráveis ($MP_{2,5}$), ozônio (O_3), óxidos de nitrogênio (NO_X), monóxido de carbono (CO) e hidrocarbonetos (HC). Além desses poluentes, alguns parâmetros meteorológicos são monitorados: direção escalar do vento (DV), desvio padrão da direção do vento (SV), velocidade escalar

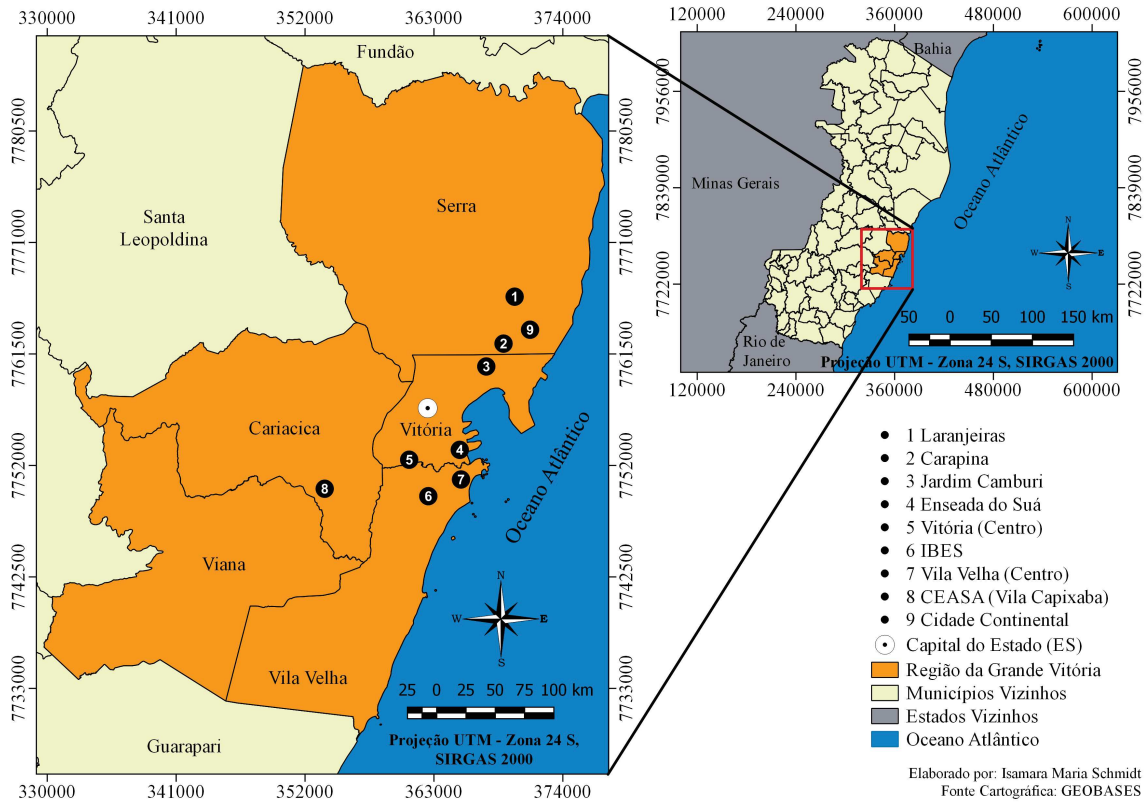


Figura 4.1 – Localização espacial das estações da RAMQAr.

do vento (VV), precipitação pluviométrica (PP), umidade relativa do ar (UR), temperatura do ar (T), pressão atmosférica (P) e radiação solar (I). Nem todos os poluentes e parâmetros meteorológicos são monitorados por todas as estações. Os poluentes e os parâmetros meteorológicos monitorados por cada estação estão mostrados na Tabela 4.1.

Tabela 4.1 – Poluentes e parâmetros meteorológicos em cada estação da RAMQAr.

Estação	**PTS**	**MP_{10}**	**$MP_{2,5}$**	**SO_2**	**CO**	**NO_X**	**HC**	**O_3**	**Meteorologia**
Laranjeiras	X	X		X	X	X		X	
Carapina	X	X							DV,SV,VV,UR,PP,P,T,I
Jardim Camburi	X	X		X		X			
Enseada do Suá	X	X	X	X	X	X	X	X	DV,SV,VV
Vitória Centro	X	X		X	X	X	X		
Ibes	X	X	X	X	X	X	X	X	DV,SV,VV,T,UR
Vila Velha Centro		X		X					
Vila Capixaba	X	X		X	X	X		X	DV,SV,VV,T,UR
Cidade Continental	X	X		X					DV,VV

A estação da Enseada do Suá e a estação do Ibes são as únicas que registram as concentrações de todos os poluentes, enquanto apenas a estação de Carapina monitora todos os parâmetros meteorológicos.

4.3 Dados

Esta tese foi realizada, à exceção do apêndice, utilizando os dados de MP_{10} medidos em estações da RAMQAr nos anos de 2018 e de 2019. Os dados são fornecidos em médias horárias de concentrações medidas em $\mu g/m^3$.

4.4 *Software* Estatístico

As metodologias consideradas e toda análise efetuada foi realizada no *software* R. Além de ser gratuito, o R possui um grande número de procedimentos estatísticos convencionais, entre eles estão os modelos lineares, modelos de regressão não linear, análise de séries temporais, testes estatísticos paramétricos e não paramétricos, análise multivariada, etc. Esse *software* dispõe ainda de uma grande quantidade de funções para o desenvolvimento de ambiente gráfico e criação de diversos tipos de apresentação de dados (REISEN; SILVA, 2011).

5 Resultados e Discussões

Nesta seção encontram-se os resultados e as discussões referentes à esta pesquisa compilados no formato de dois artigos científicos.

5.1 Robust Local Bootstrap for Weakly Stationary Time Series in the Presence of Additive Outliers

Carlo Corrêa Solci Valdério Anselmo Reisen Paulo Jorge Canas Rodrigues

Abstract

The aim of this paper is to propose a generalization of the local bootstrap for periodogram statistics to the case when weakly stationary time series are contaminated by additive outliers. In order to achieve robustness, we suggest to replace the classical version of the periodogram with the M-periodogram in the local bootstrap procedure. The robust bootstrap periodogram is implemented in the Whittle estimator to obtain confidence intervals for the parameters of a time series model. A finite sample size investigation was conducted to compare the performance of the classical local bootstrap with the one proposed in this paper, to estimate 95% confidence intervals for the parameters of autoregressive and of seasonal autoregressive time series. The results have shown that the robust estimator is resistant to additive outlier contamination and produces confidence intervals with coverage percentage closer to 95% and with lower amplitudes than the ones obtained with the classical estimator, even for small percentages and magnitudes of outliers. It was also empirically demonstrated that when the expected number of outliers is kept constant, the coverage percentages of the confidence intervals of the robust estimators tend to 95% as the sample size increases. An application to the daily mean concentration of the particulate matter with diameter smaller than 10 µm (PM_{10}) was considered to illustrate the methodologies in a real data context. All the results presented here give strong motivation to use the proposed robust methodology in practical situations in which weakly stationary time series are contaminated by additive outliers.

Keywords. Bootstrap; Periodogram; Robust estimation; Whittle estimator; PM_{10} pollutant.

5.1.1 Introduction

The bootstrap is a resampling technique that provides tools for statistical analysis without requiring rigorous structural assumptions. It was initially proposed by Efron (1979), but despite its efficiency for independent and identically distributed (i.i.d.) variables, it was shown by Singh (1981) that Efron's methodology is inadequate to the case of dependent data. Due to this fact, several approaches to perform the bootstrap in time series have been proposed, as addressed, for example in Lahiri (2003) and Kreiss & Paparoditis (2011). In time series, the bootstrap approaches can be built in the time and frequency domains.

As well-known, an important quantity for time series analysis in the frequency domain is the spectral density function which can be estimated classically by the periodogram, hence the bootstrap in this domain generates periodogram replicates. In this context, the bootstrap in the frequency domain has an advantage over the one in the time domain since, for weakly stationary processes, the periodogram ordinates are nearly independent (a more precise definition is that they are asymptotically independent). Thus, the classical bootstrap approach of drawing with replacement of Efron (1979) can be potentially applied to them. There are several bootstrap approaches in the frequency domain, some examples are the multiplicative residual bootstrap of Franke & Härdle (1992), the local bootstrap of Paparoditis & Politis (1999) and the hybrid bootstrap of Kreiss & Paparoditis (2003).

The bootstrap methodologies in the frequency domain are useful to estimate population quantities, such as the standard error and the quantiles of some statistic of interest, based on the sampling distribution of estimators that are functions of the periodogram. Among these approaches, a particularly interesting one is the local bootstrap of Paparoditis & Politis (1999) because of its simplicity to implement and its similarity to the approach of Efron (1979). Due the fact that the distribution of each periodogram ordinate is a function of its frequency, the resampling is performed locally, that is, by choosing with replacement between periodogram ordinates corresponding to frequencies which are near to the frequency of interest.

In order to use the local bootstrap to obtain confidence intervals of the parameter vector $\boldsymbol{\varphi}$ of weakly stationary time series models, it is necessary to estimate the values of these parameters as functionals of the periodogram $I_N(\lambda)$ of a sample $Y_1, Y_2 \ldots, Y_N$, as well as of the parametric spectral density $f(\lambda, \boldsymbol{\varphi})$ of the process $\{Y_t\}$, $t \in \mathbb{Z}$. This can be achieved by using an important class of estimators that are obtained through the minimization of the criterion $\int_{-\pi}^{\pi} \left\{ \log f(\lambda, \boldsymbol{\varphi}) + \frac{I_N(\lambda)}{f(\lambda, \boldsymbol{\varphi})} \right\} d\lambda$, which are well-known as the Whittle estimators and were initially proposed by Whittle (1953). The confidence intervals of $\boldsymbol{\varphi}$, computed by using local bootstrap, are obtained without having to make parametric assumptions about the form of the underlying population $\{Y_t\}$. This makes the local bootstrap an interesting alternative to estimate confidence intervals of the parameters of weakly stationary time series models.

It is important to recall that, since the periodogram is a classical estimator of the spectral density function, it does not have the property of being resistant to additive outlier contamination. Hence, the Whittle estimators have their performance deteriorated when there is presence of this kind of observation. In this situation it is more appropriate to use a robust version of the Whittle estimators which is obtained by replacing the periodogram $I_N(\lambda)$ in the criterion $\int_{-\pi}^{\pi} \left\{ \log f(\lambda, \boldsymbol{\varphi}) + \frac{I_N(\lambda)}{f(\lambda, \boldsymbol{\varphi})} \right\} d\lambda$ by a robust counterpart of $I_N(\lambda)$. In this context, there are some versions of the periodogram that are resistant to additive outlier contamination such as the Q_n-periodogram, see, for example, Molinares, Reisen & Cribari-Neto (2009), and the M-periodogram, see, for instance, Reisen, Lévy-

Leduc & Taqqu (2017), Fajardo et al. (2018). The latter has the advantage to provide an autocovariance function which is positive semidefinite and this motivates the use of the robust version of the Whittle estimators obtained by using it as the estimator of the spectral density function. Since the methodology proposed by Paparoditis & Politis (1999) is based in the resampling of the ordinates of the classical periodogram $I_N(\lambda)$ to obtain via Whittle estimators the bootstrap confidence intervals of the parameters of weakly stationary time series, these intervals are shifted to the left when there is contamination by additive outliers because of the sensitivity of $I_N(\lambda)$ to this type of outlying observation.

In this context, this paper proposes a robust alternative to the local bootstrap of Paparoditis & Politis (1999) which is resistant to additive outlier contamination since it generates confidence intervals of parameters of weakly stationary time series with a significant reduction in the aforementioned effect of left shift. The proposed robust local bootstrap is obtained by replacing the classical periodogram $I_N(\lambda)$ by the robust M-periodogram $I_{N,\psi}(\lambda)$ of Reisen, Lévy-Leduc & Taqqu (2017). Hence, the bootstrap versions of the time series parameters are obtained via the robust Whittle estimator that uses $I_{N,\psi}(\lambda)$. The finite sample properties of the robust local bootstrap for series generated by the processes AR(1) and SARMA$(1,0)\times(1,0)_4$ under scenarios with and without additive outlier contamination were investigated and compared to the ones of the methodology of Paparoditis & Politis (1999) through a Monte Carlo study. Furthermore, the daily mean concentration of the atmospheric pollutant PM_{10} (particulate matter with diameter smaller than 10 µm) in the Greater Vitória Region, in the Brazilian state of Espírito Santo, was used to illustrate the bootstrap methodologies in a real air quality area application, because it may present observations with high levels of pollutant concentrations which can be modeled as additive outliers.

The rest of the paper is organized as follows: Section 5.1.2 summarizes the well-known local bootstrap of Paparoditis & Politis (1999) and shows how to compute the classical periodogram based on a regression equation, it also discusses the robust M-periodogram of Reisen, Lévy-Leduc & Taqqu (2017) and its asymptotic properties; Section 5.1.3 introduces the proposed robust local bootstrap and discusses the Whittle estimator and its robust counterpart that uses $I_{N,\psi}(\lambda)$; Section 5.1.4 presents the results of the Monte Carlo simulation experiment; Section 5.1.5 shows the results of the application of the bootstrap methodologies to PM_{10} concentrations; Section 5.1.6 concludes the paper.

5.1.2 The Model, Assumptions, the Local Bootstrap and Spectral Estimators

Let $\{Y_t\}$, $t \in \mathbb{Z}$, be a real valued weakly stationary linear process, i.e., it satisfies the difference equation

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \tag{5.1}$$

where $\{\epsilon_t\}$, $t \in \mathbb{Z}$, is a sequence of i.i.d. random variables with $\mathsf{E}(\epsilon_t) = 0$, $\mathsf{E}(\epsilon_t^2) = \sigma^2$ and $\mathsf{E}(\epsilon_t^4) < \infty$. Moreover, $\{\psi_j\}$, $j \in \mathbb{Z}$, is a sequence of constants such that $\psi_0 = 1$ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Since the robust local bootstrap approach proposed in this paper is based on the local bootstrap method suggested in Paparoditis & Politis (1999), some of their assumptions are also considered here.

Let $Y_1, Y_2, \ldots, Y_N$, be a sample from $\{Y_t\}$ and $\lambda_j = 2\pi j/N$, $j = 0, 1, 2, \ldots, N'$, be the Fourier frequencies with $N' = [N/2]$, where $[x]$ is the integer part of x. A classical non-parametric spectral estimator is the periodogram function which is given by

$$I_N(\lambda_j) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} Y_t \exp(-i\lambda_j t) \right|^2 . \tag{5.2}$$

This definition can be extended for any $\lambda \in [-\pi, \pi]$, if we let $I_N(\lambda) = I_N\{r(N, \lambda)\}$, where for $\lambda \in [0, \pi]$ we have that $r(N, \lambda)$ is the multiple of $2\pi/N$ closest to λ (the smaller one if there are two), and for $\lambda \in [-\pi, 0)$ we set $r(N, \lambda) = r(N, -\lambda)$.

The local bootstrap procedure relies on the asymptotic independence of the periodogram ordinates as well as in the smoothness of the spectral density function. To achieve these necessary properties, $f(\lambda)$ has to fulfill the following conditions.

Remark 1. If the spectral density of Y_t in (5.1), which can be obtained by $f(\lambda) = \sigma^2(2\pi)^{-1} |\sum_{j=-\infty}^{\infty} \psi_j \exp(-ij\lambda)|^2$, satisfies $f(\lambda) > 0$ for all $\lambda \in [-\pi, \pi]$, and if $0 < \lambda_1 < \cdots < \lambda_m < \pi$, then the random vector $(I_N(\lambda_1), \ldots, I_N(\lambda_m))'$ converges in distribution to a vector of independent and exponentially distributed random variables, the i^{th} component of which has mean $f(\omega_i)$, $i = 1, \ldots, m$. Under the additional assumption of $\sum_{j=-\infty}^{\infty} |j|^{1/2} |\psi_j| < \infty$, we have that $\mathsf{Cov}(I_N(\lambda_j), I_N(\lambda_k)) = O(N^{-1})$, if $\lambda_j \neq \lambda_k$. In order to ensure the smoothness of the spectral density we assume that $f(\lambda)$ is continuously differentiable with bounded derivative in $[-\pi, \pi]$.

The asymptotic results in Remark 1 show that the periodogram, although is an unbiased estimator of the spectral density, it is not a consistent estimator, i.e, its variance $\mathsf{Var}(I_N(\lambda_j)) = O(1)$ (as $N \to \infty$). However, for any two neighboring frequencies, λ_1, λ_2, $\mathsf{Cov}(I_N(\lambda_1), I_N(\lambda_2))$ decreases as N increases. With the assumptions that the errors $\{\epsilon_t\}$ are Gaussian white noise processes and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, we have that asymptotically the set of random variables $\{2I_N(\lambda_j)/f(\lambda_j)\}$, $j = 0, 1, \ldots, N'$, are independently distributed, and for $j \neq 0, N/2$ (N even), each is asymptotically distributed as a $\chi^2_{(2)}$.

The local bootstrap scheme for the periodogram is summarized as follows (for more details, see Paparoditis & Politis (1999)).

(i) Choose a resampling width k_N where $k_N = k(N) \in \mathbb{N}$ and $k_N \leq [N'/2]$.

(ii) Define i.i.d. discrete random variables $J_1, J_2, \ldots, J_{N'}$, that assume values in the set $\{-k_N, -k_N + 1, \ldots, k_N\}$ with probability $\mathsf{P}(J_i = s) = p_{k_N,s}$ for $s = 0, \pm 1, \ldots, \pm k_N$.

(iii) The bootstrap periodogram is defined by $I_N^*(\lambda_j) = I_N(\lambda_{J_j+j})$ for $j = 1, 2, \ldots, N'$, $I_N^*(\lambda_j) = I_N^*(-\lambda_j)$ for $\lambda_j < 0$ and for $\lambda_j = 0$ we have $I_N^*(\lambda_j) = 0$.

Conditionally on the sample $Y_1, Y_2, \ldots, Y_N$, the expected value and variance of the bootstrap periodogram are, respectively, given by

$$\mathsf{E}\{I_N^*(\lambda)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s=-k_N}^{k_N} p_{k_N,s} I_N\{r(N,\lambda) + \lambda_s\} \equiv \tilde{f}(\lambda) \tag{5.3}$$

and

$$\mathsf{Var}\{I_N^*(\lambda)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s=-k_N}^{k_N} p_{k_N,s} I_N^2\{r(N,\lambda) + \lambda_s\} - \tilde{f}^2(\lambda). \tag{5.4}$$

As can be seen from Equations 5.3 and 5.4, $\tilde{f}(\lambda)$ and $\sum_{s=-k_N}^{k_N} p_{k_N,s} I_N^2\{r(N,\lambda) + \lambda_s\}$ can be thought of as kernel estimators of $f(\lambda)$ and $\mathsf{E}\{I_N^2(\lambda)\} = \{2 + \eta(\lambda)\}f^2(\lambda) + o(1)$, respectively, where

$$\eta(\lambda) = \begin{cases} 1, & \text{if } \lambda = 0 \,(\mathrm{mod}\, \pi)\,, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, in order to ensure the convergence of $I_N^*(\lambda)$, we need to let $k_N \to \infty$ as $N \to \infty$ such that $k_N = \mathrm{o}(N)$, and the sequence $\{p_{k_N,s} : -k_N \leq s \leq k_N\}$ has to satisfy $\sum_{s=-k_N}^{k_N} p_{k_N,s} = 1$, $p_{k_N,s} = p_{k_N,-s}$ and $\sum_{s=-k_N}^{k_N} p_{k_N,s}^2 \to 0$ as $k_N \to \infty$.

Under the above assumption, it follows that, in probability, $\mathsf{E}\{I_N^*(\lambda)|Y_1, Y_2, \ldots, Y_N\} \to f(\lambda)$ and $\mathsf{Var}\{I_N^*(\lambda|Y_1, Y_2, \ldots, Y_N)\} \to (1 + \eta(\lambda))f^2(\lambda)$. These show that, for a fixed j and for $N \to \infty$, the bootstrap periodogram $I_N^*(\lambda_j)$ has the same mean and variance of $I_N(\lambda_j)$. The authors also established that $I_N^*(\lambda_j) \to I_N(\lambda_j)$ in distribution.

In practical situations, $p_{k_N,s}$ is chosen based on

$$p_{k_N,s} = \frac{W(\pi s k_N^{-1})}{\sum_{s=-k_N}^{k_N} W(\pi s k_N^{-1})}. \tag{5.5}$$

where $W(\cdot)$ is a sequence of weight functions satisfying, for all λ, $W(\lambda) = W(-\lambda)$, $W(\lambda) \geq 0$, and $\int_{-\pi}^{\pi} W(\lambda)d\lambda = 1$, $\int_{-\pi}^{\pi} W^2(\lambda)d\lambda < \infty$. $W(\cdot)$ is well-known as a kernel function, and is widely used to obtain a consistent spectral estimator, i.e, the smoothed periodogram. Classical examples of $W(\cdot)$ are: Parzen kernel, Daniell kernel, Bartlett-Priestley kernel, among others (see, for instance, Taniguchi & Kakizawa (2000), Priestley (1981) for further details).

Alternatively, when comparing the results of the local bootstrap applied to samples with different sizes it may be more convenient to fix constants $\nu > 0$ and $\alpha \in (0, 1)$ in order to define a resampling bandwidth $b_N = \nu N^{-\alpha}$ as a function of N and calculate the corresponding resampling width as $k_N = [Nb_N/2]$. This yields an alternative version of (5.5) which is given by

$$p_{b_N,s} = \frac{W\{2\pi s(Nb_N)^{-1}\}}{\sum_{s=-k_N}^{k_N} W\{2\pi s(Nb_N)^{-1}\}}.$$

As addressed, for example, in Reisen, Lévy-Leduc & Taqqu (2017), Fajardo et al. (2018), the periodogram in (5.2) can also be computed based on the following regression equation

$$Y_i = c'_{Ni}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)}\cos(i\lambda_j) + \beta^{(2)}\sin(i\lambda_j) + \varepsilon_i\ ,\ 1 \leq i \leq N,\ \boldsymbol{\beta} \in \mathbb{R}^2\ , \tag{5.6}$$

where $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ and ε_i denotes the deviation of Y_i from $c'_{Ni}\boldsymbol{\beta}$. Thus, the periodogram $I_N(\lambda_j)$ is calculated from

$$I_N(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)\|^2 = \frac{N}{8\pi}\left((\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j))^2 + (\hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))^2\right) =: I_N^{\mathrm{LS}}(\lambda_j), \tag{5.7}$$

where $\|\cdot\|$ denotes the classical Euclidian norm and $\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'$ is the least-square estimator of $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model given in (5.6) computed from

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \underset{\boldsymbol{\beta}(\lambda_j)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{N}(Y_i - c'_{N,i}(\lambda_j)\boldsymbol{\beta}(\lambda_j))^2, \tag{5.8}$$

where

$$c'_{N,i}(\lambda_j) = (\cos(i\lambda_j)\ \sin(i\lambda_j))\,. \tag{5.9}$$

5.1.2.1 The M-periodogram Spectral Estimator

As it is well-known, M-estimation is an alternative robust procedure to the least-square estimation approach. Thus, based on the regression equation in (5.6), the M-regression estimator is used here to estimate the vector $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ by $\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j) = (\hat{\beta}_{N,\psi}^{(1)}(\lambda_j), \hat{\beta}_{N,\psi}^{(2)}(\lambda_j))$, which is the solution of

$$\sum_{i=1}^{N} c_{N,i}(\lambda_j)\psi(Y_i - c'_{N,i}(\lambda_j)\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)) = \mathbf{0}, \tag{5.10}$$

where $\psi(\cdot)$ was chosen as the Huber (1964) function,

$$\psi(x) = \psi_\delta(x) = \begin{cases} x, & \text{if } |x| \leq \delta, \\ \operatorname{sign}(x)\delta, & \text{if } |x| > \delta. \end{cases} \tag{5.11}$$

By analogy to (5.7), the robust periodogram $I_{N,\psi}(\lambda_j)$ is defined by

$$I_{N,\psi}(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)\|^2 = \frac{N}{8\pi}\left[(\hat{\beta}_{N,\psi}^{(1)}(\lambda_j))^2 + (\hat{\beta}_{N,\psi}^{(2)}(\lambda_j))^2\right]. \tag{5.12}$$

Similarly to $I_N(\lambda)$, this definition can also be extended for any $\lambda \in [-\pi, \pi]$, if we let $I_{N,\psi}(\lambda) = I_{N,\psi}\{r(N,\lambda)\}$ for $\lambda \in [0, \pi]$ and for $\lambda \in [-\pi, 0)$ we set $r(N,\lambda) = r(N, -\lambda)$.

Remark 2. The Huber function is chosen here because it satisfies assumptions (A1)-(A4) of Reisen et al. (2019). These authors establish that, for any fixed j and under the additional assumption that $\varepsilon_i = \sum_{j=0}^{\infty} a_j \eta_{i-j}$, where $\{\eta_j\}$, $j \in \mathbb{Z}$, is a sequence of i.i.d. standard Gaussian random variables as well as that a_j is a sequence of constants such that $a_0 = 1$ and $\sum_{j=0}^{\infty} |a_j| < \infty$, we have

$$I_{N,\psi}(\lambda_j) \xrightarrow{d} \frac{X^2 + Y^2}{4\pi(F(c) - F(-c))^2}, \quad \text{as } N \to \infty, \tag{5.13}$$

where c is a positive constant, $F(\cdot)$ is the cumulative distribution function of ε_1,

$$X \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathsf{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right), \; Y \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathsf{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right) \tag{5.14}$$

and

$$\mathsf{Cov}(X, Y) = \sum_{k\in\mathbb{Z}} \mathsf{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\sin(k\lambda_j). \tag{5.15}$$

As well-addressed in the recent literature, the M-periodogram $I_{N,\psi}(\cdot)$ becomes an alternative spectral estimator for linear time series, with short- and long-memory correlation structures, such as ARMA and ARFIMA processes, respectively. An overview of robust spectral estimators for these classes of time series is addressed in Reisen et al. (2019). In addition to its elegant asymptotic properties, $I_{N,\psi}(\cdot)$ has the interesting empirical property of being robust against outliers, while the classical periodogram $I_N(\cdot)$ of (5.7) is fully affected by this type of observations.

5.1.3 The Local Bootstrap and Whittle Estimator Using $I_{N,\psi}(\cdot)$

We now introduce the local bootstrap using $I_{N,\psi}(\cdot)$, denoted by $I^*_{N,\psi}(\cdot)$. This approach follows similar guidelines of the local bootstrap scheme discussed previously where k_N, b_N, W, $\{p_{k_N,s} : -k_N \le s \le k_N\}$, $\{p_{b_N,s} : -k_N \le s \le k_N\}$, $\{I_N(\lambda_j) : 0 \le j \le N'\}$, and $\{I^*_N(\lambda_j) : 0 \le j \le N'\}$ are replaced by $k_{N,\psi}$, $b_{N,\psi}$, W_ψ, $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \le s' \le k_{N,\psi}\}$, $\{p_{b_{N,\psi},s'} : -k_{N,\psi} \le s' \le k_{N,\psi}\}$, $\{I_{N,\psi}(\lambda_j) : 0 \le j \le N'\}$, and $\{I^*_{N,\psi}(\lambda_j) : 0 \le j \le N'\}$, respectively. The assumptions for $k_{N,\psi}$, W_ψ, and $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \le s' \le k_{N,\psi}\}$ are kept the same as of k_N, W, and $\{p_{k_N,s} : -k_N \le s \le k_N\}$, sequentially. Without loss of generality, we assume here that $k_{N,\psi} = k_N$, $b_{N,\psi} = b_N$, $W_\psi = W$, $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \le s' \le k_{N,\psi}\} = \{p_{k_N,s} : -k_N \le s \le k_N\}$, and $\{p_{b_{N,\psi},s'} : -k_{N,\psi} \le s' \le k_{N,\psi}\} = \{p_{b_N,s} : -k_N \le s \le k_N\}$.

Analogously to the local bootstrap for the classical periodogram, the first two conditional moments of the robust bootstrap periodogram $I^*_{N,\psi}(\lambda)$ are, respectively, given by

$$\mathsf{E}\{I^*_{N,\psi}(\lambda)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s'=-k_{N,\psi}}^{k_{N,\psi}} p_{k_{N,\psi},s'} I_{N,\psi}\{r(N,\lambda) + \lambda_{s'}\} \equiv \tilde{f}_\psi(\lambda) \tag{5.16}$$

and

$$\mathsf{Var}\{I^*_{N,\psi}(\lambda)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s'=-k_{N,\psi}}^{k_{N,\psi}} p_{k_{N,\psi},s'} I^2_{N,\psi}\{r(N,\lambda) + \lambda_{s'}\} - \tilde{f}^2_{\psi}(\lambda). \tag{5.17}$$

It is important to emphasize that $\tilde{f}_{\psi}(\lambda)$ and $\sum_{s'=-k_{N,\psi}}^{k_{N,\psi}} p_{k_{N,\psi},s'} I^2_{N,\psi}\{r(N,\lambda) + \lambda_{s'}\}$ can be thought of as robust kernel estimators of $f(\lambda)$ and $\mathsf{E}\{I^2_N(\lambda)\}$, respectively.

5.1.3.1 Whittle Estimators

To estimate the parameters of the model satisfying Equation 5.1, we consider the Whittle estimator initially proposed by Whittle (1953) and widely used in the literature of time series. Let $\boldsymbol{\varphi}$ be the parameter vector of the process $\{Y_t\}$ with parametric spectral density $f(\lambda, \boldsymbol{\varphi})$. The estimates of $\boldsymbol{\varphi}$, denoted by $\hat{\boldsymbol{\varphi}}_W$, are obtained by minimizing

$$\int_{-\pi}^{\pi} \left\{ \log f(\lambda, \boldsymbol{\varphi}) + \frac{I_N(\lambda)}{f(\lambda, \boldsymbol{\varphi})} \right\} d\lambda, \tag{5.18}$$

where the notation log refers to the natural logarithm and $I_N(\lambda)$ is the periodogram function defined previously and computed from the sample $Y_1, \ldots, Y_N$, of the process $\{Y_t\}$. Equivalently, the Whittle estimator $\hat{\boldsymbol{\varphi}}_W$ can be obtained by minimizing

$$\bar{\sigma}^2_N(\boldsymbol{\varphi}) = \frac{1}{N} \sum_j \frac{I_N(\lambda_j)}{g(\lambda_j, \boldsymbol{\varphi})} \tag{5.19}$$

where $g(\lambda, \boldsymbol{\varphi}) = 2\pi f(\lambda, \boldsymbol{\varphi})/\sigma^2$ and the sum is taken over all frequencies $\lambda_j = 2\pi j/N \in (-\pi, \pi]$.

The weakly stationary and invertible Autoregressive Moving Average (ARMA(p,q)) model $Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}, \{\epsilon_t\} \sim \text{IID}(0, \sigma^2)$ and $\mathsf{E}(\epsilon_t^4) < \infty$, where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$ have no common zeroes, is a particular time series model satisfying Equation 5.1. For this model, we have $g(\lambda, \boldsymbol{\varphi}) = |\theta(e^{-i\lambda})|^2/|\phi(e^{-i\lambda})|^2$.

Remark 3. Let $\boldsymbol{\varphi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ and denote by C the parameter set, $C = \{\boldsymbol{\varphi} \in \mathbb{R}^{p+q} : \phi(z)\theta(z) \neq 0 \text{ for } |z| \leq 1,\ \phi_p \neq 0,\ \theta_q \neq 0, \text{ and } \phi(\cdot), \theta(\cdot) \text{ have no common zeroes}\}$. Let $\bar{\boldsymbol{\varphi}}_N$ be the estimator in C that minimizes $\bar{\sigma}^2_N(\boldsymbol{\varphi})$ for an ARMA process $\{Y_t\}$ with true parameter values $\boldsymbol{\varphi}_0 \in C$ and $\sigma_0^2 > 0$. Then,

(i) $\bar{\boldsymbol{\varphi}}_N \xrightarrow{as} \boldsymbol{\varphi}_0$ and $\bar{\sigma}_N(\bar{\boldsymbol{\varphi}}_N) \xrightarrow{as} \sigma_0^2$, as $N \to \infty$, where $\xrightarrow{as}$ denotes almost sure convergence.

(ii) $\bar{\boldsymbol{\varphi}}_N \xrightarrow{d} \mathcal{N}(\boldsymbol{\varphi}_0, N^{-1}V^{-1}(\boldsymbol{\varphi}_0))$, as $N \to \infty$, where

$$V(\boldsymbol{\varphi}_0) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left[\frac{\partial \log g(\lambda, \boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}} \right] \left[\frac{\partial \log g(\lambda, \boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}} \right]' d\lambda,$$

with $\xrightarrow{d}$ denoting convergence in distribution.

The results of items (i) and (ii) are stated in Theorems 10.8.1 and 10.8.2 of Brockwell & Davis (1991), respectively.

(iii) Replacing $I_N(\lambda_j)$ by $I_{N,\psi}(\lambda_j)$ in Equation 5.19, it is possible to obtain the Whittle estimator of $\boldsymbol{\varphi}$ using M-periodogram, i.e, $\hat{\boldsymbol{\varphi}}_{W,\psi}$, by minimizing

$$\bar{\sigma}^2_{N,\psi}(\boldsymbol{\varphi}) = \frac{1}{N}\sum_j \frac{I_{N,\psi}(\lambda_j)}{g(\lambda_j, \boldsymbol{\varphi})}, \tag{5.20}$$

where the sum is also taken over all frequencies $\lambda_j = 2\pi j/N \in (-\pi, \pi]$.

(iv) It can be shown that

$$\hat{\boldsymbol{\varphi}}_{W,\psi} \xrightarrow{p} \boldsymbol{\varphi}_0, \text{ as } N \to \infty, \tag{5.21}$$

where $\xrightarrow{p}$ denotes convergence in probability. The proof of the above result follows similar arguments of Theorem 10.8.1 in Brockwell & Davis (1991). The full proof is given in Theorem 1 in Reisen, Lévy-Leduc & Solci (2022) (paper in compilation shown in the Appendix A).

Regarding the local bootstrap estimators discussed here, $\hat{\boldsymbol{\varphi}}^*_W$ is obtained by replacing $I_N(\lambda_j)$ by $I^*_N(\lambda_j)$ in (5.19), while one can get $\hat{\boldsymbol{\varphi}}^*_{W,\psi}$ by replacing $I_{N,\psi}(\lambda_j)$ by $I^*_{N,\psi}(\lambda_j)$ in (5.20). Whereas concerning the conditional expected values of these estimators, $\tilde{\boldsymbol{\varphi}}_W = \mathsf{E}(\hat{\boldsymbol{\varphi}}^*_W | Y_1, Y_2, \ldots, Y_N)$ can be calculated by replacing $I_N(\lambda_j)$ by $\tilde{f}(\lambda_j)$ in (5.19) while one can obtain $\tilde{\boldsymbol{\varphi}}_{W,\psi} = \mathsf{E}(\hat{\boldsymbol{\varphi}}^*_{W,\psi} | Y_1, Y_2, \ldots, Y_N)$ by replacing $I_{N,\psi}(\lambda_j)$ by $\tilde{f}_\psi(\lambda_j)$ in (5.20). The empirical properties of these estimators are discussed in the next section.

5.1.4 Monte Carlo Study

In order to investigate the impact of atypical observations on the estimates obtained from the methods discussed previously, series of weakly stationary linear processes were generated with and without outliers. Let $\{Z_t\}$ be defined as follows

$$Z_t = Y_t + \omega V_t \tag{5.22}$$

where $\{Y_t\}$ is a weakly stationary linear process that satisfies Equation 5.1, additionally, $\{V_t\}$ is a sequence of independent random variables with $\mathsf{P}(V_t = -1) = \mathsf{P}(V_t = 1) = \xi/2$ and $\mathsf{P}(V_t = 0) = 1 - \xi$, $\xi \in [0, 1)$. Moreover, for all t and s, $\{Y_t\}$ and $\{V_s\}$ are independent variables and ω is the magnitude of the outlier.

The simulation study was carried out via the generation of series of autoregressive and seasonal autoregressive processes with and without additive outliers. More specifically, the time series chosen were of AR(1) $Y_t = \phi Y_{t-1} + \epsilon_t$ with $\phi = 0.2$, 0.5, and 0.8, as well as of SARMA$(1,0) \times (1,0)_{\mathcal{S}}$ processes $Y_t = \phi Y_{t-1} + \Phi Y_{t-\mathcal{S}} - \phi\Phi Y_{t-\mathcal{S}-1} + \epsilon_t$ with $\mathcal{S} = 4$,

$\phi = 0.5$, and $\Phi = 0.2$, 0.5, and 0.7. The series $\{Y_t\}$ of both processes were contaminated by additive outliers according to Equation 5.22 with $pr_{out} = \xi = 0.005$ and 0.01, and $\omega = 0$, 4, and 7, generating the processes $\{Z_t\}$. The parameter values were chosen to achieve stationarity and low, moderate and strong correlation dependency. The sample sizes were taken as small ($N = 200$) and large ($N = 400$), which are common sample sizes in practical situations, and for the series of both processes the random variables ϵ_t were generated independently and $\mathcal{N}(0, 1)$ distributed. It is important to highlight that the value $pr_{out} = 0.01$ was used for both $N = 200$ and $N = 400$, while the value $pr_{out} = 0.005$ was used only for $N = 400$, being these choices considered to compare the results maintaining the probability and the expected number of outliers constant when the sample size increases. For the robust estimator we have chosen $\delta = 1.345$ in the Huber function (Equation 5.11) as a compromise between robustness and efficiency. Additionally, we have used $b_{N,\psi} = b_N = \nu N^{-\alpha}$, where $\nu = 0.15$ and $\alpha = 0.45$, being b_N the 'resampling bandwidth' of $I_N(\lambda_j)$, $b_{N,\psi}$ the 'robust resampling bandwidth' of $I_{N,\psi}(\lambda_j)$, with these quantities used to obtain the sets of probabilities of choosing the periodogram ordinates in the bootstrap procedure. The choice of a SARMA$(1,0) \times (1,0)_8$ process was due to the fact that one of the real data time series analyzed in the Section 5.1.5 follows a seasonal time series model. Another motivation to simulate a SARMA$(1,0) \times (1,0)_8$ process is the fact that all the theory given in Section 5.1.3.1 for an ARMA process is also valid for a SARMA process.

As a means to evaluate if the bootstrap estimates were able to mimic some features of the distributions of interest, we have calculated the estimates of the mean values $\overline{x} = \mathsf{E}(x)$, of the standard deviation $\mathsf{SD}(x) = \sqrt{\mathsf{Var}(x)}$, of the asymmetry coefficient $\gamma_1(x) = \mathsf{E}([\{x - \overline{x}\}/\mathsf{SD}(x)]^3)$, and of the 95% confidence interval $\mathrm{CI}_{95\%}(y)$ together with its amplitude $\mathrm{A}(y)$ and coverage percentage $\mathrm{P}(y)$. The value of x is $\hat{\phi}^*$ for the AR(1) model and can be $\hat{\phi}^*$ or $\hat{\Phi}^*$ for the SARMA$(1,0) \times (1,0)_8$ model, while y has the value $\overline{\hat{\phi}^*}$ for the AR(1) model and can be $\overline{\hat{\phi}}^*$ or $\overline{\hat{\Phi}}^*$ for the SARMA$(1,0) \times (1,0)_8$ model. The results of the bootstrap estimates for the parameters are shown in Tables 5.1-5.9, for the AR(1) series, and in Tables 5.10-5.18 for the SARMA$(1,0) \times (1,0)_8$ series. In the following, if a table has the column I_N or I_N^* it is to show the type of periodogram used: C denotes the classical and M designates the robust. For both models, the Bartlett-Priestley kernel was used to calculate the set of probabilities of the bootstrap. The bootstrap estimates were obtained thorough the generation of $REP_{MC} = 1000$ Monte Carlo replicates of $\{Z_t\}$ and, for each of them, $B = 5000$ bootstrap replicates of the periodogram were generated, with their related estimated parameters being denoted by $\hat{\phi}^{*(1)}, \hat{\phi}^{*(2)}, \ldots, \hat{\phi}^{*(B)}$ or by $\hat{\Phi}^{*(1)}, \hat{\Phi}^{*(2)}, \ldots, \hat{\Phi}^{*(B)}$, which were used to estimate the aforementioned characteristics of the distributions of interest.

It is important to highlight that to avoid taking average of confidence intervals in the bootstrap procedure, which would be necessary due to the fact that each Monte Carlo

replicate generates a confidence interval $\mathrm{CI}_{95\%}(x)$, where x takes the values of $\hat{\phi}^*$ or $\hat{\Phi}^*$, it was preferred to estimate the 95% bootstrap confidence interval as the 2.5% and the 97.5% percentiles of the empirical distribution of the mean values $\overline{\hat{\phi}^*} = \sum_{i=1}^{B} \hat{\phi}^{*(i)}/B$ or $\overline{\hat{\Phi}^*} = \sum_{i=1}^{B} \hat{\Phi}^{*(i)}/B$. For each Monte Carlo replicate these percentile intervals were denoted by $\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$ with amplitude $\mathrm{A}(\overline{\hat{\phi}^*})$ and coverage percentage $\mathrm{P}(\overline{\hat{\phi}^*})$, or by $\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$ with amplitude $\mathrm{A}(\overline{\hat{\Phi}^*})$ and coverage percentage $\mathrm{P}(\overline{\hat{\Phi}^*})$. The choice of this methodology to estimate the bootstrap confidence interval is due to the fact that the average of intervals of certain confidence level usually does not maintain the same confidence level of the intervals of which the average was taken. In this context, we have to emphasize that Tables 5.1-5.18, which display the results of the bootstrap estimates, have the average values for all the calculated estimates (that in the case of the confidence interval as well as of its amplitude and coverage percentage were calculated based on a single value), and between parentheses are the standard deviations only of the estimates of the mean values, of the standard deviations and of the asymmetries of the parameters. For the bootstrap confidence intervals, the coverage percentage $\mathrm{P}(x)$ was calculated as the percentage of times in which the true value of the bootstrap estimates, calculated for the uncontaminated series $\{Y_t\}$ (that can be the component referring to x of $\tilde{\boldsymbol{\varphi}}_W$ or $\tilde{\boldsymbol{\varphi}}_{W,\psi}$), is contained in the confidence interval of the bootstrap procedure $\mathrm{CI}_{95\%}(x)$ where x takes the values of $\overline{\hat{\phi}^*}$ or $\overline{\hat{\Phi}^*}$.

Tables 5.1-5.18 show that the bootstrap estimates for both the classical and the robust methodology have coverage percentages close to 95% in the scenarios without contamination, which demonstrates the efficiency of both methodologies in this scenario. However, when there is data contamination by additive outliers, only the robust methodologies are able to maintain coverage percentages close to 95%, while the classical methodologies perform worse and worse when compared to the robust ones as the value of pr_{out} or of ω increases. In this context, it is important to emphasize that the confidence intervals of the robust approaches had coverage percentages tending to 95% as the sample size increases while the expected number of outliers is kept constant, i.e., when we go from the scenario with $N = 200$ and $pr_{out} = 0.01$ to the one with $N = 400$ and $pr_{out} = 0.005$, as in this case the outlier effect is diluted with the increase of N. Moreover, it should be noted that for the scenarios with contamination, the robust methodologies generated confidence intervals that, when compared to the classical methodologies, in addition to presenting coverage percentages closer to 95%, they also presented lower amplitudes. This gives empirical evidence that the robust local bootstrap is a good alternative to estimate confidence intervals of parameters of weakly stationary time series for which there is suspect of contamination by additive outliers. When compared to the local bootstrap of Paparoditis & Politis (1999), it has similar performance when there is no outlier contamination and it generates intervals with better performance in terms of both amplitude and coverage percentage in the presence of additive outliers in the data.

Table 5.1 – Bootstrap Estimates for $\phi = 0.2$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.1816(0.0709)	0.0533(0.0076)	-0.1001(0.0743)	(0.0471,0.3160)	0.2689	0.9490
	M	0.1716(0.0713)	0.0534(0.0077)	-0.0964(0.0750)	(0.0350,0.3103)	0.2753	0.9470
4	C	0.1566(0.0724)	0.0544(0.0079)	-0.0897(0.0737)	(0.0123,0.2955)	0.2832	0.9390
	M	0.1652(0.0694)	0.0541(0.0077)	-0.0902(0.0715)	(0.0266,0.2926)	0.2660	0.9430
7	C	0.1282(0.0792)	0.0535(0.0073)	-0.0766(0.0751)	(-0.0157,0.2843)	0.3000	0.9140
	M	0.1662(0.0732)	0.0540(0.0076)	-0.0933(0.0730)	(0.0153,0.3074)	0.2921	0.9420

Table 5.2 – Bootstrap Estimates for $\phi = 0.2$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.1929(0.0490)	0.0400(0.0043)	-0.0749(0.0481)	(0.0976,0.2910)	0.1934	0.9480
	M	0.1837(0.0496)	0.0401(0.0043)	-0.0716(0.0498)	(0.0844,0.2827)	0.1983	0.9490
4	C	0.1808(0.0489)	0.0401(0.0041)	-0.0673(0.0482)	(0.0852,0.2776)	0.1924	0.9400
	M	0.1814(0.0490)	0.0400(0.0041)	-0.0673(0.0471)	(0.0851,0.2740)	0.1889	0.9460
7	C	0.1540(0.0543)	0.0402(0.0041)	-0.0615(0.0480)	(0.0448,0.2587)	0.2139	0.9290
	M	0.1757(0.0485)	0.0401(0.0044)	-0.0665(0.0468)	(0.0765,0.2712)	0.1947	0.9450

Table 5.3 – Bootstrap Estimates for $\phi = 0.2$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
4	C	0.1638(0.0497)	0.0402(0.0042)	-0.0641(0.0478)	(0.0665,0.2581)	0.1916	0.9120
	M	0.1737(0.0487)	0.0401(0.0042)	-0.0675(0.0492)	(0.0825,0.2694)	0.1869	0.9450
7	C	0.1325(0.0550)	0.0402(0.0040)	-0.0535(0.0492)	(0.0243,0.2400)	0.2157	0.8220
	M	0.1761(0.0501)	0.0401(0.0041)	-0.0685(0.0483)	(0.0772,0.2696)	0.1924	0.9400

Table 5.4 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.4745(0.0631)	0.0481(0.0091)	-0.2818(0.0989)	(0.3438,0.5873)	0.2435	0.9430
	M	0.4546(0.0667)	0.0488(0.0090)	-0.2690(0.0976)	(0.3170,0.5756)	0.2586	0.9470
4	C	0.4184(0.0748)	0.0515(0.0090)	-0.2548(0.0987)	(0.2660,0.5645)	0.2985	0.9140
	M	0.4354(0.0685)	0.0506(0.0088)	-0.2651(0.1028)	(0.2934,0.5700)	0.2766	0.9380
7	C	0.3568(0.0980)	0.0526(0.0092)	-0.2181(0.0993)	(0.1688,0.5425)	0.3737	0.8100
	M	0.4412(0.0681)	0.0499(0.0090)	-0.2616(0.0984)	(0.2996,0.5647)	0.2651	0.9360

5.1.5 An Application to the Air Quality Area

The application is based on a data set (air pollutant variables) collected at Automatic Air Quality Monitoring Network (RAMQAr) in the Greater Vitória Region (GVR) in the Brazilian state of Espírito Santo, which is composed by nine monitoring stations placed in

Table 5.5 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.4889(0.0438)	0.0356(0.0048)	-0.2078(0.0635)	(0.4012,0.5747)	0.1735	0.9460
	M	0.4689(0.0461)	0.0363(0.0049)	-0.1988(0.0606)	(0.3732,0.5593)	0.1861	0.9460
4	C	0.4597(0.0482)	0.0367(0.0049)	-0.1989(0.0613)	(0.3567,0.5509)	0.1942	0.9210
	M	0.4587(0.0448)	0.0368(0.0049)	-0.1962(0.0598)	(0.3708,0.5429)	0.1721	0.9430
7	C	0.4169(0.0644)	0.0381(0.0053)	-0.1788(0.0606)	(0.2917,0.5369)	0.2452	0.8610
	M	0.4590(0.0457)	0.0367(0.0049)	-0.1935(0.0602)	(0.3690,0.5461)	0.1771	0.9420

Table 5.6 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
4	C	0.4347(0.0499)	0.0374(0.0047)	-0.1882(0.0595)	(0.3328,0.5251)	0.1923	0.7890
	M	0.4497(0.0465)	0.0369(0.0046)	-0.1934(0.0596)	(0.3568,0.5408)	0.1840	0.9380
7	C	0.3580(0.0661)	0.0396(0.0050)	-0.1518(0.0589)	(0.2284,0.4792)	0.2508	0.3910
	M	0.4497(0.0456)	0.0371(0.0049)	-0.1924(0.0578)	(0.3595,0.5370)	0.1775	0.9300

Table 5.7 – Bootstrap Estimates for $\phi = 0.8$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.7677(0.0435)	0.0360(0.0102)	-0.6529(0.2035)	(0.6731,0.8410)	0.1679	0.9330
	M	0.7494(0.0479)	0.0377(0.0105)	-0.6260(0.1950)	(0.6410,0.8311)	0.1901	0.9400
4	C	0.7216(0.0622)	0.0420(0.0118)	-0.6117(0.2024)	(0.5812,0.8246)	0.2434	0.8470
	M	0.7259(0.0575)	0.0408(0.0114)	-0.6037(0.1934)	(0.5985,0.8275)	0.2290	0.9260
7	C	0.6509(0.0944)	0.0480(0.0144)	-0.5452(0.1932)	(0.4562,0.8127)	0.3565	0.7610
	M	0.7236(0.0569)	0.0406(0.0115)	-0.6007(0.1968)	(0.6020,0.8261)	0.2241	0.9100

Table 5.8 – Bootstrap Estimates for $\phi = 0.8$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.7822(0.0324)	0.0257(0.0058)	-0.4800(0.1221)	(0.7154,0.8388)	0.1234	0.9430
	M	0.7664(0.0358)	0.0269(0.0060)	-0.4590(0.1165)	(0.6925,0.8298)	0.1373	0.9410
4	C	0.7624(0.0374)	0.0274(0.0062)	-0.4627(0.1220)	(0.6818,0.8284)	0.1466	0.9110
	M	0.7559(0.0368)	0.0276(0.0059)	-0.4498(0.1207)	(0.6793,0.8250)	0.1457	0.9350
7	C	0.7190(0.0584)	0.0316(0.0076)	-0.4377(0.1168)	(0.5982,0.8176)	0.2194	0.8160
	M	0.7515(0.0370)	0.0284(0.0063)	-0.4490(0.1138)	(0.6768,0.8206)	0.1438	0.9320

strategic locations and accounts for the measuring of several atmospheric pollutants and meteorological variables in the area. GVR is comprised of seven cities with a population of approximately 2 million inhabitants in an area of $2319\,\mathrm{km}^2$. The region is situated along the South Atlantic coast of Brazil (latitude $20°19'15''$S, longitude $40°20'10''$W) and

Table 5.9 – Bootstrap Estimates for $\phi = 0.8$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
4	C	0.7372(0.0439)	0.0303(0.0071)	-0.4451(0.1184)	(0.6446,0.8134)	0.1688	0.7790
	M	0.7410(0.0393)	0.0296(0.0068)	-0.4437(0.1187)	(0.6588,0.8131)	0.1543	0.9020
7	C	0.6658(0.0677)	0.0356(0.0079)	-0.3945(0.1150)	(0.5231,0.7824)	0.2593	0.3940
	M	0.7379(0.0401)	0.0295(0.0063)	-0.4320(0.1134)	(0.6500,0.8113)	0.1613	0.8720

has a tropical humid climate, with average temperatures ranging from 24 °C to 30 °C. The data sets considered in this paper are of the pollutant Particulate Matter with diameter smaller than 10 µm (PM_{10}), measured hourly, in µg/m^3, collected at the stations located in Downtown Vila Velha and Jardim Camburi areas.

We will denote the PM_{10} concentrations in the stations of Downtown Vila Velha and Jardim Camburi by PM_{10}^{VV} and PM_{10}^{JC}, respectively. These data sets include daily average concentrations from January 1, 2018 to September 22, 2019, which keep a sample size, $N = 630$, multiple of the natural choice to the seasonality $\mathcal{S} = 7$ and it is equivalent to 90 full weeks. Due to skewness and some evidences of time varying variance, the natural logarithm transformation (log) was used and the plots of the $\log(PM_{10}^{VV})$ and $\log(PM_{10}^{JC})$ are displayed in Figures 5.1 and 5.2, respectively. From these figures, one can see large peaks of PM_{10} concentration which may be viewed here as outliers, and these high levels can provoke serious damage to some statistics, such as the mean and the standard deviation and, therefore, may affect the sample correlation structure as well as the periodogram of the series, causing misleading results. The existence of any outlier's effect will be assessed by comparing the results of robust and classical approaches while the presence of deterministic trends must be firstly removed from $\log(PM_{10}^{VV})$ and $\log(PM_{10}^{JC})$ before further analysis. This will be discussed in the sequence, where a linear model with errors following an AR(p) process is fitted to $\log(PM_{10}^{VV})$ and a linear model with errors following a SARMA$(\tilde{p}, 0) \times (P, 0)_{\mathcal{S}}$ process is fitted to $\log(PM_{10}^{JC})$.

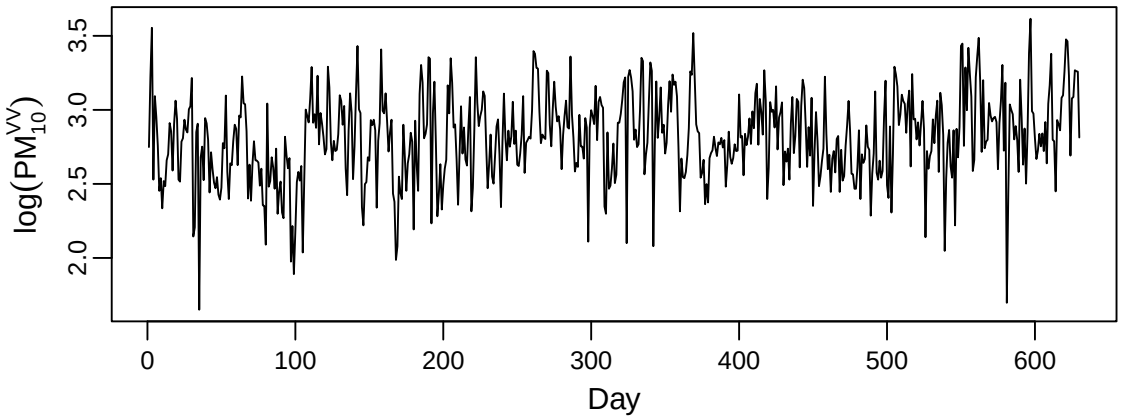


Figure 5.1 – Plot of the $\log(PM_{10}^{VV})$ time series.

Table 5.10 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.2$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4761(0.0662)	0.1875(0.0675)	0.0494(0.0088)	0.0541(0.0081)	-0.2711(0.1035)	-0.1121(0.0818)	(0.3402,0.5958)	(0.0556,0.3094)	0.2556	0.2538	0.9430	0.9450
	M	0.4565(0.0689)	0.1758(0.0669)	0.0503(0.0088)	0.0542(0.0081)	-0.2588(0.1028)	-0.1028(0.0807)	(0.3214,0.5855)	(0.0415,0.3037)	0.2641	0.2622	0.9480	0.9490
4	C	0.4241(0.0705)	0.1658(0.0695)	0.0517(0.0091)	0.0544(0.0080)	-0.2391(0.1063)	-0.0994(0.0804)	(0.2867,0.5582)	(0.0229,0.2994)	0.2715	0.2765	0.8860	0.9340
	M	0.4377(0.0650)	0.1699(0.0700)	0.0510(0.0089)	0.0544(0.0081)	-0.2460(0.1004)	-0.1014(0.0792)	(0.3079,0.5638)	(0.0360,0.3037)	0.2559	0.2677	0.9340	0.9440
7	C	0.3533(0.0946)	0.1389(0.0754)	0.0540(0.0089)	0.0539(0.0073)	-0.2011(0.0956)	-0.0847(0.0811)	(0.1709,0.5378)	(-0.0097,0.2826)	0.3669	0.2923	0.8100	0.8860
	M	0.4378(0.0662)	0.1698(0.0733)	0.0512(0.0090)	0.0546(0.0080)	-0.2470(0.0998)	-0.0983(0.0796)	(0.3057,0.5589)	(0.0201,0.3024)	0.2532	0.2823	0.9320	0.9400

Table 5.11 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.2$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4855(0.0451)	0.1940(0.0497)	0.0360(0.0047)	0.0403(0.0044)	-0.1898(0.0613)	-0.0773(0.0510)	(0.3942,0.5680)	(0.0980,0.2891)	0.1738	0.1911	0.9440	0.9470
	M	0.4653(0.0468)	0.1829(0.0501)	0.0369(0.0048)	0.0403(0.0043)	-0.1816(0.0609)	-0.0737(0.0521)	(0.3666,0.5556)	(0.0838,0.2790)	0.1890	0.1952	0.9500	0.9500
4	C	0.4602(0.0480)	0.1788(0.0490)	0.0372(0.0047)	0.0403(0.0041)	-0.1788(0.0599)	-0.0727(0.0494)	(0.3638,0.5490)	(0.0807,0.2734)	0.1852	0.1927	0.9060	0.9460
	M	0.4594(0.0466)	0.1765(0.0489)	0.0371(0.0047)	0.0404(0.0042)	-0.1764(0.0588)	-0.0718(0.0480)	(0.3667,0.5467)	(0.0801,0.2681)	0.1800	0.1880	0.9460	0.9490
7	C	0.4176(0.0624)	0.1599(0.0523)	0.0386(0.0049)	0.0400(0.0043)	-0.1629(0.0614)	-0.0660(0.0494)	(0.2964,0.5334)	(0.0564,0.2605)	0.2370	0.2041	0.8150	0.9100
	M	0.4606(0.0459)	0.1773(0.0487)	0.0371(0.0045)	0.0402(0.0042)	-0.1788(0.0616)	-0.0727(0.0498)	(0.3693,0.5442)	(0.0790,0.2698)	0.1749	0.1908	0.9410	0.9460

Table 5.12 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.2$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
4	C	0.4375(0.0527)	0.1683(0.0513)	0.0383(0.0048)	0.0404(0.0041)	-0.1725(0.0601)	-0.0686(0.0499)	(0.3261,0.5393)	(0.0638,0.2727)	0.2132	0.2089	0.8510	0.9240
	M	0.4528(0.0479)	0.1750(0.0505)	0.0375(0.0047)	0.0402(0.0041)	-0.1760(0.0602)	-0.0705(0.0516)	(0.3574,0.5473)	(0.0815,0.2779)	0.1899	0.1964	0.9300	0.9410
7	C	0.3642(0.0686)	0.1392(0.0541)	0.0404(0.0052)	0.0401(0.0040)	-0.1442(0.0619)	-0.0580(0.0506)	(0.2218,0.4905)	(0.0312,0.2433)	0.2687	0.2121	0.4840	0.8380
	M	0.4502(0.0487)	0.1718(0.0521)	0.0377(0.0046)	0.0402(0.0041)	-0.1738(0.0610)	-0.0690(0.0502)	(0.3512,0.5399)	(0.0701,0.2707)	0.1887	0.2006	0.9280	0.9400

Table 5.13 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.5$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4771(0.0655)	0.4731(0.0629)	0.0489(0.0086)	0.0484(0.0091)	-0.2654(0.1028)	-0.2901(0.1036)	(0.3421,0.5974)	(0.3306,0.5868)	0.2553	0.2562	0.9480	0.9460
	M	0.4502(0.0705)	0.4483(0.0662)	0.0504(0.0085)	0.0495(0.0090)	-0.2506(0.1006)	-0.2734(0.1016)	(0.3025,0.5803)	(0.3028,0.5675)	0.2778	0.2647	0.9510	0.9420
4	C	0.4223(0.0760)	0.4220(0.0726)	0.0535(0.0099)	0.0518(0.0094)	-0.2393(0.1060)	-0.2704(0.1057)	(0.2754,0.5649)	(0.2763,0.5616)	0.2895	0.2853	0.9070	0.8970
	M	0.4280(0.0711)	0.4282(0.0676)	0.0528(0.0098)	0.0512(0.0094)	-0.2397(0.1029)	-0.2689(0.1008)	(0.2876,0.5622)	(0.2800,0.5553)	0.2746	0.2753	0.9370	0.9300
7	C	0.3537(0.0945)	0.3508(0.0938)	0.0560(0.0101)	0.0538(0.0095)	-0.1990(0.1031)	-0.2296(0.1037)	(0.1639,0.5323)	(0.1646,0.5345)	0.3684	0.3699	0.7750	0.7930
	M	0.4295(0.0690)	0.4227(0.0705)	0.0529(0.0095)	0.0516(0.0094)	-0.2424(0.1059)	-0.2652(0.1029)	(0.2906,0.5578)	(0.2833,0.5540)	0.2672	0.2707	0.9250	0.9230

Table 5.14 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.5$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4871(0.0450)	0.4879(0.0436)	0.0359(0.0048)	0.0357(0.0048)	-0.1894(0.0636)	-0.2119(0.0635)	(0.3957,0.5724)	(0.4015,0.5671)	0.1767	0.1656	0.9490	0.9430
	M	0.4620(0.0475)	0.4626(0.0464)	0.0371(0.0048)	0.0366(0.0048)	-0.1785(0.0641)	-0.2024(0.0638)	(0.3686,0.5554)	(0.3677,0.5514)	0.1868	0.1837	0.9500	0.9500
4	C	0.4574(0.0496)	0.4565(0.0490)	0.0376(0.0048)	0.0370(0.0050)	-0.1787(0.0644)	-0.1970(0.0628)	(0.3540,0.5490)	(0.3518,0.5474)	0.1950	0.1956	0.9120	0.9030
	M	0.4503(0.0488)	0.4512(0.0476)	0.0378(0.0049)	0.0371(0.0050)	-0.1771(0.0633)	-0.1954(0.0626)	(0.3506,0.5446)	(0.3611,0.5435)	0.1940	0.1824	0.9430	0.9420
7	C	0.4138(0.0666)	0.4109(0.0645)	0.0393(0.0053)	0.0384(0.0052)	-0.1600(0.0647)	-0.1771(0.0634)	(0.2806,0.5393)	(0.2845,0.5329)	0.2587	0.2484	0.8360	0.8340
	M	0.4511(0.0492)	0.4493(0.0482)	0.0377(0.0048)	0.0370(0.0050)	-0.1768(0.0626)	-0.1933(0.0635)	(0.3541,0.5440)	(0.3537,0.5403)	0.1899	0.1866	0.9380	0.9360

Table 5.15 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.5$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
4	C	0.4328(0.0527)	0.4293(0.0522)	0.0388(0.0051)	0.0379(0.0050)	-0.1688(0.0637)	-0.1877(0.0620)	(0.3285,0.5382)	(0.3281,0.5275)	0.2097	0.1994	0.8650	0.8320
	M	0.4398(0.0497)	0.4366(0.0480)	0.0383(0.0051)	0.0375(0.0048)	-0.1714(0.0631)	-0.1870(0.0616)	(0.3385,0.5367)	(0.3385,0.5280)	0.1982	0.1895	0.9350	0.9380
7	C	0.3645(0.0683)	0.3642(0.0668)	0.0418(0.0055)	0.0401(0.0051)	-0.1390(0.0634)	-0.1551(0.0623)	(0.2391,0.4954)	(0.2380,0.4930)	0.2563	0.2550	0.5070	0.4990
	M	0.4406(0.0505)	0.4396(0.0477)	0.0389(0.0052)	0.0378(0.0048)	-0.1703(0.0611)	-0.1897(0.0599)	(0.3364,0.5376)	(0.3400,0.5305)	0.2012	0.1905	0.9220	0.9230

Table 5.16 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.7$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4784(0.0667)	0.6616(0.0533)	0.0509(0.0093)	0.0430(0.0105)	-0.2591(0.1061)	-0.4998(0.1588)	(0.3436,0.6033)	(0.5505,0.7587)	0.2597	0.2082	0.9490	0.9430
	M	0.4453(0.0737)	0.6358(0.0579)	0.0524(0.0093)	0.0443(0.0104)	-0.2418(0.1042)	-0.4694(0.1503)	(0.2966,0.5848)	(0.5084,0.7429)	0.2882	0.2345	0.9460	0.9490
4	C	0.4160(0.0791)	0.6050(0.0706)	0.0544(0.0094)	0.0471(0.0108)	-0.2285(0.1053)	-0.4484(0.1457)	(0.2584,0.5642)	(0.4563,0.7309)	0.3058	0.2746	0.9060	0.8720
	M	0.4140(0.0750)	0.6046(0.0667)	0.0540(0.0090)	0.0465(0.0102)	-0.2249(0.1005)	-0.4353(0.1432)	(0.2604,0.5614)	(0.4600,0.7227)	0.3010	0.2627	0.9350	0.9160
7	C	0.3447(0.0965)	0.5296(0.0975)	0.0573(0.0107)	0.0522(0.0120)	-0.1804(0.1109)	-0.3982(0.1527)	(0.1648,0.5414)	(0.3313,0.7179)	0.3766	0.3866	0.8170	0.8100
	M	0.4063(0.0750)	0.6024(0.0654)	0.0548(0.0099)	0.0469(0.0111)	-0.2192(0.1011)	-0.4392(0.1421)	(0.2547,0.5481)	(0.4735,0.7200)	0.2934	0.2465	0.9190	0.9100

Table 5.17 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.7$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
0	C	0.4887(0.0449)	0.6811(0.0390)	0.0360(0.0047)	0.0301(0.0054)	-0.1979(0.0607)	-0.3506(0.0888)	(0.3946,0.5706)	(0.5974,0.7486)	0.1760	0.1512	0.9500	0.9440
	M	0.4560(0.0494)	0.6568(0.0422)	0.0374(0.0047)	0.0313(0.0054)	-0.1809(0.0612)	-0.3326(0.0875)	(0.3560,0.5482)	(0.5644,0.7278)	0.1922	0.1634	0.9430	0.9450
4	C	0.4528(0.0530)	0.6513(0.0446)	0.0382(0.0053)	0.0322(0.0058)	-0.1780(0.0636)	-0.3354(0.0903)	(0.3440,0.5546)	(0.5566,0.7322)	0.2106	0.1756	0.9280	0.9110
	M	0.4394(0.0510)	0.6416(0.0439)	0.0387(0.0051)	0.0324(0.0059)	-0.1737(0.0615)	-0.3246(0.0861)	(0.3393,0.5374)	(0.5541,0.7274)	0.1981	0.1733	0.9410	0.9380
7	C	0.4055(0.0702)	0.6055(0.0641)	0.0399(0.0054)	0.0351(0.0066)	-0.1559(0.0705)	-0.3119(0.0917)	(0.2708,0.5386)	(0.4773,0.7214)	0.2678	0.2441	0.8320	0.8280
	M	0.4351(0.0540)	0.6394(0.0437)	0.0387(0.0050)	0.0326(0.0058)	-0.1714(0.0670)	-0.3256(0.0880)	(0.3331,0.5384)	(0.5497,0.7250)	0.2053	0.1753	0.9340	0.9330

Table 5.18 – Bootstrap Estimates for $\phi = 0.5$, $\Phi = 0.7$, $\mathcal{S} = 4$, $REP_{MC} = 1000$, $B = 5000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\overline{\hat{\Phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\mathsf{SD}(\hat{\Phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\gamma_1(\hat{\Phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{CI}_{95\%}(\overline{\hat{\Phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\Phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\Phi}^*})$
4	C	0.4229(0.0566)	0.6248(0.0483)	0.0394(0.0050)	0.0342(0.0058)	-0.1673(0.0635)	-0.3214(0.0876)	(0.3159,0.5336)	(0.5264,0.7148)	0.2177	0.1884	0.8170	0.7820
	M	0.4201(0.0540)	0.6261(0.0445)	0.0392(0.0049)	0.0337(0.0057)	-0.1664(0.0625)	-0.3183(0.0876)	(0.3153,0.5280)	(0.5352,0.7105)	0.2127	0.1753	0.9150	0.9010
7	C	0.3457(0.0709)	0.5456(0.0670)	0.0423(0.0053)	0.0388(0.0071)	-0.1266(0.0690)	-0.2781(0.0876)	(0.2026,0.4907)	(0.4148,0.6808)	0.2881	0.2660	0.5100	0.4370
	M	0.4156(0.0558)	0.6225(0.0445)	0.0395(0.0050)	0.0340(0.0062)	-0.1651(0.0630)	-0.3169(0.0886)	(0.3019,0.5253)	(0.5349,0.7054)	0.2234	0.1705	0.9140	0.8840

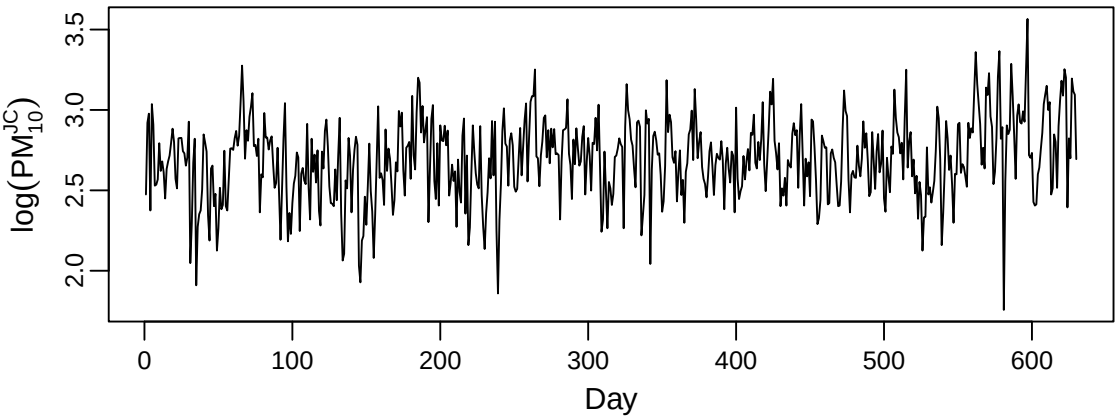


Figure 5.2 – Plot of the $\log(\text{PM}_{10}^{\text{JC}})$ time series.

From the analysis of Figures 5.1 and 5.2, it can be concluded that both time series under study have a linear trend and a more complex trend that can be modeled by cubic b-splines basis functions $B_k^3(t)$ with $d_f = 8$ and $\tilde{d}_f = 7$ degrees of freedom, for the series $\log(\text{PM}_{10}^{\text{VV}})$ and $\log(\text{PM}_{10}^{\text{JC}})$, respectively. Hence, the following model is suggested here to fit the PM_{10} concentrations of Downtown Vila Velha

$$\log(\text{PM}_{10,t}^{\text{VV}}) = \mu + \alpha t + \sum_{k=1}^{d_f} B_k^3(t)\beta_k + Y_t; \tag{5.23}$$

$$\phi_p(B)Y_t = \epsilon_t, \tag{5.24}$$

where B is the backshift operator that satisfies $B^j x_t = x_{t-j}$, additionally, we have that $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. While for the Jardim Camburi data we propose the use of

$$\log(\text{PM}_{10,t}^{\text{JC}}) = \tilde{\mu} + \tilde{\alpha} t + \sum_{k=1}^{\tilde{d}_f} B_k^3(t)\tilde{\beta}_k + \tilde{Y}_t; \tag{5.25}$$

$$\Phi_P(B^{\mathcal{S}})\tilde{\phi}_{\tilde{p}}(B)\tilde{Y}_t = \tilde{\epsilon}_t, \tag{5.26}$$

where B is the backshift operator, $\tilde{\phi}_{\tilde{p}}(B) = 1 - \tilde{\phi}_1 B - \tilde{\phi}_2 B^2 - \cdots - \tilde{\phi}_{\tilde{p}} B^{\tilde{p}}$, $\Phi_P(B^{\mathcal{S}}) = 1 - \Phi_1 B^{\mathcal{S}} - \Phi_2 B^{2\mathcal{S}} - \cdots - \Phi_P B^{P\mathcal{S}}$, and the superscript $\tilde{}$ was used to differentiate the parameters of the linear model and of the time series related to Jardim Camburi from the ones regarding Downtown Vila Velha.

The model in Equations 5.23 and 5.24 as well as the one of Equations 5.25 and 5.26 were fitted based on following two steps procedure: (i) the linear models in (5.23) and (5.25) are estimated through the ordinary least squares procedure; and (ii) the AR(p) model in (5.24) and the SARMA$(\tilde{p}, 0) \times (P, 0)_{\mathcal{S}}$ model in (5.26) are fitted to the residuals of their respective linear model in step (i), where the AR with order p as well as the AR with order $\tilde{p}$, and the seasonal AR with order P, are identified through the Schwartz Information Criterion (BIC) proposed by Schwarz (1978).

The estimated coefficients of the linear models in Equations 5.23 and 5.25, fitted in the first step, are shown in Tables 5.19 and 5.20, respectively. The residuals of the linear models did not results in rejecting the null hypothesis of level stationarity of the KPSS test, with a p-value > 0.05. In order to appropriately select the model to fit these residuals, it is important to analyze their corresponding ACFs which are displayed in Figures 5.3 and 5.4, respectively. The ACF of Figure 5.3 shows that the residuals may follow an autoregressive model because it tails off as exponential decay, while the ACF of Figure 5.4 resembles the one of a seasonal model with $\mathcal{S} = 7$ because it has peaks of autocorrelation for lags multiple of seven. These are the reasons that motivated the choices of fitting an AR(p) model and a SARMA$(\tilde{p}, 0) \times (P, 0)_{\mathcal{S}}$ model in the second step.

Table 5.19 – Estimated coefficients of the linear model for the $\log(\text{PM}_{10}^{\text{VV}})$ time series.

Parameter	μ	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Estimate	2.9350	0.0003	-0.4426	-0.0755	-0.4514	-0.1064	-0.2772	0.0034	-0.1925	-0.1756

Table 5.20 – Estimated coefficients of the linear model for the $\log(\text{PM}_{10}^{\text{JC}})$ time series.

Parameter	$\tilde{\mu}$	$\tilde{\alpha}$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$	$\tilde{\beta}_5$	$\tilde{\beta}_6$	$\tilde{\beta}_7$
Estimate	2.7880	0.0003	-0.3985	0.0454	-0.4135	-0.1085	-0.1351	-0.1177	-0.2572

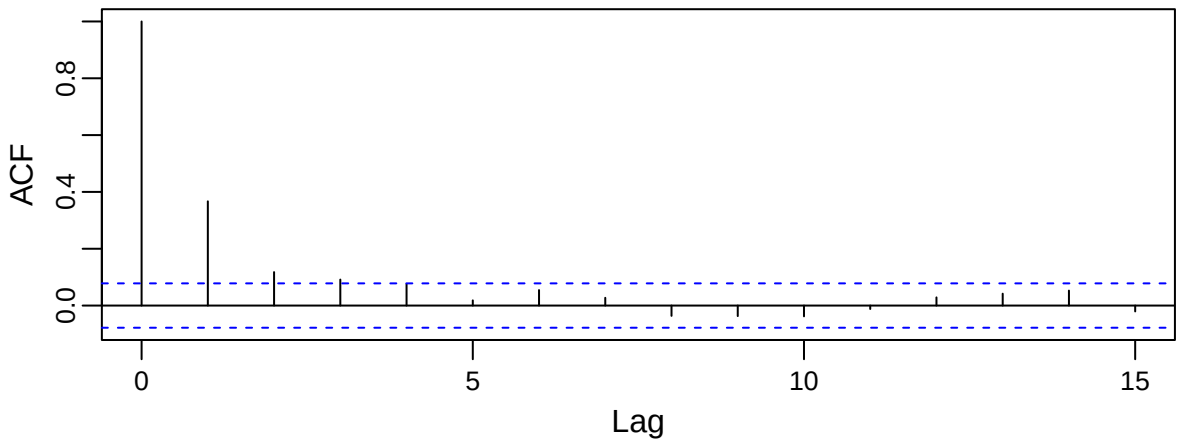


Figure 5.3 – ACF of the residuals of the linear model for the $\log(\text{PM}_{10}^{\text{VV}})$ time series.

The BIC criterion was used to identify the orders of the models and the results are displayed in Tables 5.21 and 5.22. In order to keep consistency with the simulation study, $\delta = 1.345$ was fixed in the Huber function (Equation 5.11).

The exact estimates of the AR(p) coefficients are displayed in Table 5.23 while the SARMA$(\tilde{p}, 0) \times (P, 0)_{\mathcal{S}}$ coefficients are shown in Table 5.24. Based on these results, it is clear that the robust methods always provided higher coefficient estimates. In this context, we have that for the AR(p) model the robust estimate of ϕ_1 was 10.4% bigger

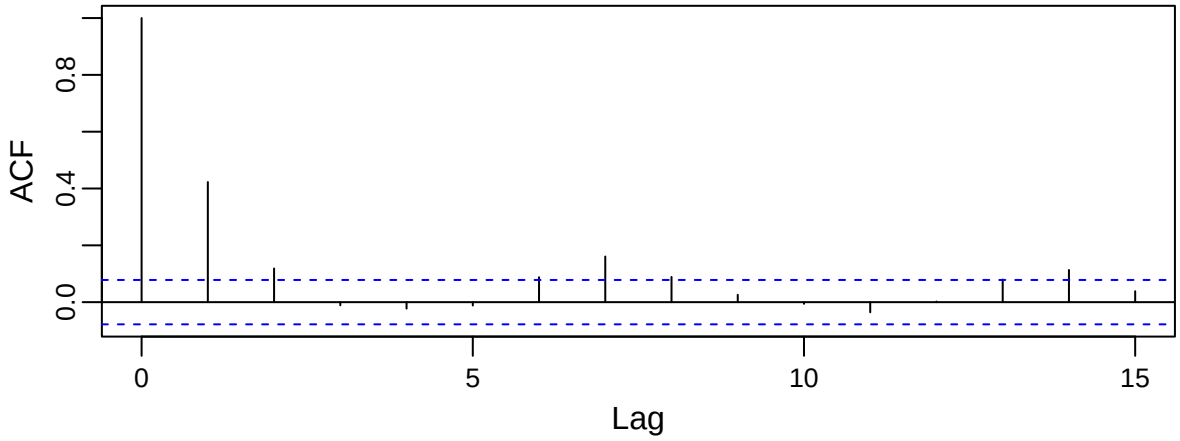


Figure 5.4 – ACF of the residuals of the linear model for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series.

Table 5.21 – Selected AR orders using the BIC for the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ time series.

I_N	BIC	p
C	-1696.445	1
M	-1695.562	1

Table 5.22 – Selected AR orders and seasonal AR orders using the BIC for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series.

I_N	BIC	$\tilde{p}$	P
C	-1935.253	1	1
M	-1934.989	1	1

than its classical counterpart, while for the SARMA$(\tilde{p},0)\times(P,0)_s$ model we have that, for instance, the robust estimate of Φ_1 was 13.5% bigger than the classical one. This indicates that the high levels of the pollutant PM_{10} presented the effects of additive outliers in both the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ and the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ series since the classical estimates suffered from memory loss while their robust counterparts were resistant to outlier contamination.

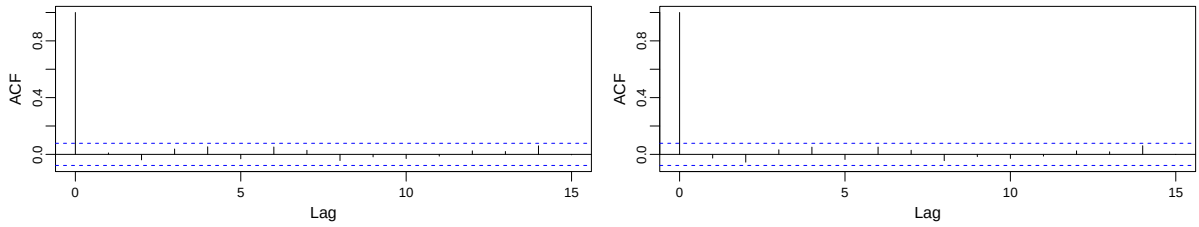
Table 5.23 – Exact estimates of the AR(p) coefficients for the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ time series.

I_N	$\hat{\phi}_1$
C	0.3642
M	0.4021

Table 5.24 – Exact estimates of the SARMA$(p,0)\times(P,0)_s$ coefficients for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series.

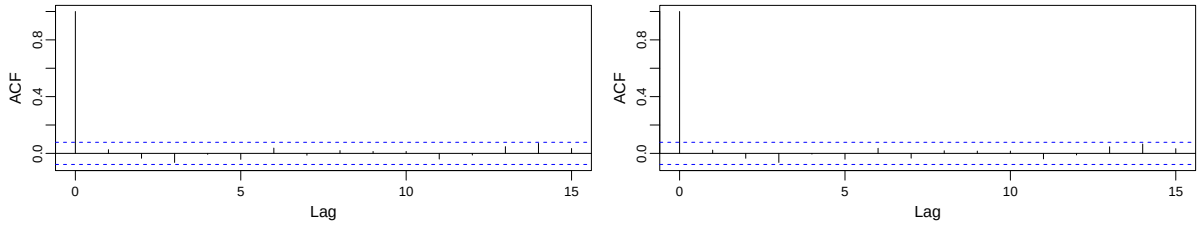
I_N	$\hat{\tilde{\phi}}_1$	$\hat{\Phi}_1$
C	0.4181	0.1451
M	0.4203	0.1647

The classical ACF of the residuals of each estimated model is shown in Figure 5.5 for the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ series, and in Figure 5.6 for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ series. It can be seen that for both series all the models were able to fully explain the correlation structure of the data, despite the eventual outliers effect. Based on the ACF of the residuals, the two estimation methods for both the AR(p) and the SARMA$(\tilde{p}, 0) \times (P, 0)_s$ models are comparable since all the estimated residuals look like a white noise process.



(a) C (b) M

Figure 5.5 – ACF of the residuals of the AR(p) fit for the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ time series.



(a) C (b) M

Figure 5.6 – ACF of the residuals of the SARMA$(\tilde{p}, 0) \times (P, 0)_s$ fit for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series.

The bootstrap estimates of the confidence intervals of the estimated coefficients for $B = 5000$ are given in Table 5.25 for the AR(p) coefficients, and in Table 5.26 for the SARMA$(\tilde{p}, 0) \times (P, 0)_s$ coefficients. It is important to highlight that, similarly to the Monte Carlo experiment, we have chosen for both models the Bartlett-Priestley kernel with $b_{N,\psi} = b_N = 0.15N^{-0.45}$ to obtain the set of probabilities to choose the periodogram ordinates in the bootstrap procedure. Based on these results, it is possible to see that the confidence intervals of the classical method have a left shift in their limits when compared to the ones of its robust counterpart. This is also evidence that both the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ and the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series suffered from the effects of additive outliers contamination.

Table 5.25 – Bootstrap estimates of the 95% confidence interval of the AR(p) coefficients for the $\log(\mathrm{PM}_{10}^{\mathrm{VV}})$ time series.

I_N^*	$\mathrm{CI}_{95\%}(\hat{\phi}_1^*)$
C	(0.3015, 0.4259)
M	(0.3402, 0.4635)

Table 5.26 – Bootstrap estimates of the 95% confidence interval of the SARMA$(\tilde{p}, 0) \times (P, 0)_s$ coefficients for the $\log(\mathrm{PM}_{10}^{\mathrm{JC}})$ time series.

I_N^*	$\mathrm{CI}_{95\%}(\hat{\tilde{\phi}}_1^*)$	$\mathrm{CI}_{95\%}(\hat{\Phi}_1^*)$
C	(0.3505, 0.4692)	(0.0034, 0.2378)
M	(0.3520, 0.4720)	(0.0195, 0.2464)

5.1.6 Conclusions

The robust version of the local bootstrap in the periodogram, presented in this paper, had its finite sample performance compared to the one of the classical bootstrap, through a Monte Carlo experiment. This empirical investigation showed that both the robust and the classical versions of the bootstrap performed well when the time series did not have outliers. However, when there was contamination by additive outliers, the classical bootstrap had its performance completely affected, while the robust one proved to be very resistant to the contamination, maintaining the coverage percentages of the confidence intervals close to 95% and presenting lower amplitudes than the classical bootstrap. The daily mean concentrations of the PM_{10} collected in the stations of Downtown Vila Velha and Jardim Camburi, in the Brazilian state of Espírito Santo, were analyzed as an application of the methodologies studied in this paper. This analysis led to the conclusion that the memory loss occurred in the classical bootstrap caused it to generate confidence intervals dislocated to the left when compared to the ones obtained by the robust bootstrap. Based on these investigations, it is possible to conclude that the robust version of the local bootstrap in the periodogram proved to be an alternative for estimating confidence intervals of parameters of models of weakly stationary time series contaminated by additive outliers.

5.2 Local Bootstrap for Weakly Stationary Time Series in the Presence of Missing Data and Additive Outliers

Carlo Corrêa Solci Valdério Anselmo Reisen Paulo Jorge Canas Rodrigues

Abstract

This paper proposes a generalization of the classical and of the robust versions of the local bootstrap for periodogram statistics to the case when weakly stationary time series have, respectively, missing data and additive outliers in the presence of missing data. In order to make it possible to work with incomplete time series, we suggest to replace the series with its amplitude modulated version. A Monte Carlo experiment was carried out to compare the performance of the bootstrap methodologies proposed in this paper, to estimate 95% confidence intervals for the parameters of autoregressive time series via the Whittle estimator. The results have shown that the exact and the bootstrap estimates provided very similar confidence intervals. For the scenario without outlier contamination and without missing data, it was demonstrated that both the proposed methodologies are as efficient as their counterparts for complete time series. When there was presence of missing data but no outlier contamination, the classical and the robust estimators for incomplete time series maintained their coverage percentages close to 95%. However, when there was contamination by additive outliers and presence of missing data, the classical bootstrap for incomplete time series had its performance completely affected, while its robust counterpart proved to be very resistant to the contamination, maintaining the coverage percentages of the confidence intervals close to 95% and presenting lower amplitudes than its classical counterpart. The daily mean concentration of the particulate matter with diameter smaller than 10 µm (PM_{10}) data was used in order to illustrate the proposed methodologies in a real application. All the results presented here give strong motivation to use the classical and the robust version of the proposed methodology in practical situations in which weakly stationary incomplete time series, respectively, are not and are contaminated by additive outliers.

Keywords. Bootstrap; Periodogram; Amplitude Modulation; Robust estimation; Whittle estimator; PM_{10} pollutant.

5.2.1 Introduction

The bootstrap is a resampling methodology proposed by Efron (1979) for independent and identically distributed (i.i.d.) variables. This approach was later extended to the time series context and this extension provides tools for statistical analysis of dependent data without requiring stringent structural assumptions. Several examples of bootstrap

methods for dependent data can be found, for example, in Lahiri (2003) and Kreiss & Paparoditis (2011). It is important to recall that these approaches can be built in the time and the frequency domains.

In this context, the bootstrap in the frequency domain generates periodogram replicates and, hence, is useful for estimating confidence intervals of statistics which are functions of the periodogram. The bootstrap in the frequency domain has an advantage over the one in the time domain since, for weakly stationary processes, the periodogram ordinates are nearly independent. Therefore, it is possible to apply to them the classical bootstrap approach of drawing with replacement proposed by Efron (1979). Some examples of bootstrap methodologies in the frequency domain are the multiplicative residual bootstrap of Franke & Härdle (1992), the local bootstrap of Paparoditis & Politis (1999) and the hybrid bootstrap of Kreiss & Paparoditis (2003).

Among the aforementioned methodologies, the local bootstrap of Paparoditis & Politis (1999) can be highlighted because it is simple and quite similar to the original idea of resampling with replacement proposed by of Efron (1979). The main difference is that the resampling is performed locally by choosing with replacement between periodogram ordinates corresponding to frequencies which are near to the frequency of interest. The reason for the resampling to be performed locally is that it is well-known that the distribution of each periodogram ordiante is a function of its frequency and, hence, resampling in the whole set of periodogram ordinates neglects this heterogeneity and leads to misleading conclusions.

As a means to use the local bootstrap to obtain confidence intervals for the vector parameter $\boldsymbol{\varphi}$ of weakly stationary time series models, one needs to estimate the values of these parameters as functionals of the periodogram $I_X(\lambda)$ of a sample $X_1, X_2 \ldots, X_N$, and of the parametric spectral density $f_X(\lambda, \boldsymbol{\varphi})$ of the process $\{X_t\}$, $t \in \mathbb{Z}$. It is possible to do this by using a class of estimators that are obtained through the minimization of the criterion $\int_{-\pi}^{\pi} \left\{ \log f_X(\lambda, \boldsymbol{\varphi}) + \frac{I_X(\lambda)}{f_X(\lambda, \boldsymbol{\varphi})} \right\} d\lambda$, which are known as the Whittle estimators (WHITTLE, 1953). Besides of its aforementioned simplicity, the local bootstrap has the advantage of providing estimates of confidence intervals of the parameters of weakly stationary time series models without requiring for assumptions about the form of the underlying population $\{X_t\}$.

It is important to emphasize that, the classical periodogram cannot be calculated when a complete sample $X_1, X_2 \ldots, X_N$, is not available and, hence, in this case the Whittle methodology is not applicable. However, it is possible to use the work of Parzen (1963) and define an amplitude modulated sequence $Y_t = a_t X_t$, for $t = 1, 2, \ldots, N$, with $a_t = 1$, if X_t is observed and $a_t = 0$, if X_t is missing. The concept of amplitude modulation was used by Dunsmuir & Robinson (1981) to obtain a version of the Whittle methodology which may be used for incomplete time series, this estimator can be obtained via the minimization of the criterion $\int_{-\pi}^{\pi} \left\{ \log f_Y(\lambda, \boldsymbol{\varphi}) + \frac{I_Y(\lambda)}{f_Y(\lambda, \boldsymbol{\varphi})} \right\} d\lambda$. The main disadvantage of the methodology

proposed by Dunsmuir & Robinson (1981) is that it is not resistant to additive outlier contamination since it uses the classical version of the periodogram. Therefore, in order to achieve robustness it is more appropriate to use the estimator obtained through the minimization of the criterion $\int_{-\pi}^{\pi} \left\{ \log f_Y(\lambda, \boldsymbol{\varphi}) + \frac{I_{Y,\psi}(\lambda)}{f_Y(\lambda,\boldsymbol{\varphi})} \right\} d\lambda$, where $I_{Y,\psi}(\lambda)$ is the robust M-periodogram of Y_t, which was proposed by Reisen, Lévy-Leduc & Taqqu (2017).

In this paper we propose a classical and a robust alternative to the local bootstrap of Paparoditis & Politis (1999) for the case when weakly stationary time series have, respectively, missing data and additive outliers in the presence of missing data. It is possible to obtain the proposed local bootstrap methodologies by replacing the classical periodogram $I_X(\lambda)$ of the complete series $X_1, X_2, \ldots, X_N$ by, respectively, the classical periodogram $I_Y(\lambda)$ and the robust M-periodogram $I_{Y,\psi}(\lambda)$ of the amplitude modulated sequence $Y_1, Y_2, \ldots, Y_N$, which is completely available even when a complete record of the series X_t is not accessible. In order to investigate the finite sample properties of the classical and of the robust local bootstrap for incomplete time series, it was conducted a Monte Carlo experiment in which time series were generated by the process AR(1) under scenarios without missing data and without additive oultlier contamination, with missing data and without additive outlier contamination, and with missing data and with additive outlier contamination. Additionally, the daily mean concentration of the atmospheric pollutant PM_{10} (particulate matter with diameter smaller than 10 µm) in the Greater Vitória Region, in the Brazilian state of Espírito Santo, was used to illustrate the usefulness of the proposed bootstrap methodologies in a real application in the air quality area, because it often has the presence of missing data and may present observations with high levels of pollutant concentrations which can be modeled as additive outliers.

The rest of this article is structured as follows: Section 5.2.2 defines a linear process in the presence of missing data; Section 5.2.3 shows how to use the methodologies of Parzen (1963) and Dunsmuir & Robinson (1981) to estimate incomplete weakly stationary time series; Section 5.2.4 presents the proposed classical version of local bootstrap for times series in the presence of missing data; Section 5.2.5 introduces the proposed robust version of local bootstrap for times series in the presence of missing data and of additive outliers; Section 5.2.6 presents the results of the Monte Carlo experiment; Section 5.2.7 shows the results of the application of the proposed bootstrap methodologies to PM_{10} concentrations; Section 5.2.8 concludes the paper.

5.2.2 Weakly Stationary Linear Process in the Presence of Missing Data

Let $\{X_t\}$, $t \in \mathbb{Z}$, be a real valued weakly stationary linear process, i.e., it is defined by the difference equation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \tag{5.27}$$

where $\{\epsilon_t\}$, $t \in \mathbb{Z}$, is a sequence of i.i.d. random variables with $\mathsf{E}(\epsilon_t) = 0$, $\mathsf{E}(\epsilon_t^2) = \sigma^2$ and $\mathsf{E}(\epsilon_t^4) < \infty$. Furthermore, $\{\psi_j\}$, $j \in \mathbb{Z}$, is a sequence of constants such that $\psi_0 = 1$ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

The weakly stationary and invertible Autoregressive Moving Average (ARMA(p,q)) model $Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}$, $\{\epsilon_t\} \sim \text{IID}(0, \sigma^2)$ and $\mathsf{E}(\epsilon_t^4) < \infty$, where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$ have no common zeroes, is a particular time series model satisfying Equation 5.27. Following Dunsmuir & Robinson (1981), we will focus on time series following an ARMA(p,q) model in the remainder of this paper.

If a sample $X_1, X_2, \ldots, X_N$ from the process $\{X_t\}$ is available, then it is possible to use the local bootstrap to estimate the confidence intervals of the parameters of a series that can be represented by (5.27) by using the classical periodogram $I_X(\lambda_j) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} X_t e^{i\lambda_j t}\right|^2$ or the robust periodogram $I_{X,\psi}(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)\|^2$ where the vector $\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j) = (\hat{\beta}^{(1)}_{N,\psi}(\lambda_j), \hat{\beta}^{(2)}_{N,\psi}(\lambda_j))$ is obtained via a M-regression estimator and is the solution of $\sum_{i=1}^{N} c_{N,i}(\lambda_j)\psi(Y_i - c'_{N,i}(\lambda_j)\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)) = \mathbf{0}$ where $\psi(\cdot)$ was chosen as the Huber (1964) function. Sometimes, such a complete record is not available and the original version of local bootstrap proposed by Paparoditis & Politis (1999) or its robust counterpart cannot be applied. In this context, this paper proposes a classical and a robust version of the local bootstrap to estimate the confidence intervals of the parameters of time series that can be written as (5.27) with the aid of the parametric estimation methodology of Dunsmuir & Robinson (1981) for the case when the observations occur at times $1 = n_1 < n_2 < \cdots < n_M = N$ and it is assumed that these observation times are independent of the process $\{X_t\}$.

5.2.3 Estimation Criteria for Weakly Stationary Time Series with Missing Data via Amplitude Modulation

An interesting approach to deal with missing observations in a weakly stationary time series is to replace our observed sequence, with its missing values, by a related series which can be handled like a stationary series of equally spaced observations. In this context, following Parzen (1963), we introduce the amplitude modulated sequence

$$Y_t = a_t X_t, \quad t = 1, 2, \ldots, N, \tag{5.28}$$

where

$$a_t = \begin{cases} 1, & \text{if } t = n_j \text{ for some } j, \\ 0, & \text{if } t \neq n_j \text{ for all } j. \end{cases} \tag{5.29}$$

The process generating the a_t's may be deterministic or random and is assumed to be independent of the process $\{X_t\}$. Now define

$$\bar{a} = N^{-1} \sum_{t=1}^{N} a_t,$$

$$C_a(l) = N^{-1} \sum_{t=1}^{N-l} a_t a_{t+l}, \tag{5.30}$$

$$C_Y(l) = N^{-1} \sum_{t=1}^{N-l} Y_t Y_{t+l}. \tag{5.31}$$

In terms of these quantities it is possible to define an estimate of $\gamma_X(l)$ as $\bar{\gamma}_X(l) = C_Y(l)/C_a(l)$. If X_t has a nonzero constant mean μ, then it may be estimated by $\bar{\mu} = \sum Y_t / \sum a_t$ and, hence, $\bar{\gamma}_X(l)$ would be modified by replacing Y_t by $Y_t - \bar{\mu}$ throughout. We also assume the following

Remark 4.

$$\lim_{N\to\infty} C_a(l) = \nu(l) \text{ almost surely, (for each fixed, finite } l)$$
$$\lim_{N\to\infty} \bar{a} = \mu \text{ almost surely}$$

Parzen (1963) described processes that satisfy the above remark as asymptotically stationary. For these processes we can define $\gamma_a(l) = \nu(l) - \mu^2 = \int_{-\pi}^{\pi} e^{il\lambda} F_a(d\lambda)$. Under Remark 4 and some additional conditions, see, for instance, Parzen (1963), Y_t is also asymptotically stationary with

$$\lim_{N\to\infty} C_Y(l) = \gamma_Y(l) \text{ almost surely,}$$

where $\gamma_Y(l) = \nu(l)\gamma_X(l)$ and $\gamma_X(l) = \mathsf{E}(X_t X_{t+l})$. As addressed, for example, in Dunsmuir & Robinson (1981) it is possible to define the asymptotic spectrum of Y_t as

$$f_Y(\lambda, \boldsymbol{\varphi}) = \int_{-\pi}^{\pi} f_X(\lambda - \omega, \boldsymbol{\varphi}) F_a(d\omega) + \mu^2 f_X(\lambda, \boldsymbol{\varphi}),$$

where $\boldsymbol{\varphi}$ is the parameter vector $\boldsymbol{\varphi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$ of Y_t. If F_a and μ are known, then one can minimize

$$L(\boldsymbol{\varphi}) = N^{-1} \sum_{j=0}^{N-1} \{ I_Y(\lambda_j) / f_Y(\lambda_j, \boldsymbol{\varphi}) + \log[2\pi f_Y(\lambda_j, \boldsymbol{\varphi})] \} \tag{5.32}$$

to estimate $\boldsymbol{\varphi}$. In the above expression

$$I_Y(\lambda_j) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} Y_t e^{it\lambda_j} \right|^2, \quad \lambda_j = 2\pi j/N,\ 0 \le j \le N-1, \tag{5.33}$$

is the periodogram of Y_t. The expression for $L(\boldsymbol{\varphi})$ can be regarded as an analog to Whittle's approximation of the Gaussian likelihood for the amplitude modulated sequence Y_t.

Unfortunately, in most of the situations F_a and μ are not completely known. However, it is possible to circumvent this problem by replacing the spectral density $f_Y(\lambda_j, \boldsymbol{\varphi})$ in $L(\boldsymbol{\varphi})$ by

$$\bar{f}_Y(\lambda_j, \boldsymbol{\varphi}) = \frac{2\pi}{N} \sum_{k=0}^{N-1} f_Y(\lambda_j - \lambda_k, \boldsymbol{\varphi}) I_a(\lambda_k), \tag{5.34}$$

where

$$I_a(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} a_t e^{it\lambda} \right|^2$$

is the periodogram of a_t. The value minimizing

$$\bar{L}(\boldsymbol{\varphi}) = \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \frac{I_Y(\lambda_j)}{\bar{f}_Y(\lambda_j, \boldsymbol{\varphi})} + \log[2\pi \bar{f}_Y(\lambda_j, \boldsymbol{\varphi})] \right\} \tag{5.35}$$

over an appropriate set of $\boldsymbol{\varphi}$ values is denoted by $\bar{\boldsymbol{\varphi}}$.

In some occasions one might be interested in finding the estimates of a concentrated vector $\boldsymbol{\xi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ which is functionally uncorrelated to σ^2. Hence, we have that $\boldsymbol{\varphi} = (\boldsymbol{\xi}', \sigma^2)'$ and $f_X(\lambda, \boldsymbol{\varphi})$ will have the form

$$f_X(\lambda, \boldsymbol{\varphi}) = \frac{\sigma^2}{2\pi} g_X(\lambda, \boldsymbol{\xi}). \tag{5.36}$$

Put $\bar{\boldsymbol{\varphi}} = (\bar{\boldsymbol{\xi}}', \bar{\sigma}^2)'$. Then $\bar{\sigma}^2 = \sigma^2(\bar{\boldsymbol{\xi}})$, where

$$\sigma^2(\boldsymbol{\xi}) = \frac{2\pi}{N} \sum_{j=0}^{N-1} \frac{I_Y(\lambda_j)}{\bar{g}_Y(\lambda_j, \boldsymbol{\xi})} \tag{5.37}$$

and $\bar{\boldsymbol{\xi}}$ is the value minimizing

$$\bar{M}(\boldsymbol{\xi}) = \log[\sigma^2(\boldsymbol{\xi})] + \frac{1}{N} \sum_{j=0}^{N-1} \log[\bar{g}_Y(\lambda_j, \boldsymbol{\xi})], \tag{5.38}$$

where

$$\bar{g}_Y(\lambda_j, \boldsymbol{\xi}) = \frac{2\pi}{N} \sum_{k=0}^{N-1} g_X(\lambda_j - \lambda_k, \boldsymbol{\xi}) I_a(\lambda_k). \tag{5.39}$$

To evaluate $\bar{g}_Y(\lambda_j, \boldsymbol{\xi})$, we need to compute the convolution (5.39) and this has to be done N times for $M(\boldsymbol{\xi})$ to be found. When N is large this could be prohibitively expensive in terms of required computational time. However, the fast Fourier transform (FFT) algorithm can speed the computations considerably. To compute the $\bar{g}_Y(\lambda_j, \boldsymbol{\xi})$ one can carry out in sequence the following four steps, in each of which the FFT may be used.

(i) $\bar{a}(\lambda_j) = N^{-1} \sum_{t=1}^{N} a_t e^{it\lambda_j}, \quad j = 0, 1, \ldots, N-1;$

(ii) $\bar{b}_k = \sum_{j=0}^{N-1} |\bar{a}(\lambda_j)|^2 e^{ik\lambda_j}, \quad k = 0, 1, \ldots, N-1;$

(iii) $\bar{c}_k = N^{-1} \sum_{j=0}^{N-1} g_X(\lambda_j, \boldsymbol{\xi}) e^{-ik\lambda_j}, \quad k = 0, 1, \ldots, N-1;$

(iv) $\bar{g}_Y(\lambda_j, \boldsymbol{\xi}) = \sum_{k=0}^{N-1} \bar{b}_k \bar{c}_k e^{ik\lambda_j}, \quad j = 0, 1, \ldots, N-1.$

One may use the above methodology of minimizing $\bar{L}(\boldsymbol{\varphi})$ to estimate the vector parameter $\boldsymbol{\varphi}$ of weakly stationary time series in the presence of missing data. However, in order to obtain estimates of the confidence intervals of these parameters, it may be necessary to make assumptions about the form of the underlying population $\{X_t\}$ of the full sample $X_1, X_2, \ldots, X_N$. The next section presents an alternative procedure to estimate these confidence intervals that does not require for structural assumptions.

5.2.4 The Local Bootstrap in the Presence of Missing Data

One possibility to deal with the presence of missing data is to apply the local bootstrap of Paparoditis & Politis (1999) to the periodogram of the amplitude modulated sequence Y_t. If we set $N' = [N/2]$ where $[x]$ is the integer part of x, then the local bootstrap approach for the periodogram of a time series that has missing data is described in the following.

(i) Choose a resampling width k_N where $k_N = k(N) \in \mathbb{N}$ and $k_N \leq [N'/2]$.

(ii) Define i.i.d. discrete random variables $J_1, J_2, \ldots, J_{N'}$, that take values in the set $\{-k_N, -k_N + 1, \ldots, k_N\}$ with probability $\mathsf{P}(J_i = s) = p_{k_N,s}$ for $s = 0, \pm 1, \ldots, \pm k_N$.

(iii) The bootstrap periodogram is defined by $I_Y^*(\lambda_j) = I_Y(\lambda_{J_j + j})$ for $j = 1, 2, \ldots, N'$, $I_Y^*(\lambda_j) = I_Y^*(2\pi - \lambda_j)$ for $N' + 1 \leq j \leq N - 1$ and $I_Y^*(\lambda_j) = 0$ for $\lambda_j = 0$.

The conditional expected value and variance of the bootstrap periodogram in the presence of missing data are, respectively, obtained by

$$\mathsf{E}\{I_Y^*(\lambda_j) | Y_1, Y_2, \ldots, Y_N\} = \sum_{s=-k_N}^{k_N} p_{k_N,s} I_Y(\lambda_j + \lambda_s) \equiv \tilde{f}_Y(\lambda_j) \tag{5.40}$$

and

$$\mathsf{Var}\{I_Y^*(\lambda_j) | Y_1, Y_2, \ldots, Y_N\} = \sum_{s=-k_N}^{k_N} p_{k_N,s} I_Y^2(\lambda_j + \lambda_s) - \tilde{f}_Y^2(\lambda_j). \tag{5.41}$$

It is possible to note from equation (5.40) that $\tilde{f}_Y(\lambda_j)$ can be viewed as a kernel estimator of $f_Y(\lambda_j)$. Hence, the convergence of $I_Y^*(\lambda_j)$ can be achieved by letting $k_N \to \infty$ as $N \to \infty$ such that $k_N = \mathrm{o}(N)$, and the sequence $\{p_{k_N,s} : -k_N \leq s \leq k_N\}$ needs to fulfill $\sum_{s=-k_N}^{k_N} p_{k_N,s} = 1$, $p_{k_N,s} = p_{k_N,-s}$ and $\sum_{s=-k_N}^{k_N} p_{k_N,s}^2 \to 0$ as $k_N \to \infty$.

One may choose $p_{k_N,s}$ in practical cases as

$$p_{k_N,s} = \frac{W(\pi s k_N^{-1})}{\sum_{s=-k_N}^{k_N} W(\pi s k_N^{-1})}, \tag{5.42}$$

where $W(\cdot)$ is a sequence of weight functions that satisfy, for all λ, $W(\lambda) = W(-\lambda)$, $W(\lambda) \geq 0$, and $\int_{-\pi}^{\pi} W(\lambda)d\lambda = 1$, $\int_{-\pi}^{\pi} W^2(\lambda)d\lambda < \infty$. It is important to recall that $W(\cdot)$ is a kernel function, which is frequently used to obtain the smoothed periodogram. In Sections 5.2.6 and 5.2.7 we will choose $W(\cdot)$ as the Bartlett-Priestley kernel.

However, if one wants to compare the results of the local bootstrap applied to samples with different sizes it might be preferable to fix constants $\nu > 0$ and $\alpha \in (0, 1)$ as a means to define a resampling bandwidth $b_N = \nu N^{-\alpha}$ as a function of N and calculate the corresponding resampling width as $k_N = [Nb_N/2]$. This generates an alternative version of (5.42) which is obtained by

$$p_{b_N,s} = \frac{W\{2\pi s(Nb_N)^{-1}\}}{\sum_{s=-k_N}^{k_N} W\{2\pi s(Nb_N)^{-1}\}}.$$

In this context, it is important to highlight that after calculating the bootstrap periodogram $I_Y^*(\lambda_j)$ of the amplitude modulated sequence Y_t, it is possible to obtain the bootstrap estimator $\bar{\boldsymbol{\xi}}^*$ of $\boldsymbol{\xi}^* = (\phi_1^*, \ldots, \phi_p^*, \theta_1^*, \ldots, \theta_q^*)'$ by replacing $I_Y(\lambda_j)$ by $I_Y^*(\lambda_j)$ in (5.37) and then minimizing (5.38). While the conditional expected value of this estimator, $\tilde{\boldsymbol{\xi}} = \mathsf{E}(\bar{\boldsymbol{\xi}}^* | Y_1, Y_2, \ldots, Y_N)$ can be calculated by replacing $I_Y(\lambda_j)$ by $\tilde{f}_Y(\lambda_j)$ in (5.37) and then minimizing (5.38).

Before introducing the robust M-periodogram in the next section, we need to emphasize that as can be seen, for example, in Reisen, Lévy-Leduc & Taqqu (2017), Fajardo et al. (2018), the periodogram of (5.33) may be also calculated based on the succeeding regression equation

$$Y_i = c'_{Ni}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)}\cos(i\lambda_j) + \beta^{(2)}\sin(i\lambda_j) + \varepsilon_i \ , \ 1 \leq i \leq N, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{5.43}$$

where $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ and ε_i is the deviation of Y_i from $c'_{Ni}\boldsymbol{\beta}$. Hence, the periodogram $I_Y(\lambda_j)$ is calculated from

$$I_Y(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)\|^2 = \frac{N}{8\pi}\left((\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j))^2 + (\hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))^2\right) =: I_Y^{\mathrm{LS}}(\lambda_j), \tag{5.44}$$

where $\|\cdot\|$ designates the classical Euclidian norm and $\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'$ denotes the least-square estimator of $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model displayed in Equation 5.43 which may be calculated from

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \underset{\boldsymbol{\beta}(\lambda_j)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - c'_{N,i}(\lambda_j)\boldsymbol{\beta}(\lambda_j))^2, \tag{5.45}$$

where

$$c'_{N,i}(\lambda_j) = (\cos(i\lambda_j) \ \sin(i\lambda_j)) \, . \tag{5.46}$$

5.2.4.1 The Robust M-periodogram

It is important to recall that M-estimation is a robust procedure which is an alternative to the classical least-square estimation methodology. Hence, we can apply the M-regression estimator to the regression equation in (5.43), in order to estimate the vector $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$ by $\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j) = (\hat{\beta}^{(1)}_{N,\psi}(\lambda_j), \hat{\beta}^{(2)}_{N,\psi}(\lambda_j))$, which is the solution of

$$\sum_{i=1}^{N} c_{N,i}(\lambda_j)\psi(Y_i - c'_{N,i}(\lambda_j)\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)) = \mathbf{0}, \tag{5.47}$$

where as the $\psi(\cdot)$ it was preferred to use the Huber (1964) function,

$$\psi(x) = \psi_\delta(x) = \begin{cases} x, & \text{if } |x| \leq \delta, \\ \text{sign}(x)\delta, & \text{if } |x| > \delta. \end{cases} \tag{5.48}$$

Analogously to (5.44), it is possible to define the robust periodogram $I_{Y,\psi}(\lambda_j)$ by

$$I_{Y,\psi}(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_{N,\psi}(\lambda_j)\|^2 = \frac{N}{8\pi}\left[(\hat{\beta}^{(1)}_{N,\psi}(\lambda_j))^2 + (\hat{\beta}^{(2)}_{N,\psi}(\lambda_j))^2\right]. \tag{5.49}$$

Remark 5. The Huber function is chosen here because it satisfies assumptions (A1)-(A4) of Reisen et al. (2019).

In this context, it is important to highlight that in addition to being an alternative periodogram for the uncontaminated time series scenario, since it has small loss of efficiency when compared to its classical counterpart $I_Y(\cdot)$ of (5.44), the M-periodogram $I_{Y,\psi}(\cdot)$ has the interesting empirical property, which is not shared by the classical periodogram $I_Y(\cdot)$, of being robust against additive outlier contamination.

5.2.5 The Robust Local Bootstrap in the Presence of Missing Data

The robust version of the local bootstrap in the presence of missing data will be denoted in the following by $I^*_{Y,\psi}(\cdot)$. This methodology is similar to the local bootstrap approach discussed previously where k_N, b_N, W, $\{p_{k_N,s} : -k_N \leq s \leq k_N\}$, $\{p_{b_N,s} : -k_N \leq s \leq k_N\}$, $\{I_Y(\lambda_j) : 0 \leq j \leq N-1\}$, and $\{I^*_Y(\lambda_j) : 0 \leq j \leq N-1\}$ are replaced by $k_{N,\psi}$, $b_{N,\psi}$, W_ψ, $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \leq s' \leq k_{N,\psi}\}$, $\{p_{b_{N,\psi},s'} : -k_{N,\psi} \leq s' \leq k_{N,\psi}\}$, $\{I_{Y,\psi}(\lambda_j) : 0 \leq j \leq N-1\}$, and $\{I^*_{Y,\psi}(\lambda_j) : 0 \leq j \leq N-1\}$, respectively. The assumptions for $k_{N,\psi}$, W_ψ, and $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \leq s' \leq k_{N,\psi}\}$ are maintained the same as of k_N, W, and $\{p_{k_N,s} : -k_N \leq s \leq k_N\}$, sequentially. It is assumed, without loss of generality, that $k_{N,\psi} = k_N$, $b_{N,\psi} = b_N$, $W_\psi = W$, $\{p_{k_{N,\psi},s'} : -k_{N,\psi} \leq s' \leq k_{N,\psi}\} = \{p_{k_N,s} : -k_N \leq s \leq k_N\}$, and $\{p_{b_{N,\psi},s'} : -k_{N,\psi} \leq s' \leq k_{N,\psi}\} = \{p_{b_N,s} : -k_N \leq s \leq k_N\}$.

Similarly to the local bootstrap for the classical periodogram, the first two conditional moments of the robust bootstrap periodogram $I^*_{Y,\psi}(\lambda)$ are, respectively, obtained by

$$\mathsf{E}\{I^*_{Y,\psi}(\lambda_j)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s'=-k_{N,\psi}}^{k_{N,\psi}} p_{k_{N,\psi},s'} I_{Y,\psi}(\lambda_j + \lambda_{s'}) \equiv \tilde{f}_{Y,\psi}(\lambda_j) \tag{5.50}$$

and

$$\mathsf{Var}\{I^*_{Y,\psi}(\lambda_j)|Y_1, Y_2, \ldots, Y_N\} = \sum_{s'=-k_{N,\psi}}^{k_{N,\psi}} p_{k_{N,\psi},s'} I^2_{Y,\psi}(\lambda_j + \lambda_{s'}) - \tilde{f}^2_{Y,\psi}(\lambda_j). \tag{5.51}$$

It is important to highlight that $\tilde{f}_{Y,\psi}(\lambda_j)$ can be seen as a robust kernel estimator of $f_Y(\lambda_j)$ and that one may obtain the robust estimate $\bar{\boldsymbol{\xi}}_\psi$ of $\boldsymbol{\xi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ by replacing $I_Y(\lambda_j)$ by $I_{Y,\psi}(\lambda_j)$ in (5.37) and then minimizing (5.38).

In this context, the robust version of the bootstrap periodogram $I^*_{Y,\psi}(\lambda_j)$ of the amplitude modulated sequence Y_t, may be used to obtain the robust bootstrap estimator $\bar{\boldsymbol{\xi}}^*_\psi$ of $\boldsymbol{\xi}^* = (\phi^*_1, \ldots, \phi^*_p, \theta^*_1, \ldots, \theta^*_q)'$ by replacing $I_Y(\lambda_j)$ by $I^*_{Y,\psi}(\lambda_j)$ in (5.37) and then minimizing (5.38). While the conditional expected value of this estimator, $\tilde{\boldsymbol{\xi}}_\psi = \mathsf{E}(\bar{\boldsymbol{\xi}}^*_\psi|Y_1, Y_2, \ldots, Y_N)$ may be obtained by replacing $I_Y(\lambda_j)$ by $\tilde{f}_{Y,\psi}(\lambda_j)$ in (5.37) and then minimizing (5.38). The next section discusses the empirical properties of this bootstrap estimator and of its classical counterpart.

5.2.6 Monte Carlo Experiment

The impact of atypical observations on the estimates obtained from the approaches discussed previously was investigated through the generation with and without outliers of time series from weakly stationary linear processes. Let $\{Z_t\}$ be defined as

$$Z_t = Y_t + \omega V_t \tag{5.52}$$

where $\{Y_t\}$ is a amplitude modulated sequence that satisfies Equation 5.28 with $\{X_t\}$ being a weakly stationary linear process satisfying Equation 5.27 and $\{a_t\}$ being a sequence of independent Bernoulli trials with $\mathsf{P}(a_t = 1) = pr_{nm}$ and $\mathsf{P}(a_t = 0) = 1 - pr_{nm}$, $pr_{nm} \in (0, 1]$, while $\{V_t\}$ is a sequence of independent random variables with $\mathsf{P}(V_t = -1) = \mathsf{P}(V_t = 1) = pr_{out}/2$ and $\mathsf{P}(V_t = 0) = 1 - pr_{out}$, $pr_{out} \in [0, 1)$. Moreover, for all t and s, $\{Y_t\}$ and $\{V_s\}$ are independent variables and ω is the magnitude of the outlier.

The Monte Carlo study was conducted through the generation of series of autoregressive processes with and without additive outliers. More specifically, we have chosen time series of AR(1) processes $X_t = \phi X_{t-1} + \epsilon_t$ with $\phi = 0.5$. The series $\{Y_t\}$ were generated by choosing $pr_{nm} = 1$ and 0.95, being contaminated by additive outliers according to Equation 5.52 with $pr_{out} = 0.005$ and 0.01, and $\omega = 0$ and 4, generating the processes

$\{Z_t\}$. The parameter value was chosen in order to attain stationarity and moderate correlation dependency. We have chosen small ($N = 200$) and large ($N = 400$) sample sizes, and the random variables ϵ_t were generated independently and $\mathcal{N}(0, 1)$ distributed. It is important to emphasize that the value $pr_{out} = 0.01$ was used for both $N = 200$ and $N = 400$, while the value $pr_{out} = 0.005$ was used only for $N = 400$, the reason for these choices was to compare the results keeping the probability and the expected number of outliers constant as the sample size increases. As a compromise between robustness and efficiency, we have chosen $\delta = 1.345$ in the Huber function (Equation 5.48) for the robust estimator. Furthermore, in order to obtain the sets of probabilities of choosing the periodogram ordinates in the bootstrap procedure, we have used $b_{N,\psi} = b_N = \nu N^{-\alpha}$, where $\nu = 0.15$ and $\alpha = 0.45$, being b_N the 'resampling bandwidth' of $I_Y(\lambda_j)$, $b_{N,\psi}$ the 'robust resampling bandwidth' of $I_{Y,\psi}(\lambda_j)$.

With the aim to evaluate if the bootstrap estimates could mimic some characteristics of the distributions of interest, we have calculated the estimates for the mean values $\overline{x} = \mathsf{E}(x)$, the standard deviation $\mathsf{SD}(x) = \sqrt{\mathsf{Var}(x)}$, the asymmetry coefficient $\gamma_1(x) = \mathsf{E}([\{x-\overline{x}\}/\mathsf{SD}(x)]^3)$, and the 95% confidence interval $\mathrm{CI}_{95\%}(y)$ together with its amplitude $\mathrm{A}(y)$ and coverage percentage $\mathrm{P}(y)$. The value of x can be $\hat{\phi}$ or $\hat{\phi}^*$, while y can assume the values of $\hat{\phi}$ or $\overline{\hat{\phi}^*}$. For the exact estimates it was also considered the estimation of the root mean squared error (RMSE) of the parameter. The results of the exact estimates of this parameter are displayed in Tables 5.27-5.31, while the results of the bootstrap estimates of the parameter are shown in Tables 5.32-5.36. In the following, if a table has the column I_N or I_N^* it is to show the type of periodogram used: C denotes the classical periodogram $I_X(\lambda_j)$ and M designates the robust periodogram $I_{X,\psi}(\lambda_j)$, which are defined only for the case when the complete sample $X_1, X_2, \ldots, X_N$ from the process $\{X_t\}$ is available; while C_m denotes the classical periodogram $I_Y(\lambda_j)$ and M designates the robust periodogram $I_{Y,\psi}(\lambda_j)$, which are also defined for the case when the observations occur at times $1 = n_1 < n_2 < \cdots < n_M = N$, since they are calculated for the amplitude modulated sequence $Y_1, Y_2, \ldots, Y_N$. We have chosen the Bartlett-Priestley kernel to calculate the set of probabilities of the bootstrap. The empirical results were obtained thorough the generation of $REP_{exac} = 10000$ Monte Carlo replicates of $\{Z_t\}$ for the exact estimates, while for the bootstrap estimates, we have generated $REP_{MC} = 1000$ Monte Carlo replicates of $\{Z_t\}$ and, for each of them, $B = 5000$ bootstrap replicates of the periodogram were generated, with their related estimated parameters being denoted by $\hat{\phi}^{*(1)}, \hat{\phi}^{*(2)}, \ldots, \hat{\phi}^{*(B)}$, these quantities were used to estimate the previously mentioned features of the distributions of interest.

It is necessary to emphasize that to avoid taking average of confidence intervals in the bootstrap procedure, which would be necessary since each Monte Carlo replicate generates a confidence interval $\mathrm{CI}_{95\%}(\hat{\phi}^*)$, it was preferred to use the 2.5% and the 97.5% percentiles of the empirical distribution of the mean values $\overline{\hat{\phi}^*} = \sum_{i=1}^{B} \hat{\phi}^{*(i)}/B$ as an

estimate of the 95% bootstrap confidence interval. For each Monte Carlo replicate these percentile intervals were designated by $\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$ with amplitude $\mathrm{A}(\overline{\hat{\phi}^*})$ and coverage percentage $\mathrm{P}(\overline{\hat{\phi}^*})$. This methodology to estimate the bootstrap confidence interval was chosen because the average of intervals of certain confidence level in most of the cases does not maintain the same confidence level of the intervals of which the average is taken. Hence, it is important to state that Tables 5.32-5.36, which display the results of the bootstrap estimates, have the average values for all the calculated estimates (that for the confidence interval together with its amplitude and its coverage percentage were calculated based on a single value), and between parentheses are shown the standard deviations just of the estimates of the mean values, of the standard deviations, and of the asymmetries of the parameters. It is also important to highlight that the coverage percentage of the exact intervals was estimated as the percentage of times that their REP_{exac} exact estimates of ϕ are contained in the theoretical asymptotic interval that each one of them possesses. These intervals can be seen in Taniguchi (1987), and since an asymptotic distribution for the theoretical interval of the estimators for incomplete time series and of the robust estimators for complete time series is not available in the literature, the coverage percentage of them could not be calculated. On the other hand, the coverage percentage $\mathrm{P}(\overline{\hat{\phi}^*})$ of the bootstrap confidence intervals was estimated as the percentage of times in which the true value of the bootstrap estimates, calculated for the uncontaminated series $\{Y_t\}$ (that may be $\tilde{\boldsymbol{\xi}}$ or $\tilde{\boldsymbol{\xi}}_{\psi}$), is contained in the confidence interval of the bootstrap procedure $\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$.

An important aspect of Tables 5.27-5.31 is that they give empirical evidence that the estimators C, M, C_m and M_m are consistent since there is a reduction in the bias and in the standard error of them when the sample size is increased in the scenario without outlier contamination and without missing data. Moreover, in this scenario, increasing the sample size makes the amplitude of the confidence intervals reduce and the coverage percentage of the interval of the classical estimator for complete time series C tend to approximate to 95%. It is also possible to note that when there is presence of missing data but not of outlier contamination, the confidence intervals for the classical C_m and for the robust M_m estimators for incomplete time series are very similar to the ones obtained when $pr_{nm} = 1$ and $pr_{out} = 0$, having a slightly higher amplitude due to the fact that the available sample size is lower when $pr_{nm} = 0.95$. Furthermore, it is important to highlight that when there is missing data and outlier contamination, the amplitudes of classical methodology for incomplete time series C_m are higher than the ones of its robust counterpart M_m.

Tables 5.32-5.36 show that the bootstrap methodology was efficient in estimating the values of the mean, of the standard deviation, and of the asymmetry of the parameters. The biases of the bootstrap mean values were slightly larger than the ones of the exact estimates in most of the cases but the standard deviations of the bootstrap estimates

were lower than the exact ones. The asymmetries of the bootstrap procedure had absolute values larger than the asymmetries of the exact methodology in most of the cases. Furthermore, it is important to highlight that is clear the similarity between the confidence intervals obtained by the exact and the bootstrap estimates, this shows that the bootstrap was efficient in mimicking the behavior of the distribution of the parameters. It is possible to note that the bootstrap methodology generates confidence intervals for the classical estimator for complete time series C with coverage percentages closer to 95% than the ones of the exact estimates. As expected, the bootstrap estimates for the methodologies C, M, C_m and M_m have coverage percentages close to 95% in the scenario without outlier contamination and without missing data, which demonstrates the efficiency of these methodologies in this scenario. It is also possible to note that when there is presence of missing data but not of outlier contamination, the classical C_m and the robust M_m estimators for incomplete time series have coverage percentages close to 95% and similar to the ones obtained when $pr_{nm} = 1$ and $pr_{out} = 0$, having in most of the cases a slightly higher amplitude due to the fact that the available sample size is lower when $pr_{nm} = 0.95$. This gives empirical evidence that the local bootstrap in the periodogram for incomplete time series is a good alternative to estimate confidence intervals of parameters of weakly stationary time series in the presence of missing data. However, when there is data contamination by additive outliers and presence of missing data, only the robust methodology for incomplete time series M_m is able to maintain coverage percentages close to 95%, while its classical counterpart C_m performs worse and worse when compared to the robust one as the value of pr_{out} increases. In this context, it is important to emphasize that the confidence intervals of the robust approach for incomplete time series M_m had coverage percentages tending to 95% as the sample size increases while the expected number of outliers is kept constant, i.e., when we go from the scenario with $N = 200$, $pr_{nm} = 0.95$ and $pr_{out} = 0.01$ to the one with $N = 400$, $pr_{nm} = 0.95$ and $pr_{out} = 0.005$, as in this case the outlier effect is diluted with the increase of N. Moreover, it should be noted that for the scenarios with outlier contamination and with missing data, the robust methodology for incomplete time series M_m generated confidence intervals that, when compared to its classical counterpart C_m, in addition to presenting coverage percentages closer to 95%, they also presented lower amplitudes. This gives empirical evidence that the robust local bootstrap in the periodogram for incomplete time series is a good alternative to estimate confidence intervals of parameters of weakly stationary time series in the presence of missing data and of contamination by additive outliers.

5.2.7 Application to Air Pollution Data

The application is performed on a data set of air pollutant variables collected at Automatic Air Quality Monitoring Network (RAMQAr) in the Greater Vitória Region (GVR)

Table 5.27 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{nm} = 1$, $pr_{out} = 0$ and $N = 200$.

ω	I_N	$\overline{\hat{\phi}}$	SD($\hat{\phi}$)	RMSE($\hat{\phi}$)	$\gamma_1(\hat{\phi})$	CI$_{95\%}$($\hat{\phi}$)	A($\hat{\phi}$)	P($\hat{\phi}$)
0	C	0.4826	0.0621	0.0645	-0.1851	(0.3541,0.5982)	0.2441	0.9354
	M	0.4629	0.0646	0.0745	-0.1478	(0.3329,0.5830)	0.2501	-
	C_m	0.4851	0.0616	0.0634	-0.1782	(0.3589,0.6006)	0.2417	-
	M_m	0.4655	0.0642	0.0729	-0.1425	(0.3364,0.5857)	0.2493	-

Table 5.28 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{nm} = 0.95$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N	$\overline{\hat{\phi}}$	SD($\hat{\phi}$)	RMSE($\hat{\phi}$)	$\gamma_1(\hat{\phi})$	CI$_{95\%}$($\hat{\phi}$)	A($\hat{\phi}$)	P($\hat{\phi}$)
0	C_m	0.4837	0.0657	0.0677	-0.1854	(0.3507,0.6088)	0.2581	-
	M_m	0.4612	0.0686	0.0788	-0.1812	(0.3238,0.5905)	0.2667	-
4	C_m	0.4335	0.0767	0.1015	-0.2714	(0.2720,0.5757)	0.3037	-
	M_m	0.4456	0.0695	0.0882	-0.2174	(0.3047,0.5741)	0.2694	-

Table 5.29 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{nm} = 1$, $pr_{out} = 0$ and $N = 400$.

ω	I_N	$\overline{\hat{\phi}}$	SD($\hat{\phi}$)	RMSE($\hat{\phi}$)	$\gamma_1(\hat{\phi})$	CI$_{95\%}$($\hat{\phi}$)	A($\hat{\phi}$)	P($\hat{\phi}$)
0	C	0.4920	0.0436	0.0443	-0.1422	(0.4044,0.5742)	0.1698	0.9446
	M	0.4722	0.0455	0.0533	-0.1151	(0.3795,0.5587)	0.1792	-
	C_m	0.4932	0.0435	0.0440	-0.1400	(0.4056,0.5752)	0.1696	-
	M_m	0.4735	0.0453	0.0525	-0.1127	(0.3814,0.5593)	0.1779	-

Table 5.30 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{nm} = 0.95$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N	$\overline{\hat{\phi}}$	SD($\hat{\phi}$)	RMSE($\hat{\phi}$)	$\gamma_1(\hat{\phi})$	CI$_{95\%}$($\hat{\phi}$)	A($\hat{\phi}$)	P($\hat{\phi}$)
0	C_m	0.4925	0.0464	0.0470	-0.1608	(0.3987,0.5791)	0.1804	-
	M_m	0.4697	0.0482	0.0569	-0.1405	(0.3732,0.5588)	0.1856	-
4	C_m	0.4655	0.0511	0.0617	-0.1403	(0.3614,0.5628)	0.2014	-
	M_m	0.4620	0.0486	0.0617	-0.1195	(0.3657,0.5554)	0.1897	-

Table 5.31 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{nm} = 0.95$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N	$\overline{\hat{\phi}}$	SD($\hat{\phi}$)	RMSE($\hat{\phi}$)	$\gamma_1(\hat{\phi})$	CI$_{95\%}$($\hat{\phi}$)	A($\hat{\phi}$)	P($\hat{\phi}$)
4	C_m	0.4412	0.0545	0.0802	-0.1865	(0.3287,0.5442)	0.2155	-
	M_m	0.4544	0.0496	0.0674	-0.1408	(0.3539,0.5491)	0.1952	-

in the Brazilian state of Espírito Santo, which is made up by nine monitoring stations placed in strategic locations and is responsible for measuring of several atmospheric pollutants and meteorological variables in the area. GVR is comprised of seven cities with a population of approximately 2 million inhabitants in an area of 2319 km^2. The re-

Table 5.32 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{nm} = 1$, $pr_{out} = 0$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.4754(0.0610)	0.0482(0.0086)	-0.2857(0.0921)	(0.3449,0.5882)	0.2433	0.9450
	M	0.4546(0.0636)	0.0492(0.0088)	-0.2729(0.0933)	(0.3241,0.5774)	0.2533	0.9450
	C_m	0.4779(0.0606)	0.0479(0.0086)	-0.2878(0.0968)	(0.3469,0.5901)	0.2432	0.9480
	M_m	0.4572(0.0632)	0.0490(0.0087)	-0.2703(0.0913)	(0.3266,0.5785)	0.2519	0.9450

Table 5.33 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{nm} = 0.95$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C_m	0.4772(0.0641)	0.0520(0.0105)	-0.2941(0.0988)	(0.3411,0.5975)	0.2564	0.9470
	M_m	0.4546(0.0671)	0.0529(0.0106)	-0.2763(0.0981)	(0.3124,0.5751)	0.2627	0.9460
4	C_m	0.4269(0.0757)	0.0545(0.0110)	-0.2600(0.1011)	(0.2705,0.5638)	0.2933	0.8950
	M_m	0.4359(0.0690)	0.0540(0.0106)	-0.2625(0.0994)	(0.2974,0.5646)	0.2672	0.9420

Table 5.34 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{nm} = 1$, $pr_{out} = 0$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C	0.4855(0.0448)	0.0354(0.0049)	-0.2053(0.0605)	(0.3950,0.5697)	0.1747	0.9470
	M	0.4662(0.0467)	0.0360(0.0049)	-0.1952(0.0609)	(0.3627,0.5540)	0.1913	0.9470
	C_m	0.4867(0.0448)	0.0353(0.0049)	-0.2049(0.0608)	(0.3965,0.5708)	0.1743	0.9460
	M_m	0.4675(0.0467)	0.0360(0.0049)	-0.1953(0.0563)	(0.3631,0.5550)	0.1919	0.9460

Table 5.35 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{nm} = 0.95$, $pr_{out} = 0.005$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
0	C_m	0.4885(0.0463)	0.0385(0.0057)	-0.2118(0.0646)	(0.3912,0.5736)	0.1824	0.9460
	M_m	0.4661(0.0479)	0.0394(0.0057)	-0.2003(0.0604)	(0.3707,0.5562)	0.1855	0.9450
4	C_m	0.4610(0.0514)	0.0394(0.0056)	-0.2000(0.0607)	(0.3596,0.5549)	0.1953	0.9070
	M_m	0.4569(0.0481)	0.0395(0.0055)	-0.1918(0.0601)	(0.3611,0.5485)	0.1874	0.9440

Table 5.36 – Bootstrap Estimates for $\phi = 0.5$ with $REP_{MC} = 1000$, $B = 5000$, $pr_{nm} = 0.95$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N^*	$\overline{\hat{\phi}^*}$	$\mathsf{SD}(\hat{\phi}^*)$	$\gamma_1(\hat{\phi}^*)$	$\mathrm{CI}_{95\%}(\overline{\hat{\phi}^*})$	$\mathrm{A}(\overline{\hat{\phi}^*})$	$\mathrm{P}(\overline{\hat{\phi}^*})$
4	C_m	0.4410(0.0544)	0.0406(0.0056)	-0.1868(0.0600)	(0.3281,0.5442)	0.2161	0.8610
	M_m	0.4544(0.0496)	0.0398(0.0055)	-0.1916(0.0555)	(0.3570,0.5454)	0.1884	0.9270

gion is situated along the South Atlantic coast of Brazil (latitude 20°19′15″S, longitude 40°20′10″W) and has a tropical humid climate, with average temperatures ranging from 24 °C to 30 °C. In this paper, we considered the data set of the pollutant Particulate Matter with diameter smaller than 10 µm (PM_{10}), measured hourly, in µg/m^3, collected

at the station located in Downtown Vitória area.

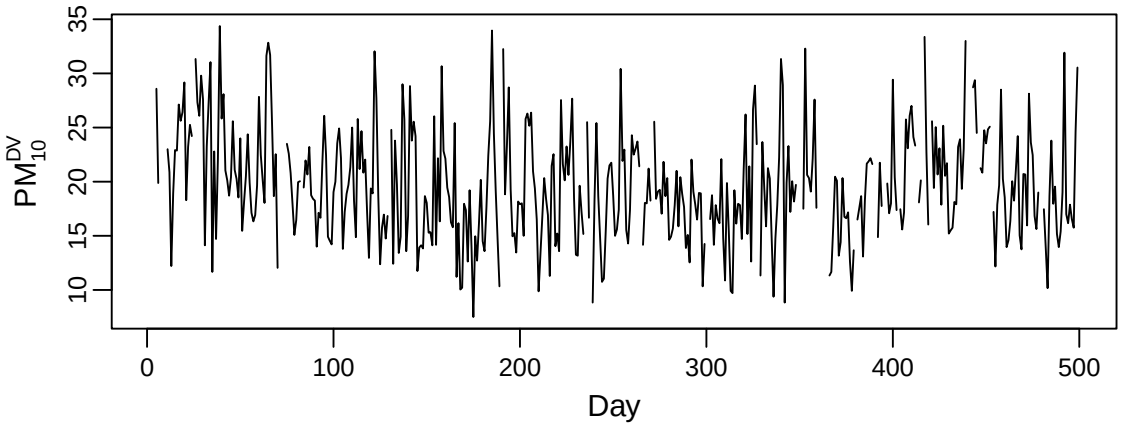


Figure 5.7 – Plot of the PM_{10}^{DV} time series.

In the following the PM_{10} concentrations in the station of Downtown Vitória will be denoted by PM_{10}^{DV}. This data set is composed by daily average concentrations from January 1, 2018 to May 15, 2019, which results in a sample size of $N = 500$. The PM_{10}^{DV} time series plot is shown in Figure 5.7. From this figure, one can see the presence of missing observations, which correspond of 9% of the sample size, as well as of large peaks of PM_{10} concentration which may be analyzed as outliers and, these high concentrations can cause significant damage to some statistics, such as the mean and the standard deviation and, hence, may affect the sample correlation structure as well as the periodogram of the series, provoking misleading results. In order to appropriately choose a time series model to fit PM_{10}^{DV}, we can use (5.30) and (5.31) to obtain via amplitude modulation its ACF and PACF which are shown in Figures 5.8 and 5.9, respectively. Since its ACF exponentially decays and its PACF shows a spike at lag 1 and then cuts off, we have chosen to fit an AR(1) model to the series PM_{10}^{DV}.

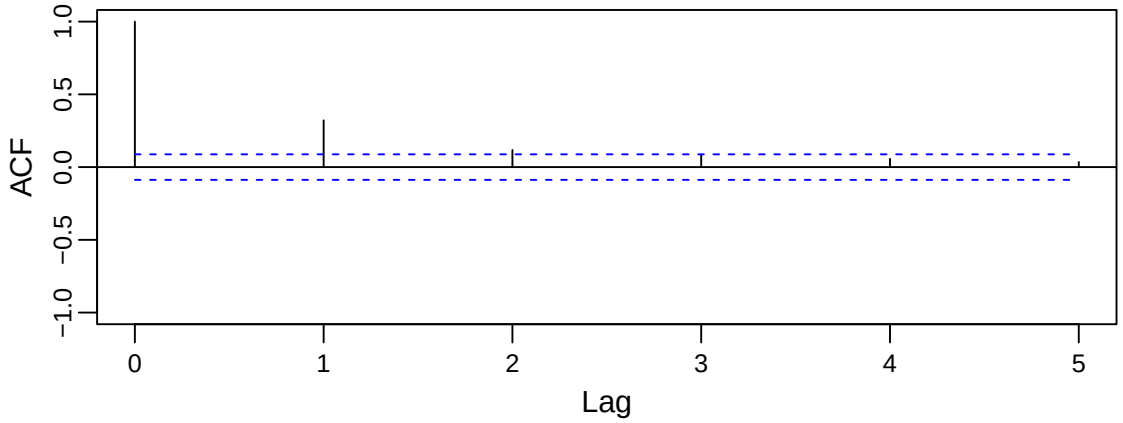


Figure 5.8 – ACF of the amplitude modulated version of the PM_{10}^{DV} time series.

The exact estimates of the AR(1) coefficient are given in Table 5.37. It is important to highlight that, in order to keep consistency with the simulation study, $\delta = 1.345$ was

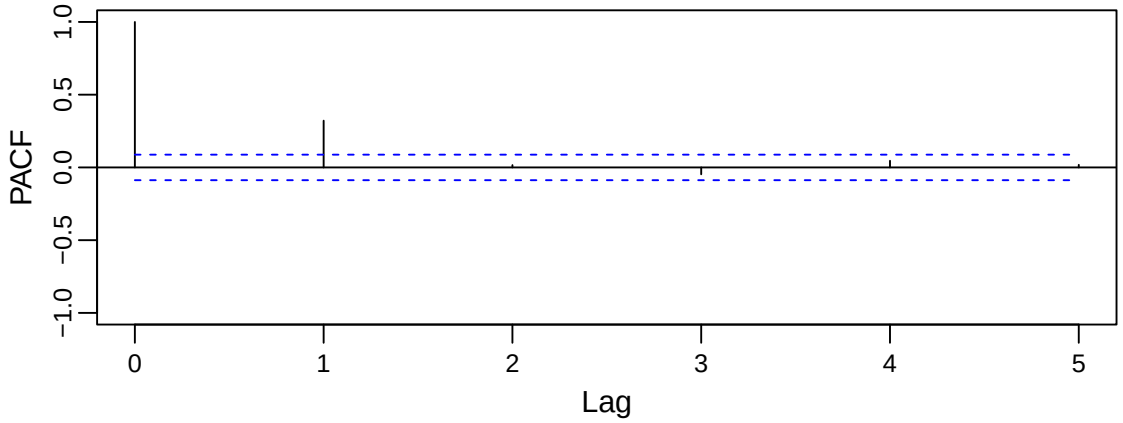


Figure 5.9 – PACF of the amplitude modulated version of the PM_{10}^{DV} time series.

fixed in the Huber function (Equation 5.48). These results clearly showed that the robust method provided a higher coefficient estimate. In this context, it is possible to note that for the AR(1) model the robust estimate of ϕ was 14.7% higher than the classical one. This is evidence that the high concentrations of the pollutant PM_{10} presented the effects of additive outliers in the PM_{10}^{DV} series because there was memory loss in the classical estimate and resistance to outlier contamination in its robust counterpart.

Table 5.37 – Exact estimates of the AR(1) coefficients for the PM_{10}^{DV} time series.

I_Y	$\hat{\phi}$
C_m	0.2932
M_m	0.3364

The bootstrap estimates of the confidence intervals of the estimated AR(1) coefficient for $B = 5000$ are displayed in Table 5.38. It is necessary to emphasize that, similarly to the Monte Carlo study, we have chosen the Bartlett-Priestley kernel with $b_{N,\psi} = b_N = 0.15N^{-0.45}$ to obtain the set of probabilities to choose the periodogram ordinates in the bootstrap procedure. These results show that the confidence interval of the classical method has a left shift in its limits when compared to its robust counterpart. This also gives evidence that the PM_{10}^{DV} time series suffered from the effects of contamination by additive outliers.

Table 5.38 – Bootstrap estimates of the 95% confidence interval of the AR(1) coefficients for the PM_{10}^{DV} time series.

I_Y^*	$CI_{95\%}(\hat{\phi}^*)$
C_m	(0.2109, 0.3649)
M_m	(0.2486, 0.4043)

5.2.8 Conclusions

The classical and the robust versions of the local bootstrap in the periodogram for time series with missing data, presented in this paper, had their finite sample performance compared with each other via a Monte Carlo study. The simulation study had results also for the exact estimates, and it was demonstrated that the exact and the bootstrap estimates provided very similar confidence intervals, which is evidence of the efficiency of the boostrap in mimicking the distribution of the parameters. For the scenario without outlier contamination and without missing data, besides the classical and the robust local bootstrap for incomplete time series, it was considered the classical and the robust local bootstrap for complete time series, the results showed that all the methodologies had coverage percentages close to 95%, which demonstrates that all of them are efficient in this scenario.

It is also important to highlight that when there was presence of missing data but not of outlier contamination, the classical and the robust estimators for incomplete time series had coverage percentages close to 95% and similar to the ones obtained when $pr_{nm} = 1$ and $pr_{out} = 0$. However, when there was contamination by additive outliers in the presence of missing data, the performance of classical bootstrap for incomplete time series was totally affected, while its robust counterpart demonstrated to be very resistant to the contamination, since it was able to maintain the coverage percentages of the confidence intervals close to 95% and presented lower amplitudes than the classical bootstrap for incomplete time series. An analysis of the daily mean concentrations of the PM_{10} collected in the station of Downtown Vitória, in the Brazilian state of Espírito Santo, was considered as an application of the methodologies studied in this article. The conclusion of this analysis was that the memory loss suffered by the classical bootstrap for incomplete time series was the reason for it generate confidence intervals dislocated to the left when compared to the ones obtained by its robust counterpart. Based on these investigations, it is possible to conclude that the classical and the robust version of the local bootstrap in the periodogram for incomplete time series proved to be an alternative for estimating confidence intervals of parameters of models of weakly stationary time series in the presence of missing data which, respectively, are not and are contaminated by additive outliers.

6 Conclusões e Trabalhos Futuros

Nesta tese foram propostas metodologias de bootstrap no domínio da frequência para séries temporais fracamente estacionárias na presença de observações faltantes e/ou de contaminação por observações atípicas aditivas. Os procedimentos de bootstrap sugeridos seguem o princípio do bootstrap local de Paparoditis & Politis (1999), com a robustez sendo atingida pela substituição do periodograma clássico pelo M-periodograma de Reisen, Lévy-Leduc & Taqqu (2017) e quando há presença de dados faltantes se substitui a série temporal original pela sua versão de amplitude modulada proposta por Parzen (1963).

As investigações empíricas por meio de experimentos de Monte Carlo mostraram que as metodologias de bootstrap propostas são eficientes em lidar com a presença de dados faltantes e/ou de outliers aditivos. Isso se deve ao fato de que quando houve contaminação por observações atípicas aditivas, as metodologias clássicas tiveram sua performance completamente afetada, enquanto os procedimentos robustos propostos nesta tese foram capazes de manter porcentagens de cobertura próximas de 95% e apresentaram amplitudes menores que as dos clássicos. Além disso, as metodologias propostas para lidar com a presença de observações faltantes foram capazes de manter porcentagens de cobertura próximas a 95% e similares às obtidas para séries completas em cenários nos quais foram geradas séries incompletas nas simulações de Monte Carlo.

As metodologias de bootstrap propostas nesta tese foram aplicadas em séries temporais de médias diárias do poluente MP_{10} coletado por estações da RAMQAr. Essa análise levou à conclusão que a perda de memória ocorrida nas metodologias clássicas as leva a gerarem intervalos de confiança deslocados para a esquerda quando comparados aos procedimentos robustos. Além disso, é importante ressaltar que para utilizar as metodologias de bootstrap que não foram desenvolvidas para séries incompletas, é necessária a utilização de técnicas de imputação para obter uma série temporal completa, enquanto que as metodologias de bootstrap para séries incompletas não requerem o uso dessas técnicas.

Este estudo cria diversas linhas de pesquisa promissoras e que podem ser perseguidas no futuro, tais como: o estudo da teoria assintótica das metodologias de bootstrap propostas; a extensão dos procedimentos propostos para processos periodicamente estacionários; e a utilização das metodologias propostas para a obtenção de testes de hipótese de periodicidade por meio do bootstrap para cenários com presença de observações faltantes e/ou de contaminação por observações atípicas aditivas.

Referências

ANDERSON, P. L.; MEERSCHAERT, M. M. Parameter estimation for periodically stationary time series. *Journal of Time Series Analysis*, Wiley Online Library, v. 26, n. 4, p. 489–518, 2005. 82

BAIRD, C. *Química Ambiental*. 2. ed. Porto Alegre: Bookman, 2002. 16

BARAKAT, H. M.; NIGM, E. M.; KHALED, O. M.; MOMENKHAN, F. A. Bootstrap method for order statistics and modeling study of the air pollution. *Communications in Statistics - Simulation and Computation*, Taylor & Francis, v. 44, n. 6, p. 1477–1491, 2015. 29

BARBOSA, G. C. *O modelo aditivo generalizado e a técnica de bootstrap: uma associação entre o número de atendimento hospitalar por causas respiratórias e a qualidade do ar*. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, 2009. 28

BASAWA, I.; LUND, R. Large sample properties of parameter estimates for periodic arma models. *Journal of Time Series Analysis*, Wiley Online Library, v. 22, n. 6, p. 651–663, 2001. 82, 85, 87

BLOOMFIELD, P.; HURD, H. L.; LUND, R. B. Periodic correlation in stratospheric ozone data. *Journal of Time Series Analysis*, Wiley Online Library, v. 15, n. 2, p. 127–150, 1994. 82

BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day, 1970. 82

BRAGA, B. et al. *Introdução à Engenharia Ambiental: o Desafio do Desenvolvimento Sustentável*. 2. ed. São Paulo: Editora Pearson, 2005. 16

BREUER, P.; MAJOR, P. Central limit theorems for non-linear functionals of gaussian fields. *Journal of Multivariate Analysis*, v. 13, n. 3, p. 425–441, 1983. ISSN 0047-259X. 108

BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods*. 2nd. ed. New York: Springer Science, 1991. 42, 82, 84, 101, 104, 106, 108

CHOI, Y.-S.; HO, C.-H.; CHEN, D.; NOH, Y.-H.; SONG, C.-K. Spectral analysis of weekly variation in PM_{10} mass concentration and meteorological conditions over China. *Atmospheric Environment*, Elsevier, v. 42, n. 4, p. 655–666, 2008. 27

CONAMA. Dispõe sobre padrães de qualidade do ar. *Diário Oficial da República Federativa do Brasil, edição 223, seção 1*, Brasília, DF, p. 155–156, nov. 2018. 17, 22

DUNSMUIR, W.; ROBINSON, P. Estimation of time series models in the presence of missing data. *Journal of the American Statistical Association*, Taylor & Francis, v. 76, n. 375, p. 560–568, 1981. 11, 12, 23, 57, 58, 59, 60

EFRON, B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 1979. 17, 18, 28, 34, 35, 56, 57

FAJARDO, F. A.; REISEN, V. A.; LÉVY-LEDUC, C.; TAQQU, M. S. M-periodogram for the analysis of long-range-dependent time series. *Statistics*, Taylor & Francis, v. 52, n. 3, p. 665–683, 2018. 27, 36, 39, 63, 91, 100

FLORES, B. E. A pragmatic view of accuracy measurement in forecasting. *Omega*, Elsevier, v. 14, n. 2, p. 93–98, 1986. 96

FOX, A. J. Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 350–363, 1972. 19

FRANKE, J.; HÄRDLE, W. On bootstrapping kernel spectral estimates. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 20, n. 1, p. 121–145, 1992. 35, 57

FULLER, W. A. *Introduction to Statistical Time Series*. New York: Wiley, 1976. 82

GARDNER, W. A.; FRANKS, L. E. Characterization of cyclostationary random signal processes. *IEEE Transactions on Signal Processing*, v. 21, p. 4–14, 1975. 82

GLADYSHEV, E. G. Periodically correlated random sequences. *Sov. Math.*, v. 2, p. 385–388, 1961. 82

GODISH, T. *Air Quality*. 3. ed. Boca Raton: CRC Press, LLC, 1997. 16

GÓMEZ-CARRACEDO, M. P.; ANDRADE, J. M.; LÓPEZ-MAHÍA, P.; MUNIATEGUI, S.; PRADA, D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 134, p. 23–33, 2014. 25

HANNA, S. R. Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmospheric Environment (1967)*, Elsevier, v. 23, n. 6, p. 1385–1398, 1989. 28

HIES, T.; TREFFEISEN, R.; SEBALD, L.; REIMER, E. Spectral analysis of air pollutants. Part 1: elemental carbon time series. *Atmospheric Environment*, Elsevier, v. 34, n. 21, p. 3495–3502, 2000. 26

HOLGATE, S. T.; KOREN, H. S.; SAMET, J. M.; MAYNARD, R. L. *Air Pollution and Health*. San Diego: Academic Press, 1999. 16

HUBER, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 35, n. 1, p. 73–101, 1964. 39, 59, 64

HURD, H. L.; GERR, N. L. Graphical methods for determining the presence of periodic correlation. *Journal of Time Series Analysis*, Wiley Online Library, v. 12, n. 4, p. 337–350, 1991. 82

IBGE. *Sinopse do Censo Demográfico 2010*. Rio de Janeiro, RJ, 2011. 30

IEMA. *Relatório da Qualidade do Ar da Região da Grande Vitória - 2005*. Cariacica, ES, 2006. 30

IGLESIAS, P.; JORQUERA, H.; PALMA, W. Data analysis using regression models with missing observations and long-memory: an application study. *Computational Statistics & Data Analysis*, Elsevier, v. 50, n. 8, p. 2028–2043, 2006. 24

IJSN. *Perfil Regional - Região Metropolitana da Grande Vitória*. Vitória, ES, 2008. 30

JACOBSON, M. *Atmospheric Pollution: History, Science, and Regulation*. [S.l.]: Cambridge University Press, 2002. ISBN 9780521010443. 16

JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, Elsevier, v. 6, n. 3, p. 255–259, 1980. 97

JHUN, I.; COULL, B. A.; SCHWARTZ, J.; HUBBELL, B.; KOUTRAKIS, P. The impact of weather changes on air quality and health in the United States in 1994–2012. *Environmental Research Letters*, IOP Publishing, v. 10, n. 8, p. 1–11, 2015. 28

JUNGER, W. L.; LEON, A. P. d. Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, Elsevier, v. 102, p. 96–104, 2015. 25

KREISS, J.-P.; PAPARODITIS, E. Autoregressive-aided periodogram bootstrap for time series. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 31, n. 6, p. 1923–1955, 2003. 35, 57

KREISS, J.-P.; PAPARODITIS, E. Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, Elsevier, v. 40, n. 4, p. 357–378, 2011. 34, 57

LAHIRI, S. N. *Resampling Methods for Dependent Data*. New York: Springer-Verlag, 2003. 34, 57

LI, T.-H. Laplace periodogram for time series analysis. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 482, p. 757–768, 2008. 99

LI, T.-H. A nonlinear method for robust spectral analysis. *IEEE Transactions on Signal Processing*, IEEE, v. 58, n. 5, p. 2466–2474, 2010. 99

LUND, R.; BASAWA, I. Asymptotics, nonparametrics, and time series. In: _____. [S.l.]: CRC Press, 1999. cap. Modeling and Inference for Periodically Correlated Time Series, p. 37. 85

LUND, R. B.; HURD, H. L.; BLOOMFIELD, P.; SMITH, R. L. Climatological time series with periodic correlation. *Journal of Climate*, American Meteorological Society, v. 8, n. 11, p. 2787–2809, 1995. 82

MA, Y.; GENTON, M. G. Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, Wiley Online Library, v. 21, n. 6, p. 663–684, 2000. 86

MARTIN, M. A.; ROBERTS, S. Bootstrap model averaging in time series studies of particulate matter air pollution and mortality. *Journal of Exposure Science & Environmental Epidemiology*, Nature Publishing Group, v. 16, n. 3, p. 242–250, 2006. 28

MCLEOD, A. Diagnostic checking of periodic autoregression models with application. *Journal of Time Series Analysis*, Wiley Online Library, v. 15, n. 2, p. 221–223, 1994. 84, 95, 98

MILLER, L. et al. Evaluation of missing value methods for predicting ambient BTEX concentrations in two neighbouring cities in Southwestern Ontario Canada. *Atmospheric Environment*, Elsevier, v. 181, p. 126–134, 2018. 26

MOLINARES, F. F.; REISEN, V. A.; CRIBARI-NETO, F. Robust estimation in long-memory processes under additive outliers. *Journal of Statistical Planning and Inference*, Elsevier, v. 139, n. 8, p. 2511–2525, 2009. 35

NOAKES, D. J.; MCLEOD, A. I.; HIPEL, K. W. Forecasting monthly riverflow time series. *International Journal of Forecasting*, Elsevier, v. 1, n. 2, p. 179–190, 1985. 82

PAPARODITIS, E.; POLITIS, D. N. The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, Wiley Online Library, v. 20, n. 2, p. 193–222, 1999. 11, 12, 18, 19, 20, 21, 35, 36, 37, 44, 57, 58, 59, 62, 74

PARZEN, E. On spectral analysis with missing observations and amplitude modulation. *Sankhyā: The Indian Journal of Statistics, Series A*, Indian Statistical Institute, p. 383–392, 1963. 11, 12, 20, 57, 58, 59, 60, 74

POLITIS, D. N.; ROMANO, J. P. The stationary bootstrap. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 89, n. 428, p. 1303–1313, 1994. 28

PRIESTLEY, M. B. *Spectral Analysis and Time Series*. London: Academic Press, 1981. 38, 82

RAO, S. T. et al. Resampling and extreme value statistics in air quality model performance evaluation. *Atmospheric Environment (1967)*, Elsevier, v. 19, n. 9, p. 1503–1518, 1985. 28

REISEN, V. A. et al. An overview of robust spectral estimators. In: *Applied Condition Monitoring*. [S.l.]: Springer International Publishing, 2019. v. 16, p. 204–224. 40, 64

REISEN, V. A.; LÉVY-LEDUC, C.; SOLCI, C. C. Asymptotic properties of the M-regression spectral Whittle estimator for ARMA models. *Paper in compilation*, 2022. 42

REISEN, V. A.; LÉVY-LEDUC, C.; TAQQU, M. S. An M-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference*, Elsevier, v. 187, p. 44–55, 2017. 11, 12, 19, 20, 21, 36, 39, 58, 63, 74, 91, 99

REISEN, V. A. et al. Robust estimation of fractional seasonal processes: Modeling and forecasting daily average SO_2 concentrations. *Mathematics and Computers in Simulation*, Elsevier, v. 146, p. 27–43, 2018. 27, 91

REISEN, V. A.; SARNAGLIA, A. J. Q.; JR, N. C. Reis; LÉVY-LEDUC, C.; SANTOS, J. M. Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility. *Environmental Modelling & Software*, v. 51, p. 286–95, 2014. 83, 91

REISEN, V. A.; SILVA, A. N. *O uso da linguagem R para cálculos de estatística básica*. Vitória: EDUFES, 2011. 32

ROBERTS, S.; MARTIN, M. A. Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies. *Environmental Health Perspectives*, National Institute of Environmental Health Sciences, v. 118, n. 1, p. 131–136, 2010. 28

ROUSSEEUW, P. J.; CROUX, C. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, Taylor & Francis Group, v. 88, n. 424, p. 1273–1283, 1993. 86

SAKAI, H. Circular lattice filtering using pagano's method. *IEEE transactions on acoustics, speech, and signal processing*, v. 30, n. 2, p. 279–287, 1982. 85

SARNAGLIA, A. J. Q.; REISEN, V. A.; BONDON, P. Periodic ARMA models: Application to particulate matter concentrations. In: IEEE. *Signal Processing Conference (EUSIPCO), 2015 23rd European*. [S.l.], 2015. p. 2181–2185. 82, 91

SARNAGLIA, A. J. Q.; REISEN, V. A.; BONDOU, P.; LÉVY-LEDUC, C. A robust estimation approach for fitting a PARMA model to real data. In: IEEE. *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. [S.l.], 2016. p. 1–5. 27, 83

SARNAGLIA, A. J. Q.; REISEN, V. A.; LÉVY-LEDUC, C. Robust estimation of periodic autoregressive processes in the presence of additive outliers. *Journal of Multivariate Analysis*, Elsevier, v. 101, n. 9, p. 2168–2183, 2010. 83, 84, 85, 86, 91, 95, 98

SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. 51, 95

SEBALD, L.; TREFFEISEN, R.; REIMER, E.; HIES, T. Spectral analysis of air pollutants. Part 2: ozone time series. *Atmospheric Environment*, Elsevier, v. 34, n. 21, p. 3503–3509, 2000. 26

SHANG, H. L. Bootstrap methods for stationary functional time series. *Statistics and Computing*, Springer, v. 28, n. 1, p. 1–10, 2018. 29

SHAO, Q. Robust estimation for periodic autoregressive time series. *Journal of Time Series Analysis*, Wiley Online Library, v. 29, n. 2, p. 251–263, 2008. 83, 85, 86, 87, 88, 98

SHAO, Q.; LUND, R. Computation and characterization of autocorrelations and partial autocorrelations in periodic arma models. *Journal of Time Series Analysis*, Wiley Online Library, v. 25, n. 3, p. 359–372, 2004. 85

SHUMWAY, R. H.; STOFFER, D. S. *Time Series Analysis and Its Applications – With R Examples*. 4th. ed. New York: Springer, Cham, 2017. (Springer Texts in Statistics). 82

SINGH, K. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 9, n. 6, p. 1187–1195, 1981. 34

SOUZA, J. B. et al. Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 67, n. 2, p. 453–480, 2018. 83, 91

SOUZA, J. B. d.; REISEN, V. A.; SANTOS, J. M.; FRANCO, G. C. Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. *Revista de Saúde Pública*, SciELO Public Health, v. 48, n. 3, p. 451–458, 2014. 17

TANIGUCHI, M. Minimum contrast estimation for spectral densities of stationary processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 49, n. 3, p. 315–325, 1987. 67

TANIGUCHI, M.; KAKIZAWA, Y. *Asymptotic Theory of Statistical Inference for Time Series*. New York, NY: Springer-Verlag, 2000. 38

TIAO, G.; GRUPE, M. Hidden periodic autoregressive-moving average models in time series data. *Biometrika*, JSTOR, p. 365–373, 1980. 82

WHITTLE, P. Estimation and information in stationary time series. *Arkiv för Matematik*, Springer, v. 2, n. 5, p. 423–434, 1953. 11, 12, 35, 41, 57

WHO - World Health Organization. *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide*. [S.l.], 2005. 16, 17

WU, W. B. M-estimation of linear models with dependent errors. *The Annals of Statistics*, The Institute of Mathematical Statistics, v. 35, n. 2, p. 495–521, 2007. 100, 107

A Estudos Adicionais

Neste apêndice são apresentados dois estudos adicionais desta tese, sendo que um está compilado no formato de artigo científico e o outro é um artigo em compilação.

A.1 Empirical Study of Robust Estimation Methods for PAR Models with Application to the Air Quality Area

Carlo Corrêa Solci Valdério Anselmo Reisen
Alessandro José Queiroz Sarnaglia Pascal Bondon

Abstract

This paper compares three estimators for periodic autoregressive (PAR) models. The first is the classical periodic Yule-Walker estimator (YWE). The second is a robust version of YWE (RYWE) which uses the robust autocovariance function in the periodic Yule-Walker equations, and the third is the robust least squares estimator (RLSE) based on iterative least squares with robust versions of the original time series. The daily mean particulate matter concentration (PM_{10}) data is used to illustrate the methodologies in a real application, that is, in the Air Quality area.

Keywords. Robust estimation; PAR models; Outliers; PM_{10} pollutant.

A.1.1 Introduction

Classical time series analysis generally relies on stationarity assumptions, see, e.g., Fuller (1976), Priestley (1981), Brockwell & Davis (1991), Shumway & Stoffer (2017). Despite the broad use of stationary tools, in some cases, this requirement is too restrictive. Examples of non-stationary phenomena are unit roots, deterministic trends, heteroskedasticity, among others.

The periodic correlation (PC) or cyclostationarity property, introduced by the seminal paper of Gladyshev (1961), deserves special attention due to the fact that this phenomenon is not revealed by usual stationary tools, which may lead to a model misspecification (TIAO; GRUPE, 1980). Due to this fact, special methods to identify the presence of PC in time series have been proposed, see, for example, Hurd & Gerr (1991) and Bloomfield, Hurd & Lund (1994). PC appears in many areas of application: Gardner & Franks (1975) investigate cyclostationarity in electrical engineering; Lund et al. (1995) have found PC in climatological time series; Noakes, McLeod & Hipel (1985) have discovered PC in time series of monthly river flows, among others.

One of the most popular models for PC is the PAR model, which is a generalization of the well-known AR model introduced by Box & Jenkins (1970) where the coefficients and orders vary periodically in time. Estimation methods for the PAR model parameters have been studied by many authors among, Basawa & Lund (2001), Anderson & Meerschaert (2005), Sarnaglia, Reisen & Bondon (2015).

Although the PAR model has been applied in several fields, to the best of our knowledge, it is still relatively unexplored in the air quality research field, especially in the context of contaminated data. Among the air pollutants, the particulate matter with diameter smaller than 10 µm (PM_{10}), is recognized for its effects on human health and is one of the most common and important pollution variables collected by an air quality monitoring network, see, for example, Reisen et al. (2014), Souza et al. (2018) and references therein. Note that, the air quality data usually present asymmetric distributions and large peaks of concentrations. Classical estimators such the sample mean, variance and autocovariance functions are affected by these observations. This suggests to consider robust estimators for PAR parameters, like the ones proposed by Shao (2008), Sarnaglia, Reisen & Lévy-Leduc (2010), Sarnaglia et al. (2016). One issue of this paper is to fit robust PAR models to PM_{10} concentrations.

To the best our knowledge, there are no empirical studies in the literature of cyclostationary processes investigating the behaviour of robust estimators under asymmetric errors and observations which can be identified as atypical or outliers. This paper aims to investigate finite sample properties of such estimators under asymmetric errors and atypical observations through a Monte Carlo study. In addition, a pollutant PM_{10} concentration data set is used as an example of application since it may present observations with high levels of pollutant concentrations that may produce sample distributions with heavy tails.

The rest of the paper is organized as follows: Section A.1.2 introduces the well-known PAR model and describes three estimation methods; Section A.1.3 presents and discusses the results of the Monte Carlo experiment; Section A.1.4 illustrates the use of the estimation methodologies with an application to fit a regression model with PAR errors to PM_{10} concentrations; Section A.1.5 concludes the paper.

A.1.2 The PAR Model and its Estimation Methods

Let $\{Y_t\}$, $t \in \mathbb{Z}$, be a stochastic process with $\mathsf{E}(Y_t^2) < \infty$, $\mu_t = \mathsf{E}(Y_t)$ and autocovariance $\gamma_t(h) = \mathsf{Cov}(Y_t, Y_{t-h})$. Process $\{Y_t\}$ has PC or is a periodically stationary process with period $\mathcal{S} \in \mathbb{N}$ ($PS_{\mathcal{S}}$), if

$$\mu_{t+\mathcal{S}} = \mu_t \quad \text{and} \quad \gamma_{t+\mathcal{S}}(h) = \gamma_t(h), \quad t, h \in \mathbb{Z}, \tag{A.1}$$

$\mathcal{S}$ being the smallest integer satisfying (A.1). Now, let $t = r\mathcal{S} + \nu$, where $r \in \mathbb{Z}$ and $\nu = 1, \ldots, \mathcal{S}$. It follows from (A.1) that $\mu_{r\mathcal{S}+\nu} = \mu_\nu$ and $\gamma_{r\mathcal{S}+\nu}(h) = \gamma_\nu(h)$. Therefore, the autocorrelation $\rho_t(h) = \mathsf{Corr}(Y_t, Y_{t-h})$ satisfies $\rho_{r\mathcal{S}+\nu}(h) = \rho_\nu(h)$. In addition, the partial autocorrelation (PACF) defined as

$$\alpha_t(h) = \mathsf{Corr}(Y_t, Y_{t-h} | Y_{t-1}, \ldots, Y_{t-h+1}) \quad t \in \mathbb{Z}, h \in \mathbb{N},$$

see, e.g., Brockwell & Davis (1991), is also periodic in time, i.e., $\alpha_{r\mathcal{S}+\nu}(h) = \alpha_\nu(h)$. Clearly, the above functions only depend on the period ν and the lag h. When they do not depend on ν, $\{Y_t\}$ is a standard stationary time series in the terminology of Box-Jenkins.For more details, see for example, McLeod (1994), Sarnaglia, Reisen & Lévy-Leduc (2010) and references therein.

The standard stationary linear model can be extended to the $\text{PS}_\mathcal{S}$ process $\{Y_{r\mathcal{S}+\nu}\}$ via

$$Y_{r\mathcal{S}+\nu} = \sum_{j\in\mathbb{Z}} \psi_j(\nu)\epsilon_{r\mathcal{S}+\nu-j},$$

where $\sum_{j\in\mathbb{Z}} |\psi_j(\nu)| < \infty$ for $\nu = 1, \ldots, \mathcal{S}$. The model is causal when $\psi_j(\nu) = 0$ for $j < 0$. In the same way, the model is invertible when

$$\sum_{j\geq 0} \pi_j(\nu) Y_{r\mathcal{S}+\nu-j} = \epsilon_{r\mathcal{S}+\nu},$$

where $\sum_{j\geq 0} |\pi_j(\nu)| < \infty$ for $\nu = 1, \ldots, \mathcal{S}$, see Sarnaglia, Reisen & Lévy-Leduc (2010) and references therein.

The PAR model is a generalization of the well-known AR process and is one of the most used models to fit a $\text{PS}_\mathcal{S}$ time series. The PAR model is given in the following definition.

Definition 1. A zero-mean $\text{PS}_\mathcal{S}$ process $\{Y_{r\mathcal{S}+\nu}\}$ follows a PAR(p_ν) model if it satisfies the difference equation

$$Y_{r\mathcal{S}+\nu} - \sum_{i=1}^{p_\nu} \phi_i(\nu) Y_{r\mathcal{S}+\nu-i} = \sigma_\nu \epsilon_{r\mathcal{S}+\nu}, \tag{A.2}$$

where $\{\epsilon_t\}$ is a sequence of uncorrelated random variables with $\mathsf{E}(\epsilon_t) = 0$, $\mathsf{E}(\epsilon_t^2) = 1$ and, for each cycle $\nu = 1, \ldots, \mathcal{S}$, $\boldsymbol{\phi}_\nu = (\phi_1(\nu), \ldots, \phi_{p_\nu}(\nu))'$ is the AR coefficient vector with order p_ν and σ_ν^2 is the error variance.

Conditions to ensure causality of a PAR model can be derived from its vector AR representation, see, e.g., Sarnaglia, Reisen & Lévy-Leduc (2010). In particular, for a PAR(1) model, the causality condition is

$$\left| \prod_{\nu=1}^{\mathcal{S}} \phi_1(\nu) \right| < 1. \tag{A.3}$$

It is assumed here that $p_1 = \cdots = p_\mathcal{S} = p$ and the following notation $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_\mathcal{S}')' = (\phi_1(1), \ldots, \phi_p(1), \ldots, \phi_1(\mathcal{S}), \ldots, \phi_p(\mathcal{S}))'$.

A.1.2.1 The Yule-Walker Estimator (YWE)

Let $Y_1, \ldots, Y_{n\mathcal{S}}$ be a sample from the process $\{Y_t\}$. The estimates of the periodic mean μ_ν and autocovariance function $\gamma_\nu(h)$ for $\nu = 1, \ldots, \mathcal{S}$ are, respectively, $\bar{Y}_\nu = \frac{1}{n}\sum_{r=0}^{n-1} Y_{r\mathcal{S}+\nu}$, and

$$\hat{\gamma}_\nu(h) = \frac{1}{n}\sum_{r=0}^{n-1}(Y_{r\mathcal{S}+\nu} - \bar{Y}_\nu)(Y_{r\mathcal{S}+\nu-h} - \bar{Y}_{\nu-h}),$$

where $Y_{r\mathcal{S}+\nu-h}$ is set to zero whenever $r\mathcal{S}+\nu-h<1$ or $r\mathcal{S}+\nu-h>n\mathcal{S}$. Therefore, the sample ACF is

$$\hat{\rho}_\nu(h) = \frac{\hat{\gamma}_\nu(h)}{[\hat{\gamma}_\nu(0)\hat{\gamma}_{\nu-h}(0)]^{\frac{1}{2}}}.$$

The sample PACF can be obtained as in Sakai (1982) and Shao & Lund (2004).

The YWE of the PAR coefficients are obtained through the linear equations system

$$\sum_{i=1}^{p}\phi_i(\nu)\gamma_{\nu-i}(h-i) = \gamma_\nu(h), \quad h = 1, \ldots, p, \tag{A.4}$$

in which $\gamma_\nu(h)$ is replaced by $\hat{\gamma}_\nu(h)$. The YWE of $\boldsymbol{\phi}$ is defined as $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}_1', \ldots, \hat{\boldsymbol{\phi}}_{\mathcal{S}}')'$ where, for each ν, $\hat{\boldsymbol{\phi}}_\nu = (\hat{\phi}_1(\nu), \ldots, \hat{\phi}_p(\nu))'$. Asymptotics results for YWE can be derived under the following assumption.

Assumption 1. $\{Y_t\}$ is a zero-mean causal PAR process and $\mathsf{E}(Y_t^4) < \infty$, $t \in \mathbb{Z}$.

The following result is due to Sarnaglia, Reisen & Lévy-Leduc (2010): Under Assumption 1, the YWE estimator $\hat{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \rightsquigarrow \mathcal{N}_{p\mathcal{S}}(0, G), \quad n \to \infty,$$

where the $\rightsquigarrow$ symbol denotes convergence in distribution.

Other asymptotically equivalent estimators of PAR coefficients are the Least Squares Estimator (LSE) (BASAWA; LUND, 2001) and the Maximum Likelihood Estimator (MLE) (LUND; BASAWA, 1999). Due to this equivalence, these estimators will not be considered in this paper.

As well known, the estimators YWE, LSE and MLE, are not resistant in the presence of atypical observations (outliers) see, for example, Shao (2008), Sarnaglia, Reisen & Lévy-Leduc (2010). The lack of robustness of classical estimators has motivated the investigation of robust approaches in the literature. In the following subsections the robust methods introduced by Sarnaglia, Reisen & Lévy-Leduc (2010) and Shao (2008) are summarized and the empirical comparison with YWE estimator is presented in the Simulation Section.

A.1.2.2 The Robust Yule-Walker Estimator (RYWE)

Sarnaglia, Reisen & Lévy-Leduc (2010) have proposed a Robust Yule-Walker Estimator (RYWE) which is based on the autocovariance function proposed in Ma & Genton (2000). Let $\mathsf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the robust scale estimator of y proposed by Rousseeuw & Croux (1993), $Q_n(\mathsf{y})$, is defined as the following kth order statistic

$$Q_n(\mathsf{y}) = d\left\{|y_i - y_j|\,; 1 \leq i < j \leq n\right\}_{(k)}, \tag{A.5}$$

where d is a constant factor to ensure Fisher-consistency and $k = \binom{c}{2} \approx 0.25\binom{n}{2}$, where $c = [n/2] + 1$ is half of the size n of the vector y. For Gaussian random variables, $d = 2.2191$. Given a $\mathrm{PS}_{\mathcal{S}}$ time series, $Y_1, \ldots, Y_{n\mathcal{S}}$, based on (A.5) Sarnaglia, Reisen & Lévy-Leduc (2010) define the robust sample (periodic) ACV function by

$$\tilde{\gamma}_\nu(h) = \frac{1}{4}[Q^2_{n-r+1}(\mathsf{u}_\nu + \mathsf{v}_\nu) - Q^2_{n-r+1}(\mathsf{u}_\nu - \mathsf{v}_\nu)], \quad 0 \leq h < [(n-1)\mathcal{S} + \nu], \tag{A.6}$$

where $\mathsf{u}_\nu = (Y_{r\mathcal{S}+\nu-h}, \ldots, Y_{(n-1)\mathcal{S}+\nu-h})$, $\mathsf{v}_\nu = (Y_{r\mathcal{S}+\nu}, \ldots, Y_{(n-1)\mathcal{S}+\nu})$. Note that, $\hat{\gamma}_\nu(h)$ does not posses the positive definite property. The RYWE is defined similarly to the YWE. For each $\nu = 1, \ldots, \mathcal{S}$, replace, in (A.4), the theoretical ACV, $\gamma_\nu(h)$, by its sample robust estimator in Equation A.6, $\tilde{\gamma}_\nu(h)$, and solve the resulting linear equations system

$$\sum_{i=1}^{p} \phi_i(\nu)\tilde{\gamma}_{\nu-i}(h-i) = \tilde{\gamma}_\nu(h), \quad k = 1, \ldots, p, \tag{A.7}$$

for $\phi_1(\nu), \ldots, \phi_p(\nu)$. The RYWE estimator is defined as $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\phi}}'_1, \ldots, \tilde{\boldsymbol{\phi}}'_{\mathcal{S}})'$. The white noise variance σ^2_ν can also be robustly estimated by the same argument as in Equation A.4 with $h = 0$.

Assumption 2. For any $\nu = 1, \ldots, \mathcal{S}$, $\{Y_{r\mathcal{S}+\nu};\ r \in \mathbb{Z}\}$ is a mean-zero Gaussian process with strong mixing coefficients α_n satisfying: $\alpha_n \leq Cn^{-a}$, for some $a > 1$ and $C \geq 1$.

The following result is due to Sarnaglia, Reisen & Lévy-Leduc (2010). Under Assumptions 1 and 2, the RYWE estimator $\tilde{\phi}_i(\nu)$ satisfies

$$\tilde{\phi}_i(\nu) - \phi_i(\nu) = O_p(n^{-1/2}),$$

for $i = 1, \ldots, p$, $\nu = 1, \ldots, \mathcal{S}$.

A.1.2.3 The Robust Least Squares Estimator (RLSE)

Shao (2008) proposes an alternative to the conditional Least Squares Estimator (LSE). The LSE of $\boldsymbol{\phi}$ can be defined as the solution of the $p\mathcal{S}$-dimensional estimating equations

$$S_n(\boldsymbol{\phi}_\nu) = \frac{1}{\sigma_\nu}\sum_{r=0}^{n-1} \epsilon_{r\mathcal{S}+\nu} Y_{r\mathcal{S}+\nu-i} = 0, \quad 1 \leq i \leq p, \quad 1 \leq \nu \leq \mathcal{S}, \tag{A.8}$$

where the error terms are given by $\epsilon_{r\mathcal{S}+\nu} = \epsilon_{r\mathcal{S}+\nu}(\boldsymbol{\phi}_\nu) = (Y_{r\mathcal{S}+\nu} - \sum_{i=1}^{p} \phi_i(\nu) Y_{r\mathcal{S}+\nu-i})/\sigma_\nu$, $0 \leq r < n$, $1 \leq \nu \leq \mathcal{S}$. Asymptotic properties for the LSE have been studied by Basawa & Lund (2001).

Shao (2008) aims to achieve robustness by replacing, in Equation (A.8), $\epsilon_{r\mathcal{S}+\nu}$ and $Y_{r\mathcal{S}+\nu}$ by their robust versions $\breve{\epsilon}_{r\mathcal{S}+\nu}$ and $\breve{Y}_{r\mathcal{S}+\nu}$, respectively, which are defined as

$$\breve{\epsilon}_{r\mathcal{S}+\nu} = \psi(\epsilon_{r\mathcal{S}+\nu}) \tag{A.9}$$

and

$$\breve{Y}_{r\mathcal{S}+\nu} = \begin{cases} Y_{r\mathcal{S}+\nu}, & \text{if } \breve{\epsilon}_{r\mathcal{S}+\nu} = \epsilon_{r\mathcal{S}+\nu}, \\ \sum_{i=1}^{p} \phi_i(\nu) \breve{Y}_{r\mathcal{S}+\nu-i} + \sigma_\nu \breve{\epsilon}_{r\mathcal{S}+\nu}, & \text{if } \breve{\epsilon}_{r\mathcal{S}+\nu} \neq \epsilon_{r\mathcal{S}+\nu}. \end{cases} \tag{A.10}$$

Therefore, the Robust LSE (RLSE) is the solution of the robustified estimating equations

$$\breve{S}_n(\boldsymbol{\phi}_\nu) = \frac{1}{\sigma_\nu} \sum_{r=0}^{n-1} \psi\left(\frac{Y_{r\mathcal{S}+\nu} - \sum_{i=1}^{p} \phi_i(\nu) Y_{r\mathcal{S}+\nu-i}}{\sigma_\nu} \right) \breve{Y}_{r\mathcal{S}+\nu-i} = 0, \quad 1 \leq i \leq p, \quad 1 \leq \nu \leq \mathcal{S}. \tag{A.11}$$

Shao (2008) has considered $\psi(\cdot)$ as the Huber type function, defined by

$$\psi(x) = \psi_c(x) = \begin{cases} x, & \text{if } |x| \leq c, \\ c\,\text{sign}(x), & \text{if } |x| > c, \end{cases} \tag{A.12}$$

because it is monotonic, which ensures existence and uniqueness of solution to Equation A.11. Nevertheless, any odd bounded and differentiable function can be a candidate for the $\psi(\cdot)$ function, including the so-called redescending functions, such as Bisquare, Hampel, Generalized Gauss-Weight, among others. However, they have many zeroes, which may lead to non-optimal solutions.

Assumption 3. The marginal density function $f_\epsilon(\cdot)$ of the error $\epsilon_{r\mathcal{S}+\nu}$ in Equation A.2 is symmetric about the origin.

The following result has been derived by Shao (2008): Under Assumptions 1 and 3, the RLSE $\breve{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\breve{\boldsymbol{\phi}} - \boldsymbol{\phi}) \rightsquigarrow \mathcal{N}_{p\mathcal{S}}(0, A), \quad n \to \infty,$$

where the covariance matrix A is given in Equation 14 of Shao (2008).

In practice, the estimates are obtained using the following iterative procedure starting with an appropriate initial guess value for the RLSE. Suppose $\breve{\boldsymbol{\phi}}^{(l)}$ represents the vector of estimates at the lth iteration. Then, at the $(l+1)$th iteration, calculate the residuals

$$e_{r\mathcal{S}+\nu}^{(l)} = Y_{r\mathcal{S}+\nu} - \sum_{i=1}^{p} \breve{\phi}_i^{(l)}(\nu) Y_{r\mathcal{S}+\nu-i}, \quad 1 \leq \nu \leq \mathcal{S},$$

where $Y_t = 0$, $t \leq 0$, estimate the white noise standard deviation at the period ν, σ_ν, by

$$\breve{\sigma}_\nu^{(l)} = \text{Median}\left(\left| e_\nu^{(l)} \right|, \left| e_{\mathcal{S}+\nu}^{(l)} \right|, \ldots, \left| e_{(n-1)\mathcal{S}+\nu}^{(l)} \right| \right), \quad 1 \leq \nu \leq \mathcal{S},$$

calculate the robust version of $e_{r\mathcal{S}+\nu}^{(l)}$ through (A.9), $\breve{e}_{r\mathcal{S}+\nu}^{(l)} = \psi(e_{r\mathcal{S}+\nu}^{(l)})$, obtain $\breve{Y}_{r\mathcal{S}+\nu}^{(l)}$ by (A.10) with $\breve{\epsilon}_{r\mathcal{S}+\nu}$ substituted for the robust residual $\breve{e}_{r\mathcal{S}+\nu}^{(l)}$ and σ_ν replaced with $\breve{\sigma}_\nu^{(l)}$, and evaluate the solution $\breve{\boldsymbol{\phi}}^{(l+1)}$ of the robustified estimating equations in (A.11) replacing $\breve{Y}_{r\mathcal{S}+\nu}$ with $\breve{Y}_{r\mathcal{S}+\nu}^{(l)}$ and σ_ν with $\breve{\sigma}_\nu^{(l)}$. Stop the procedure according to some convergence criterion.

A.1.3 Monte Carlo Study

In order to investigate the impact of atypical observations on the estimates obtained from the methods discussed previously, series of periodically stationary processes were generated with and without additive outliers. Let $\{X_{r\mathcal{S}+\nu}\}$ be defined as follows

$$X_{r\mathcal{S}+\nu} = Y_{r\mathcal{S}+\nu} + \omega V_{r\mathcal{S}+\nu} \tag{A.13}$$

where $\{Y_{r\mathcal{S}+\nu}\}$ is a PAR model with $\mathcal{S} = 4$ and coefficients given in Table A.1. The parameter values were chosen to have examples of time series models with low (Model 1) and strong (Model 2) correlation dependencies, that is, Model 2 is closer to the non-causality region than Model 1. $\{V_t\}$ is a sequence of independent random variables with $\mathsf{P}(V_t = -1) = \mathsf{P}(V_t = 1) = \xi/2$ and $\mathsf{P}(V_t = 0) = 1 - \xi$, $0 \leq \xi < 1$, Y_t and V_s are independent processes for all t, s and ω is the magnitude of the outlier. The sample sizes were taken as small ($N = 400$) and large ($N = 1600$), .i.e., $n = 100$ and $n = 400$ cycles, respectively, which are common sample sizes in practical situation. The initial value for RLSE was taken as the true parameter vector.

The simulation study is divided in these two cases; uncontaminated and contaminated series. The contaminated series were generated from model in (A.13) with the following specifications: $\omega = 7$ and $\xi = 0.01$. The effect of the normality departure in the white noise sequence was also studied by generating the random variables ϵ_t such that $\epsilon_t \sim \mathcal{N}(0, 1)$ and $\sqrt{2}\epsilon_t + 1 \sim \chi^2_{(1)}$. In both cases, $\mathsf{E}(\epsilon_t) = 0$ and $\mathsf{E}(\epsilon_t^2) = 1$. For each of these scenarios, 1000 replicates of $\{Y_t\}$ were generated to compute the mean of the empirical Bias and Root Mean Squared Error (RMSE). For the RLSE, according to Shao (2008), $c = 3.06$ was fixed in the Huber function (Equation A.12) such that, residuals greater than 3.06 (in absolute value) are regarded as outliers. Other model configurations were also considered in the simulation study such as heavy-tailed distributions, different period lengths, sample sizes and coefficient values and other outlier magnitudes. However, in general, the results leaded to similar conclusions and are not displayed here, but they are available upon request.

Tables A.2 and A.3 display the Bias and RMSE for Models 1 and 2, respectively. For illustration purpose, the empirical distributions of $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$, $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ and $\sqrt{n}(\breve{\phi}_i(\nu) - \phi_i(\nu))$ for Model 1, with $n = 400$, under the uncontaminated and contaminated Gaussian scenarios, are presented in Figures A.1 and A.2, respectively. The same

Table A.1 – Parameters of PAR(1) models used in the simulation.

Parameter	Model 1	Model 2
$\phi_1(1)$	0.9	1.5
$\phi_1(2)$	0.8	0.8
$\phi_1(3)$	0.7	1.2
$\phi_1(4)$	0.6	0.5
λ	0.3024	0.7200

plots for the uncontaminated case with asymmetric errors are displayed in Figure A.3. For Model 1 (Table A.2), under the uncontaminated Gaussian case, the reduction of the Bias and RMSE when increasing the sample size suggests that all estimators are consistent. It is observed that, in this scenario, YWE and RLSE present better results. The findings for the asymmetric uncontaminated case show that the RYWE is extremely affected by skewness of the data, presenting a persistent Bias which does not seem to vanish by increasing the sample size. The results for heavy tailed errors are similar, which show that the RYWE method is very sensitive to departures from normality. This indicates that the normality requirement in Assumption 2 is crucial to ensure asymptotic properties of the RYWE. The YWE and RLSE do not seem to be affected by non-Gaussian errors (both asymmetric and heavy tailed). This gives empirical evidence that the technical requirement of symmetric errors (Assumption 3) may be over restrictive to ensure asymptotic normality of the RLSE. As expected, atypical observations increase the Bias and RMSE of the YWE. Under normally distributed errors, both RYWE and RLSE show robustness with Bias and RMSE almost unchanged with the presence of outliers. For asymmetric errors, the RLSE is the only one which presents good performance.

The conclusions for Model 2 (Table A.3) are similar to the previous case. It is worth noting that in this stronger dependence scenario, there is a overall reduction of the Bias and RMSE. Another remarkable fact is that, in this case, the RYWE does not seem to be strongly affected by asymmetric errors in both uncontaminated and contaminated scenarios.

From Figure A.1, it can be seen that the empirical distributions are virtually the same and they have shape very close to the N(0,1) distribution. This corroborates the asymptotic results of the standardized estimators even for a small sample size. However, the scale of the RYWE distribution is slightly greater than of those of the YWE and RLSE. Figure A.2 shows the robustness to outliers of RYWE and RLSE methods under Gaussian errors, while the distribution of YWE is shifted to the left due to the well-known memory loss property. Its scale is also increased as a result of the contamination. Figure A.3 illustrates the prominent shift to the right of the RYWE distribution caused by the skewness of the errors.

Table A.2 – Bias and RMSE for Model 1 and outliers with probability $\xi = 0.01$.

ω	ϵ_t	n	$\phi_1(\nu)$	YWE		RYWE		RLSE	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
0	$\mathcal{N}(0,1)$	100	0.9	-0.007	0.077	-0.003	0.103	0.003	0.079
			0.8	-0.002	0.065	0.004	0.084	-0.002	0.065
			0.7	0.000	0.063	-0.001	0.083	-0.001	0.063
			0.6	-0.005	0.066	-0.003	0.083	-0.005	0.067
		400	0.9	-0.001	0.037	-0.001	0.047	0.002	0.038
			0.8	-0.001	0.031	0.000	0.038	-0.001	0.031
			0.7	-0.001	0.032	0.001	0.038	-0.001	0.032
			0.6	0.000	0.032	0.000	0.039	0.000	0.033
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	0.9	-0.006	0.076	0.178	0.209	0.006	0.063
			0.8	-0.007	0.065	0.117	0.147	-0.003	0.055
			0.7	-0.004	0.065	0.088	0.119	-0.003	0.052
			0.6	-0.005	0.069	0.077	0.108	-0.004	0.056
		400	0.9	-0.001	0.037	0.179	0.185	0.002	0.030
			0.8	0.000	0.033	0.115	0.122	0.000	0.026
			0.7	-0.001	0.033	0.089	0.096	-0.001	0.026
			0.6	-0.001	0.034	0.085	0.093	0.000	0.028
7	$\mathcal{N}(0,1)$	100	0.9	-0.181	0.247	0.014	0.120	-0.031	0.095
			0.8	-0.118	0.176	0.012	0.096	-0.027	0.076
			0.7	-0.105	0.157	0.015	0.091	-0.024	0.077
			0.6	-0.097	0.151	0.012	0.091	-0.027	0.081
		400	0.9	-0.183	0.203	0.017	0.055	-0.027	0.050
			0.8	-0.129	0.144	0.012	0.046	-0.021	0.041
			0.7	-0.108	0.124	0.013	0.044	-0.019	0.041
			0.6	-0.103	0.119	0.014	0.043	-0.022	0.043
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	0.9	-0.172	0.243	0.213	0.251	-0.019	0.076
			0.8	-0.126	0.180	0.142	0.175	-0.021	0.063
			0.7	-0.105	0.158	0.112	0.144	-0.018	0.061
			0.6	-0.096	0.151	0.103	0.134	-0.023	0.067
		400	0.9	-0.182	0.202	0.211	0.219	-0.021	0.041
			0.8	-0.129	0.145	0.142	0.149	-0.017	0.033
			0.7	-0.112	0.127	0.111	0.119	-0.017	0.033
			0.6	-0.106	0.121	0.106	0.114	-0.019	0.037

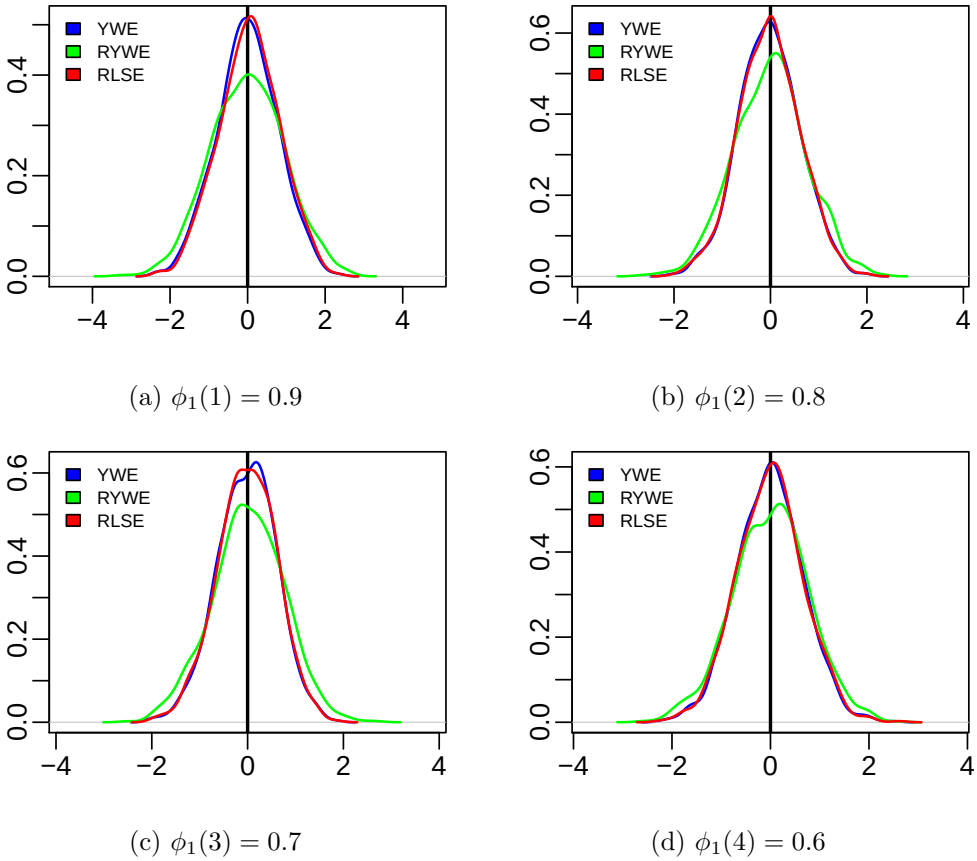
A.1.4 An Application to the Air Quality Area (the PM_{10} Data)

The application is based on a data set (air pollutant variables) collected at Automatic Air Quality Monitoring Network (RAMQAr) in the Greater Vitória Region GVR-ES, Brazil, which is composed by nine monitoring stations placed in strategic locations and accounts for the measuring of several atmospheric pollutants and meteorological variables in the area. GVR is comprised of seven cities with a population of approximately 1.9 million inhabitants in an area of 2319 km^2. The region is situated along the South Atlantic coast of Brazil (latitude 20°19′15″S, longitude 40°20′10″W) and has a tropical humid

Table A.3 – Bias and RMSE for Model 2 and outliers with probability $\xi = 0.01$.

				YWE		RYWE		RLSE	
ω	ϵ_t	n	$\phi_1(\nu)$	Bias	RMSE	Bias	RMSE	Bias	RMSE
0	$\mathcal{N}(0,1)$	100	1.5	-0.009	0.055	-0.009	0.100	0.000	0.055
			0.8	-0.004	0.033	-0.005	0.050	-0.004	0.034
			1.2	-0.006	0.040	-0.008	0.066	-0.006	0.040
			0.5	-0.008	0.033	-0.008	0.043	-0.008	0.033
		400	1.5	-0.003	0.026	-0.003	0.037	-0.001	0.026
			0.8	-0.001	0.016	-0.001	0.021	-0.001	0.016
			1.2	-0.002	0.019	-0.002	0.027	-0.002	0.020
			0.5	-0.002	0.015	-0.003	0.018	-0.003	0.015
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	1.5	-0.009	0.055	0.025	0.098	0.001	0.043
			0.8	-0.007	0.036	0.011	0.048	-0.005	0.028
			1.2	-0.007	0.040	0.019	0.069	-0.005	0.033
			0.5	-0.009	0.033	-0.004	0.039	-0.006	0.027
		400	1.5	-0.002	0.026	0.032	0.050	0.000	0.021
			0.8	-0.001	0.016	0.016	0.026	-0.001	0.013
			1.2	-0.002	0.019	0.023	0.035	-0.002	0.015
			0.5	-0.003	0.015	0.003	0.017	-0.002	0.012
7	$\mathcal{N}(0,1)$	100	1.5	-0.174	0.246	-0.023	0.122	-0.014	0.061
			0.8	-0.047	0.076	-0.011	0.058	-0.014	0.040
			1.2	-0.079	0.125	-0.013	0.079	-0.013	0.048
			0.5	-0.029	0.056	-0.010	0.046	-0.013	0.038
		400	1.5	-0.167	0.189	-0.020	0.052	-0.013	0.032
			0.8	-0.042	0.050	-0.007	0.025	-0.007	0.019
			1.2	-0.080	0.093	-0.015	0.038	-0.008	0.022
			0.5	-0.023	0.032	-0.006	0.021	-0.007	0.018
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	1.5	-0.180	0.253	0.022	0.120	-0.013	0.054
			0.8	-0.049	0.081	0.011	0.057	-0.008	0.032
			1.2	-0.082	0.126	0.016	0.083	-0.009	0.036
			0.5	-0.029	0.053	-0.003	0.041	-0.010	0.028
		400	1.5	-0.172	0.193	0.022	0.054	-0.010	0.025
			0.8	-0.041	0.050	0.015	0.028	-0.005	0.015
			1.2	-0.083	0.097	0.012	0.039	-0.007	0.018
			0.5	-0.024	0.033	0.003	0.020	-0.005	0.014

climate, with average temperatures ranging from 24 °C to 30 °C. This data set has been previously investigated in different contexts of time series modeling, such as periodic models, robustness in long-memory models, heteroskedastic long memory process, time series regression with principal component analysis, among others. See, for instance, Sarnaglia, Reisen & Bondon (2015), Reisen et al. (2014), Sarnaglia, Reisen & Lévy-Leduc (2010), Souza et al. (2018), Fajardo et al. (2018), Reisen et al. (2018), Reisen, Lévy-Leduc & Taqqu (2017) and references therein. The data set considered in this paper is the pollutant Particulate Matter with diameter smaller than 10 µm (PM_{10}), measured hourly, in µg/m^3, collected at the station located in Enseada do Suá area.

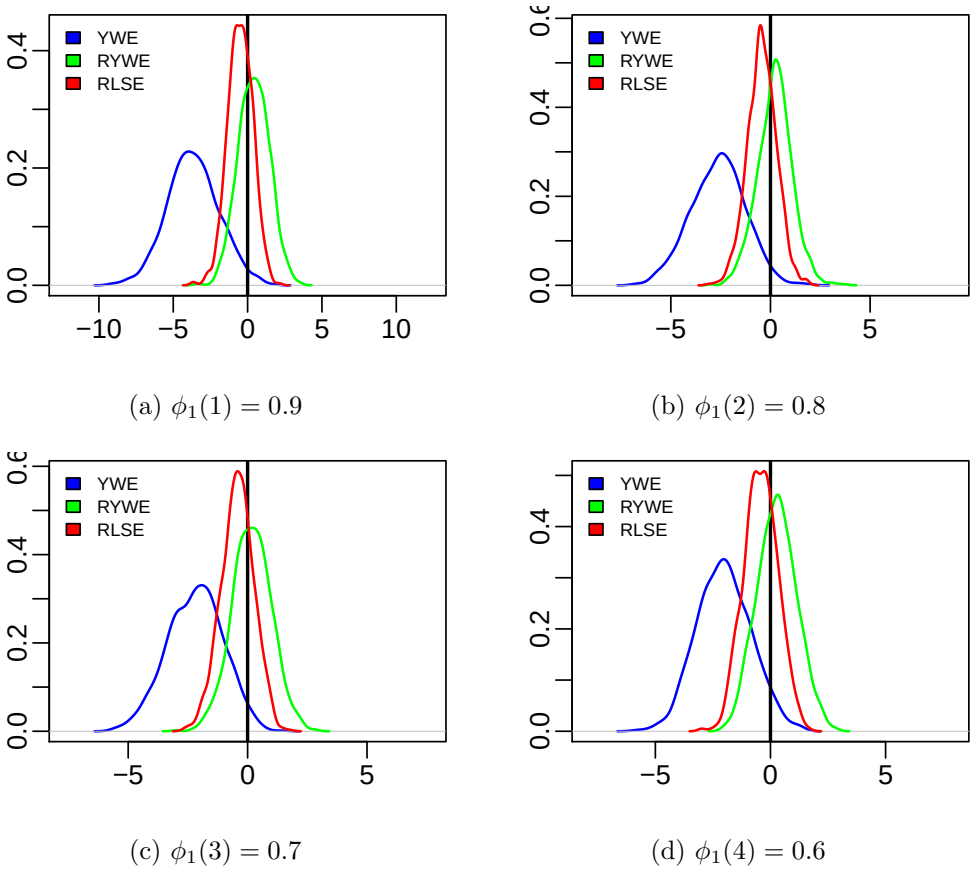


(a) $\phi_1(1) = 0.9$ (b) $\phi_1(2) = 0.8$

(c) $\phi_1(3) = 0.7$ (d) $\phi_1(4) = 0.6$

Figure A.1 – Empirical Distributions of $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (blue lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (green lines) and the $\sqrt{n}(\breve{\phi}_i(\nu) - \phi_i(\nu))$ (red lines) for Model 1 with $\omega = 0$, $n = 400$ and normal errors.

The PM_{10} data set corresponds to daily average concentrations from January 1st, 2014 to December 29th, 2015 which kept the sample size multiple of the natural choice to the period length $\mathcal{S} = 7$. Due to skewness and some evidences of time varying variance, the natural logarithm transformation (log) was used and the plot of the $\log(PM_{10})$ is displayed in Figure A.4. From this figure, one can see large peaks of PM_{10} concentration which may be viewed here as outliers and, as mentioned previously, these high levels can provoke serious damage to some statistics, such as the mean and the standard deviation and, therefore, may affect the sample correlation structure of the series, causing misleading results. The existence of any outlier's effect will be discussed in the estimation parameter model (next subsection). It can also be seen the presence of sinusoidal deterministic trends. Analysis of the periodogram (Figure A.5) corroborates to this result and indicates that the frequency $2/N$, corresponding approximately to a yearly cycle, has a large contribution to the overall variance of the data. The high frequency peaks of the periodogram correspond to weekly periodicity and, according to the daily periodic boxplots displayed in Figure A.6, they can be explained by a level decrease in the weekends. This is an expected finding due to fact that the traffic and civil construction decrease in



(a) $\phi_1(1) = 0.9$ (b) $\phi_1(2) = 0.8$

(c) $\phi_1(3) = 0.7$ (d) $\phi_1(4) = 0.6$

Figure A.2 – Empirical Distributions of $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (blue lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (green lines) and the $\sqrt{n}(\breve{\phi}_i(\nu) - \phi_i(\nu))$ (red lines) for Model 1 with $\omega = 7$, $n = 400$ and normal errors.

the region in the weekend days.

The above preliminary analysis of the series suggests that a deterministic trend must be firstly removed from $\log(\text{PM}_{10})$ before further analysis and this is discussed in the next subsection in which a linear model with errors following a PAR model is fitted to the series.

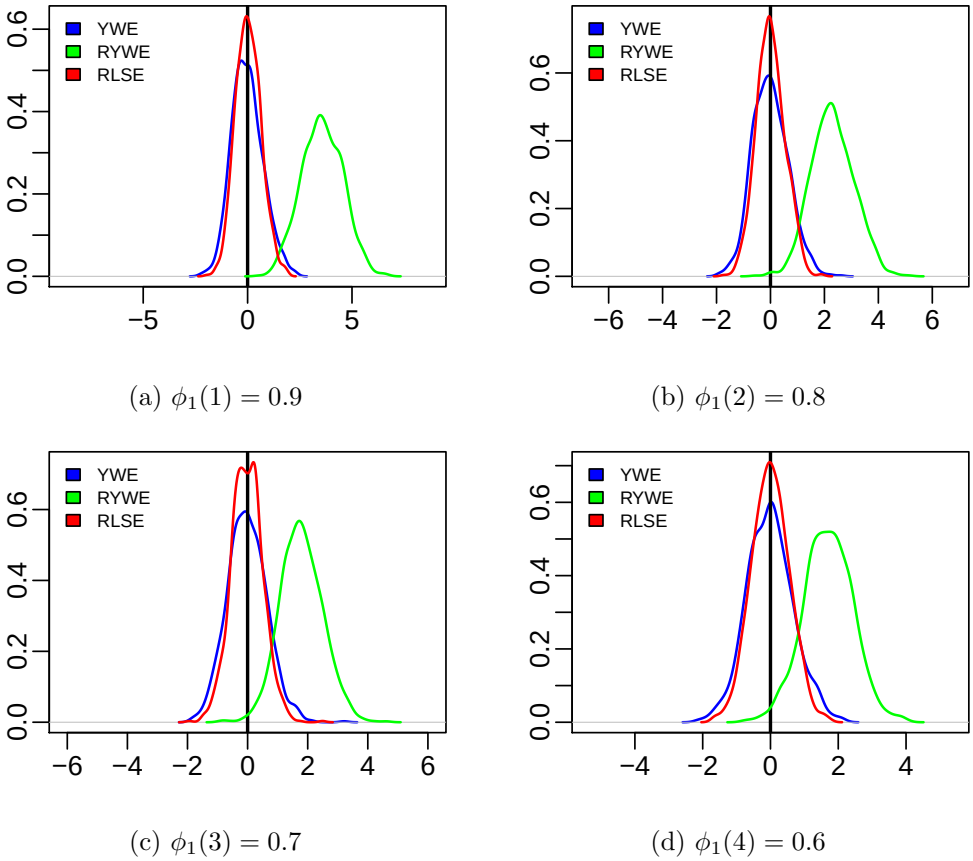
A.1.4.1 Estimated Model

According to the previous statistical analysis of the $\log(\text{PM}_{10})$ series, the following model is suggested here to fit the data

$$\log(\text{PM}_{10,t}) = \mu + \alpha_1 \text{sat}_t + \alpha_2 \text{sun}_t + \beta_1 \cos_t + \beta_2 t + Y_t; \tag{A.14}$$

$$Y_t = \sum_{i=1}^{p_t} \phi_i(t) Y_{t-i} + \sigma_t \epsilon_t, \tag{A.15}$$

with the sinusoidal covariate: $\cos_t = \cos(\frac{2\pi t}{365.25})$, $t = 1, \ldots, 728$; the linear term t; and a "day of the week" factor with the levels: Week (the reference level); Saturday (represented by the dummy variable sat_t which takes value 1 for Saturdays); and Sunday (sun_t which



(a) $\phi_1(1) = 0.9$ (b) $\phi_1(2) = 0.8$

(c) $\phi_1(3) = 0.7$ (d) $\phi_1(4) = 0.6$

Figure A.3 – Empirical Distributions of $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (blue lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (green lines) and the $\sqrt{n}(\breve{\phi}_i(\nu) - \phi_i(\nu))$ (red lines) for Model 1 with $\omega = 0$, $n = 400$ and asymmetric errors.

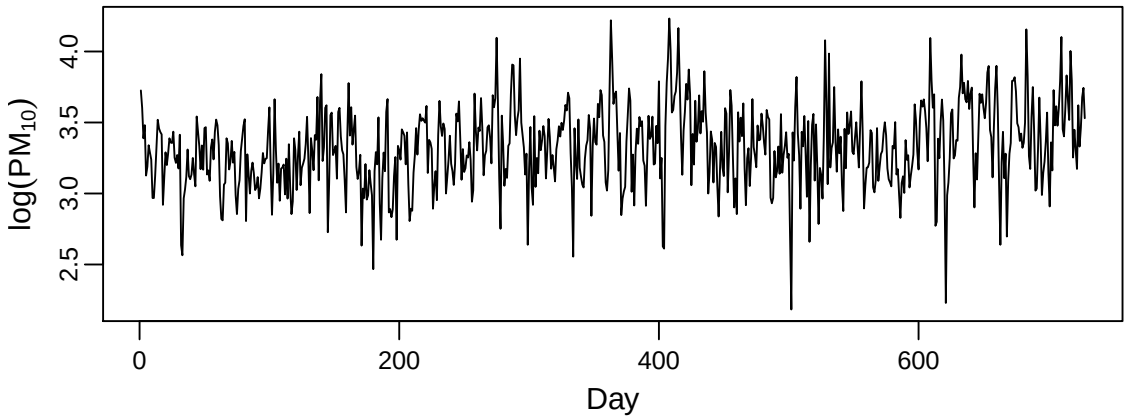


Figure A.4 – Plot of the $\log(\text{PM}_{10})$ time series.

takes value 1 for Sundays) with $\mathcal{S} = 7$. We also considered $\sin_t = \sin(\frac{2\pi t}{365.25})$, $t = 1, \ldots, N$, which turn out to be insignificant. The above model means that in the business days the regular level of $\log(\text{PM}_{10})$ is μ, on Saturdays it suffers an increase of α_1 and on Sundays it is increased by α_2 and it has a long-run cyclic trend, represented by $\cos_t$ and a linear term t.

The model in Equations A.14 and A.15 will be fitted based on following two steps

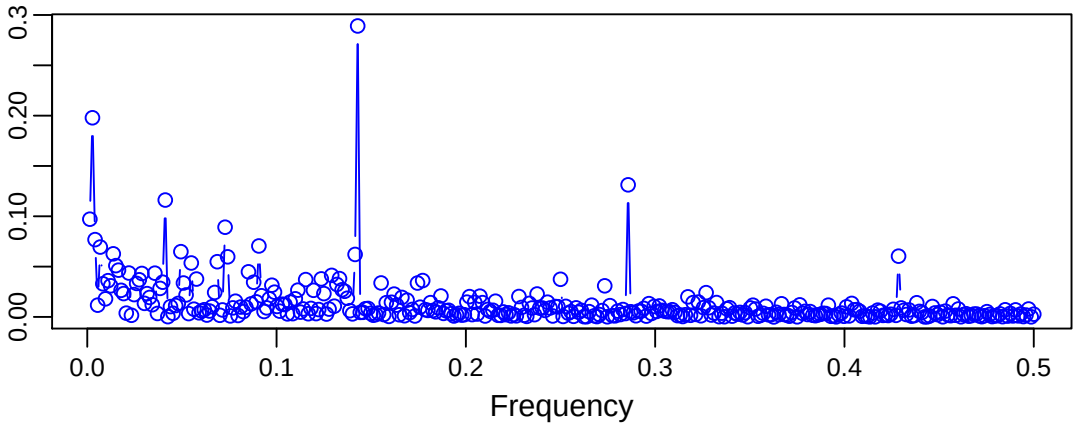


Figure A.5 – Periodogram of the $\log(\text{PM}_{10})$ time series.

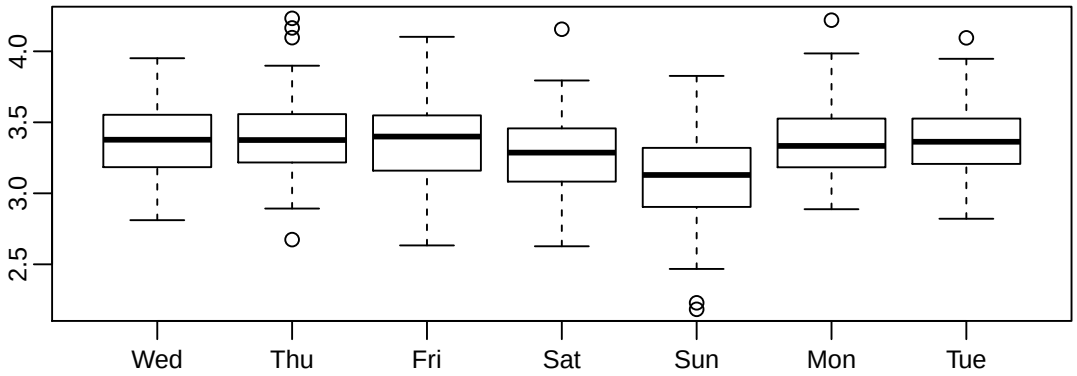


Figure A.6 – Daily box-plots of the $\log(\text{PM}_{10})$ time series.

procedure: (1) the linear model in (A.14) will be estimated through the ordinary least squares procedure; and (2) the PAR model in (A.15) will be fitted to the residuals of the linear model in step (1), where the AR orders $p_1, \ldots, p_s$ will be identified through the Schwartz Information Criterion (BIC) proposed by Schwarz (1978) and adapted to the periodic scenario by McLeod (1994).

At the first step, the linear model in Equation A.14 was fitted and the estimated coefficients are displayed in Table A.4. As expected, there were negative effects of Saturday and Sunday, which led to a decrease of $\log(\text{PM}_{10})$ levels during the weekends.

Table A.4 – Estimated coefficients of the linear model.

Parameter	μ	α_1	α_2	β_1	β_2
Estimate	3.2650	-0.0921	-0.2579	0.0640	0.0003

The BIC criterion was used to identify the order of the model (see Sarnaglia, Reisen & Lévy-Leduc (2010) for more details) and the results are displayed in Table A.5. In order to keep consistency with the simulation study, $c = 3.06$ was fixed in the Huber function (Equation A.12). Note that the PAR model with better (smaller) BIC was obtained by

the RYWE, which indicates that this estimator provides a good compromise between adjustment and parsimony.

Table A.5 – Selected AR orders using the BIC.

Estimator	BIC	p_1	p_2	p_3	p_4	p_5	p_6	p_7
YWE	−2156.25	1	1	1	1	1	1	1
RYWE	−2205.94	1	1	4	2	2	1	1
RLSE	−2200.15	1	1	4	1	2	1	1

The estimates of the PAR coefficients provided by YWE, RYWE and RLSE methods are given in Table A.6. Based on these results, it is clear the presence of Periodic Correlation in the data, since the AR coefficients and orders are not constant over the seasons. In general, the methods selected different orders and presented quite different coefficient estimates. This indicate that the high levels of the pollutant were stronger enough to provoke changes in the parameter estimates, that is, this reveals that the high levels of the pollutant PM_{10} presented the effects of additive outliers according to the discussion presented in the Simulation Section.

Table A.6 – Estimates of the AR coefficients for YWE, RYWE and RLSE.

Estimator	i	ν						
		1	2	3	4	5	6	7
$\hat{\phi}_i(\nu)$	1	0.626	0.532	0.485	0.374	0.595	0.312	0.498
$\tilde{\phi}_i(\nu)$	1	0.614	0.482	0.630	0.529	0.595	0.361	0.474
	2	0.000	0.000	-0.107	-0.143	-0.258	0.000	0.000
	3	0.000	0.000	0.451	0.000	0.000	0.000	0.000
	4	0.000	0.000	-0.374	0.000	0.000	0.000	0.000
$\breve{\phi}_i(\nu)$	1	0.661	0.513	0.479	0.376	0.638	0.293	0.522
	2	0.000	0.000	-0.018	0.000	-0.167	0.000	0.000
	3	0.000	0.000	0.271	0.000	0.000	0.000	0.000
	4	0.000	0.000	-0.240	0.000	0.000	0.000	0.000

The fitting performance will be accessed through the in-sample Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), symmetric MAPE (sMAPE) and Median of Absolute Deviation (MAD). The RMSE and the MAD are well-known and, for a discussion of MAPE and sMAPE quantities see Flores (1986). The results are presented in Table A.7. As can be seen, the RLSE are RYWE very competitive by presenting very similar results and they are slightly smaller than YWE method. This may corroborate the previous discussion related the effect of high level concentrations on the model estimation.

Figures A.7, A.8 and A.9 present the classic ACF of the residuals of each model. It can be seen that all the models were able to fully explain the correlation structure of the

Table A.7 – Fitting performance of the estimated models.

	Estimator		
Statistic	YWE	RYWE	RLSE
RMSE	0.2246	0.2245	0.2235
MAPE	5.3420	5.2545	5.2707
sMAPE	2.6454	2.6024	2.6104
MAD	0.2120	0.1988	0.2078

data, despite the eventual outliers effect. Based on the ACF of the residuals, the three estimation methods are comparable since all the estimated residuals look like a white noise process.

Finally, for all models, the residuals have not passed the Jarque-Bera normality test (JARQUE; BERA, 1980), presenting p-values < 0.05 which is an expected result due to the skewness revealed in the data.

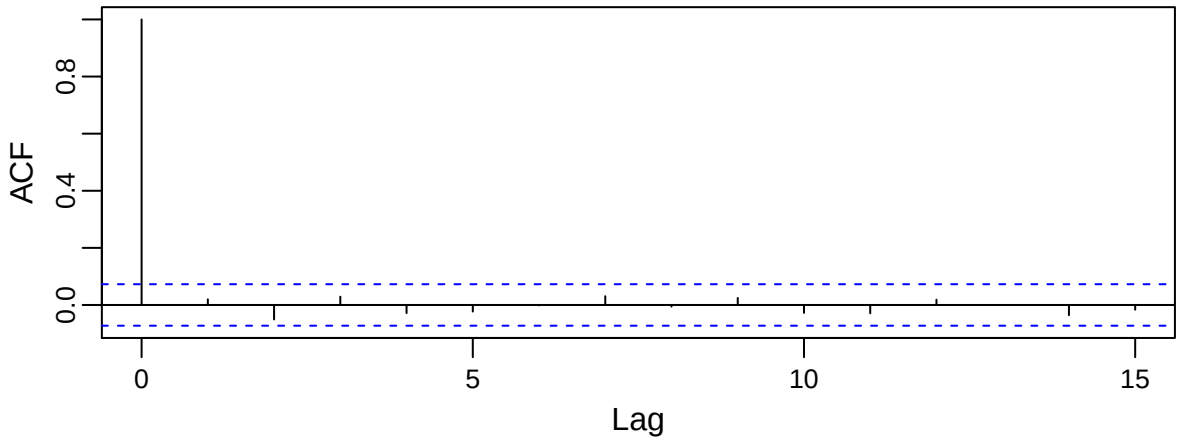


Figure A.7 – ACF of the residuals of the YWE fit.

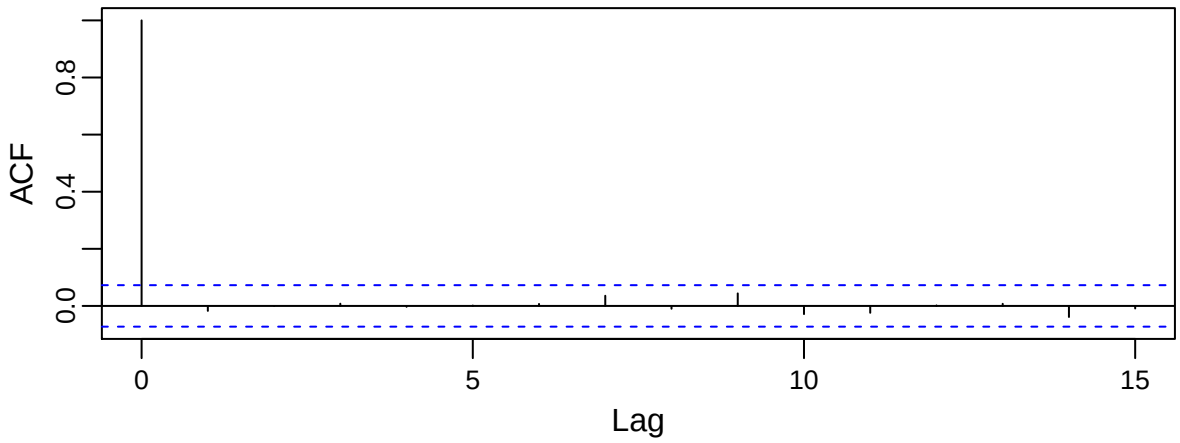


Figure A.8 – ACF of the residuals of the RYWE fit.

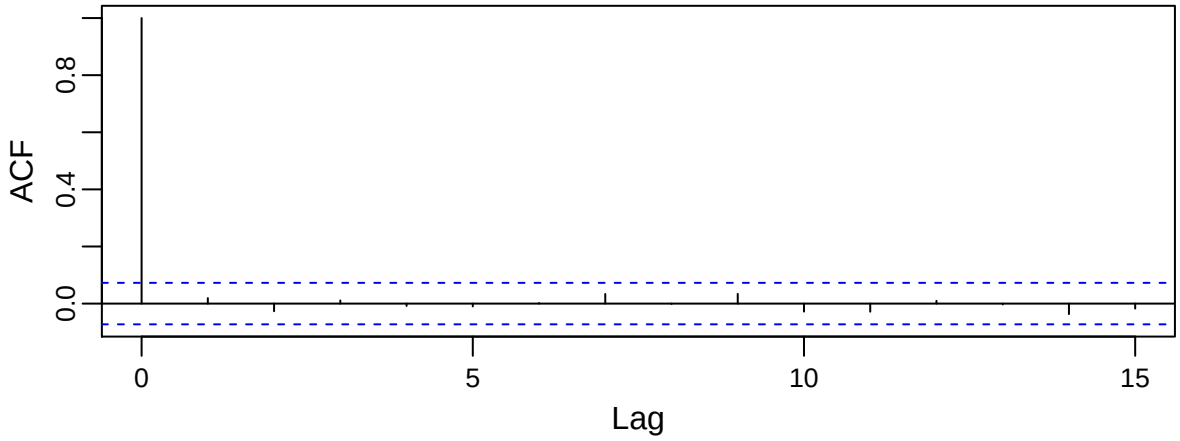


Figure A.9 – ACF of the residuals of the RLSE fit.

A.1.5 Conclusions

This paper reviews different estimation methodologies for PAR models. More specifically, the method considered are: the so-called YWE (MCLEOD, 1994), the RYWE (SARNAGLIA; REISEN; LÉVY-LEDUC, 2010) and the RLSE (SHAO, 2008). The finite sample performance of these methods was compared through a Monte Carlo experiment. The performance of RLSE is remarkably good under uncontaminated and contaminated scenarios, even under asymmetric errors, which violates Assumption 3. The RYWE is quite resistant to outliers, however it has a poor performance under asymmetric errors, mainly under weak correlation scenarios. As expected, YWE empirical distribution is resistant to departures from normality, however this estimator is completely affected by the presence of outliers. In order to illustrate the methodologies considered in this paper, the daily mean PM_{10} concentrations collected at the air quality monitoring station, located at Enseada do Suá, ES, Brazil, was considered as an application. The estimation and modelling results revealed outlier effects on the estimates.

A.2 Asymptotic properties of the M-regression spectral Whittle estimator for ARMA models.

Valdério Anselmo Reisen Céline Lévy-Leduc Carlo Corrêa Solci

Paper in Compilation

Abstract

A.2.1 Introduction

A.2.2 Statistical framework

Let $\{X_t\}_{t\in\mathbb{Z}} := \{X_t\}$ be a real valued stochastic process satisfying $\mathbb{E}(X_t^2) < \infty$ for all $t \in \mathbb{Z}$ with $\mu = \mathbb{E}(X_t)$ and

$$\gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)] \tag{A.16}$$

Suppose that the aucovariance function $\gamma_X(.)$ is absolutely summable, *i.e.*,

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty.$$

The spectral density of (X_t) is defined by

$$f(\lambda) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} \gamma_X(h)e^{-ih\lambda}, \text{ for all } \lambda \in [-\pi, \pi]. \tag{A.17}$$

The periodogram function is the standard spectral estimator of Equation A.17. For a given time series sample $X_1, \ldots, X_N$ from the process $\{X_t\}$, the classical periodogram is defined by

$$I_N(\lambda_j) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} X_k \exp(\mathrm{i}k\lambda_j)\right|^2 .$$

where $\lambda_j = 2\pi j/N$, $j = 1, .., [N/2]$, are the Fourier frequencies and $[x]$ denotes the integer part of x.

As is well-known, $I_N(\lambda_j)$ is related to the least-square estimate $\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)$ of a two-dimensional vector $\boldsymbol{\beta}$ in the linear regression model

$$X_i = c_{N,i}^T\boldsymbol{\beta} + \varepsilon_i \ , \ 1 \le i \le N, \ \boldsymbol{\beta} \in \mathbb{R}^2, \tag{A.18}$$

where

$$c_{N,i}^T(\lambda_j) = (\cos(i\lambda_j) \ \sin(i\lambda_j)) \, , \tag{A.19}$$

and where ε_i denotes the deviation of X_i from $c_{N,i}^T\boldsymbol{\beta}$ see, for example, Li (2008), Li (2010), Reisen, Lévy-Leduc & Taqqu (2017) .

Then

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\mathrm{Arg\,min}} \sum_{i=1}^{N} (X_i - c_{N,i}^T(\lambda_j)\boldsymbol{\beta})^2. \tag{A.20}$$

Indeed, the solution of (A.20) is

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (C^TC)^{-1}C^T\mathbf{X} = \frac{2}{N}C^T\mathbf{X} = \frac{2}{N}\left(\sum_{k=1}^{N} X_k\cos(k\lambda_j) \quad \sum_{k=1}^{N} X_k\sin(k\lambda_j)\right)^T, \tag{A.21}$$

where $\mathbf{X} = (X_1, \ldots, X_N)^T$ and where C and C^TC are defined by

$$C = \begin{pmatrix} \cos(\lambda_j) & \sin(\lambda_j) \\ \cos(2\lambda_j) & \sin(2\lambda_j) \\ \vdots & \vdots \\ \cos(N\lambda_j) & \sin(N\lambda_j) \end{pmatrix} \tag{A.22}$$

and

$$C^TC = \begin{pmatrix} \sum_{k=1}^{N}\cos(k\lambda_j)^2 & \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) \\ \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) & \sum_{k=1}^{N}\sin(k\lambda_j)^2 \end{pmatrix} = \frac{N}{2}\,\mathrm{Id}_2, \tag{A.23}$$

respectively, with $\mathrm{Id}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Hence

$$I_N(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)\|^2 =: I_N^{\mathrm{LS}}(\lambda_j), \tag{A.24}$$

where $\|\cdot\|$ denotes the classical Euclidian norm. It can thus be seen from (A.24) that there is a connection between the classical periodogram and the estimator of $\boldsymbol{\beta}$ in the linear regression model (A.18). Similar estimation strategy are the roots to obtain a M-periodogram spectral estimator. This is discussed as follows.

A M-periodogram based on an M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ of the regression coefficient $\boldsymbol{\beta}$ in (A.18) is built. The M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ is defined as the solution $\mathbf{t}$ of

$$\sum_{i=1}^{N} c_{N,i}\,\psi(X_i - c_{N,i}^T\mathbf{t}) = 0, \tag{A.25}$$

where, as in Wu (2007),

$$\psi(x) = \max(\min(x, c), -c),\ c > 0, \tag{A.26}$$

is the first derivative of the classical Huber's function. Following Fajardo et al. (2018), the M-periodogram is defined by

$$I_N^M(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}(\lambda_j)\|^2. \tag{A.27}$$

In the next section the Wittle estimator and its asymptotic proprieties are introduced. The method is based on the M-estimator for the spectral function defined in A.27.

A.2.3 The M-estimator for ARMA parameters and its asymptotic properties

Assume that $X_1, \ldots, X_n$ is a sample from a zero-mean stationary ARMA(p, q) process defined as

$$\sum_{j=0}^{p} \phi_j X_{t-j} = \sum_{k=0}^{q} \theta_k \, \eta_{t-k}, \tag{A.28}$$

where p and q are the AR and MA orders, respectively, $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the AR and MA coefficients, respectively, and $\phi_0 = \theta_0 = 1$. The sequence $\{\eta_t\}$ is zero-mean, Gaussian white noise process $\mathbb{E}(\eta_t^2) = \sigma^2$. As is well-known the ACFs of Model A.28 satisfies the Assumption A.17.

Let $\boldsymbol{\varphi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ be the parameter vector of model (A.28). We propose to estimate $\boldsymbol{\varphi}$ by minimizing:

$$\zeta_N(\boldsymbol{\varphi}) = \frac{1}{N} \sum_{j=1}^{N} \frac{I_N^M(\lambda_j)}{f(\lambda_j, \boldsymbol{\varphi})} \tag{A.29}$$

with respect to $\boldsymbol{\varphi}$, where

$$f(\lambda, \boldsymbol{\varphi}) = \frac{\left|\Theta(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2}{\left|\Phi(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2}, \tag{A.30}$$

and $\Phi(.)$ and $\Theta(.)$ are the following polynomials:

$$\Phi(z) = 1 + \phi_1 z + \cdots + \phi_p z^p \text{ and } \Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q. \tag{A.31}$$

Theorem 1. *Assume that $X_1, \ldots, X_n$ is a sample from a stationary ARMA(p, q) process defined in Equation (A.28) with parameter $\boldsymbol{\varphi}$ and such that the spectral density $f(.)$ defined in (A.30) satisfies for all $\lambda \in [-\pi, \pi]$, $f(\lambda, \varphi) \geq a$ for some positive a. Then, $\widehat{\boldsymbol{\varphi}}$ defined in (A.29) is a consistent estimator of $\boldsymbol{\varphi}$, as N tends to infinity, namely,*

$$\widehat{\boldsymbol{\varphi}} \xrightarrow{p} \boldsymbol{\varphi}, \text{ as } N \to \infty. \tag{A.32}$$

Proof of Theorem 1. The model (A.28) is causal and invertible. In order to prove (A.32), we shall use the same arguments as those given in Theorem 10.8.1 of Brockwell & Davis (1991). In this book, the propositions which cannot directly be used in the proof of Theorem 10.8.1 are Propositions 10.8.2 and 10.8.1 which have to be replaced by Propositions 2 and 5 given below, respectively. □

Proposition 2. *Let $X_1, \ldots, X_n$ be a sample from an ARMA(p, q) process satisfying the assumptions of Theorem 1. Let C be the set of parameters φ such that Φ and Θ have no common zeroes in the unit disk, then for every $\varphi \in C$,*

$$\frac{1}{N} \sum_{j=1}^{N} \frac{I_N^M(\lambda_j)}{g(\lambda_j, \varphi} \xrightarrow{p} \frac{1}{2\pi\kappa^2} \sum_{p \geq 1} \frac{c_p(\psi)^2}{p!} \int_{-\pi}^{\pi} \frac{g_(\lambda, \varphi^0)^{\star, p}}{g(\lambda, \varphi)} \mathrm{d}\lambda, \text{ as } N \to \infty, \tag{A.33}$$

where $\kappa = F(c) - F(-c)$, F being the c.d.f. of Z, c being defined in (A.26),

$$g(\lambda, \varphi) = \frac{\left|\Theta(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2}{\left|\Phi(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2},$$

with Θ and Φ defined in (A.31) and where $g^{\star,p}$ denotes the pth convolution product. Moreover, let us define the function g_δ for every positive δ as follows:

$$g_\delta(\lambda,\varphi)=\frac{\left|\Theta(\mathrm{e}^{-\mathrm{i}\lambda})+\delta\right|^2}{\left|\Phi(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2},$$

then

$$\frac{1}{N}\sum_{j=1}^{N}\frac{I_N^M(\lambda_j)}{g_\delta(\lambda_j,\varphi)}\xrightarrow{p}\frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\xi^p\int_{-\pi}^{\pi}\frac{g(\lambda,\varphi^0)^{\star,p}}{g_\delta(\lambda,\varphi)}\mathrm{d}\lambda,\ \text{as } N\to\infty, \tag{A.34}$$

uniformly in $\varphi\in\bar{C}$, where $\bar{C}$ denotes the closure of C and

$$\xi=\frac{\sigma^2}{2\pi\,\mathrm{Var}(X_1)}. \tag{A.35}$$

The proof of Proposition 2 is based on Lemmas 3 and 4 given below and the proofs of these are given in Section A.2.5.

Lemma 3. *Under the assumptions of Theorem 1 and if $\boldsymbol{\beta}=\mathbf{0}$ in (A.18). Then, $\hat{\boldsymbol{\beta}}_N^M$ defined by (A.25) satisfies*

$$\sqrt{\frac{N}{2}}(F(c)-F(-c))\hat{\boldsymbol{\beta}}_N(\lambda_j)=\sqrt{\frac{2}{N}}\sum_{r=1}^{N}\psi(X_r)\begin{pmatrix}\cos(r\lambda_j)\\ \sin(r\lambda_j)\end{pmatrix}+o_P(1),\ \text{as } N\to\infty,$$

where $F(.)$ is the c.d.f. of X_t, c is defined in (A.26) and $o_P(1)$ does not depend on j.

Lemma 4. *Under the assumptions of Theorem 1, let $I_N^M(\lambda_j)$ be defined in (A.27) for any fixed j. Then,*

$$I_N^M(\lambda_j)\xrightarrow{d}\frac{X^2+Y^2}{4\pi(F(c)-F(-c))^2},\quad \text{as } N\to\infty\ , \tag{A.36}$$

where F is the c.d.f. of X_t, c is defined in (A.26),

$$X\sim\mathcal{N}\left(0,\sum_{k\in\mathbb{Z}}\mathbb{E}\{\psi(X_1)\psi(X_{k+1})\}\cos(k\lambda_j)\right),\ Y\sim\mathcal{N}\left(0,\sum_{k\in\mathbb{Z}}\mathbb{E}\{\psi(X_1)\psi(X_{k+1}\}\cos(k\lambda_j)\right) \tag{A.37}$$

and

$$\mathrm{Cov}(X,Y)=\sum_{k\in\mathbb{Z}}\mathbb{E}\{\psi(X_1)\psi(X_{k+1})\}\sin(k\lambda_j). \tag{A.38}$$

Hence,

$$I_N^M(\lambda_j)=O_p(1),\ \text{as } N\to\infty,$$

where $O_p(1)$ does not depend on j.

Proposition 5. *Let C be the set of parameters $\boldsymbol{\varphi}$ such that $\Phi(.)$ and $\Theta(.)$ have no common zeroes in the unit disk and let $\boldsymbol{\varphi}$ be a fixed vector in C. Then, if there exists some positive constant a such that $f(\lambda, \varphi) \geq a$, for all $\lambda \in [-\pi, \pi]$,*

$$\frac{1}{2\pi\kappa^2} \sum_{p \geq 1} \frac{c_p(\psi)^2}{p!} \xi^p \int_{-\pi}^{\pi} \frac{f(\lambda, \boldsymbol{\varphi})^{\star, p}}{f(\lambda, \boldsymbol{\varphi})} \mathrm{d}\lambda$$

is minimized over C for $\boldsymbol{\varphi}$, where ξ is defined in (A.35).

The proof of Proposition 5 is given in Section A.2.5.

A.2.4 Simulation

Table A.8 – Exact Estimates for $\phi = 0.2$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.1891	0.0712
	M	0.1791	0.0735
4	C	0.1613	0.0808
	M	0.1702	0.0755
7	C	0.1301	0.1032
	M	0.1708	0.0759

Table A.9 – Exact Estimates for $\phi = 0.2$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.1941	0.0494
	M	0.1844	0.0518
4	C	0.1678	0.0600
	M	0.1775	0.0543
7	C	0.1329	0.0864
	M	0.1773	0.0549

Table A.10 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.4819	0.0654
	M	0.4618	0.0759
4	C	0.4312	0.1008
	M	0.4466	0.0856
7	C	0.3605	0.1692
	M	0.4457	0.0864

Table A.11 – Exact Estimates for $\phi = 0.5$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.4908	0.0444
	M	0.4713	0.0539
4	C	0.4393	0.0798
	M	0.4556	0.0645
7	C	0.3631	0.1531
	M	0.4553	0.0648

Table A.12 – Exact Estimates for $\phi = 0.8$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 200$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.7775	0.0509
	M	0.7599	0.0644
4	C	0.7320	0.0919
	M	0.7353	0.0861
7	C	0.6606	0.1687
	M	0.7315	0.0895

Table A.13 – Exact Estimates for $\phi = 0.8$ with $REP_{exac} = 10000$, $pr_{out} = 0.01$ and $N = 400$.

ω	I_N	$\mathsf{Mean}(\hat{\phi})$	$\mathrm{RMSE}(\hat{\phi})$
0	C	0.7886	0.0336
	M	0.7717	0.0454
4	C	0.7453	0.0689
	M	0.7492	0.0638
7	C	0.6698	0.1468
	M	0.7444	0.0683

A.2.5 Proofs

Proof of Proposition 2. We shall focus on the proof of (A.34) since the proof of (A.33) is similar. Following the same lines as those given in the proof of Proposition 10.8.2 in Brockwell & Davis (1991), let $q_m(\lambda, \varphi)$ be the Cesaro mean of the first m Fourier approximation to $g_\delta(\lambda, \varphi)^{-1}$, given by

$$q_m(\lambda, \varphi) = \sum_{|k|<m} \left(1 - \frac{|k|}{m}\right) b_k \mathrm{e}^{-\mathrm{i}k\lambda},$$

where $b_k = (2\pi)^{-1} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} g_\delta(\lambda, \varphi)^{-1} \mathrm{d}\lambda$. Then, for any positive ε, there exists an m such that

$$\left| q_m(\lambda, \varphi) - g_\delta(\lambda, \varphi)^{-1} \right| < \varepsilon,$$

for all $(\lambda, \varphi) \in [-\pi, \pi] \times \bar{C}$. Then, for all $\varphi \in \bar{C}$,

$$\left| \frac{1}{N} \sum_{j=1}^{N} \frac{I_N^M(\lambda_j)}{g_\delta(\lambda_j, \varphi)} - \frac{1}{N} \sum_{j=1}^{N} I_N^M(\lambda_j) q_m(\lambda_j, \varphi) \right| \leq \varepsilon N^{-1} \sum_{j=1}^{N} I_N^M(\lambda_j). \tag{A.39}$$

By Lemma 4, $I_N^M(\lambda_j) = O_p(1)$, which does not depend on j. Hence, the l.h.s of (A.39) tends to zero in probability as N tends to infinity.

Let us now prove that

$$\frac{1}{N} \sum_{j=1}^{N} I_N^M(\lambda_j) q_m(\lambda_j, \varphi) \xrightarrow{p} \frac{1}{2\pi\kappa^2} \sum_{p \geq 1} \frac{c_p(\psi)^2}{p!} \int_{-\pi}^{\pi} \frac{g(\lambda, \varphi^0)^{\star, p}}{g_\delta(\lambda, \varphi)} \mathrm{d}\lambda, \text{ as } N \to \infty.$$

Let $\hat{\boldsymbol{\beta}}_N(\lambda_j) = (\hat{\beta}_N^{(1)}(\lambda_j), \hat{\beta}_N^{(2)}(\lambda_j))'$. Then, by (A.27), we get that

$$\frac{1}{N} \sum_{j=1}^{N} I_N^M(\lambda_j) q_m(\lambda_j, \varphi) = \frac{1}{8\pi} \sum_{j=1}^{N} \left(\hat{\beta}_N^{(1)}(\lambda_j)^2 + \hat{\beta}_N^{(2)}(\lambda_j)^2 \right) \sum_{|k|<m} \left(1 - \frac{|k|}{m} \right) b_k \mathrm{e}^{-\mathrm{i}k\lambda_j}.$$

Then, by Lemma 3 and the proof of Lemma 4, we get that, as N tends to infinity,

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1}^{N} I_N^M(\lambda_j) q_m(\lambda_j, \varphi) \\
= \frac{1}{8\pi} \sum_{j=1}^{N} &\left(\frac{4}{N^2\kappa^2} \sum_{1 \leq r \neq r' \leq N} \psi(X_r)\psi(X_{r'}) \cos(r\lambda_j) \cos(r'\lambda_j) \right. \\
&\left. + \frac{4}{N^2\kappa^2} \sum_{1 \leq r \neq r' \leq N} \psi(X_r)\psi(X_{r'}) \sin(r\lambda_j) \sin(r'\lambda_j) \right) \sum_{|k|<m} d_k \mathrm{e}^{-\mathrm{i}k\lambda_j} + o_P(1) \\
&= \frac{1}{2\pi\kappa^2 N^2} \sum_{|k|<m} d_k \sum_{1 \leq r \neq r' \leq N} \psi(X_r)\psi(X_{r'}) \sum_{j=1}^{N} \cos\left((r - r')\lambda_j\right) \mathrm{e}^{-\mathrm{i}k\lambda_j} + o_P(1),
\end{aligned}$$

where $d_k = (1 - |k|/m)\, b_k$. Since

$$\sum_{j=1}^{N} \cos\left((r - r')\lambda_j\right) \mathrm{e}^{-\mathrm{i}k\lambda_j} = \begin{cases} \frac{N}{2}, \text{ if } r - r' = k \text{ or } -k \\ 0, \text{ else,} \end{cases}$$

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1}^{N} I_N^M(\lambda_j) q_m(\lambda_j, \varphi) \\
&= \frac{1}{2\pi\kappa^2} \sum_{|k|<m} d_k \left(\frac{1}{2N} \sum_{\substack{1 \leq r \neq r' \leq N \\ \text{s.t. } r - r' = k \text{ or } r - r' = -k}} \psi(X_r)\psi(X_{r'}) \right) + o_P(1) \\
&= \frac{1}{2\pi\kappa^2} \sum_{|k|<m} d_k \left\{ \frac{1}{2N} \sum_{r=1}^{N} \left(\psi(X_r)\psi(X_{r-k}) + \psi(X_r)\psi(X_{r+k}) \right) \right\} + o_P(1).
\end{aligned}$$

By the ergodic theorem, it is enough to study

$$\frac{1}{4\pi\kappa^2}\sum_{|k|<m} d_k \mathbb{E}\left(\psi(X_1)\psi(X_{1-k}) + \psi(X_1)\psi(X_{1+k})\right).$$

Using the Hermite expansion (A.40), we get

$$\mathbb{E}\left(\psi(X_1\psi(X_{1-k}) + \psi(X_1)\psi(X_{1+k})\right) = 2\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\rho(k)^p,$$

where $\psi = \psi(vx)$, $\rho(k) = \mathrm{Corr}(X_1, X_{(k+1)})$ and $v = \sqrt{\mathrm{Var}(X_1)}$. Hence,

$$\frac{1}{N}\sum_{j=1}^{N} I_N^M(\lambda_j)q_m(\lambda_j,\varphi) \xrightarrow{p} \frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\sum_{|k|<m}\left(1-\frac{|k|}{m}\right)b_k\rho(k)^p, \text{ as } N\to\infty,$$

uniformly in $\varphi \in \bar{C}$. By Lemma 6 given below,

$$\rho(k)^p = \left(\frac{\sigma^2}{2\pi\,\mathrm{Var}(X_1)}\right)^p \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} g^{\star,p}(\lambda,\varphi^0)\mathrm{d}\lambda = \xi^p \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} g^{\star,p}(\lambda,\varphi^0)\mathrm{d}\lambda.$$

Moreover,

$$\begin{aligned}
&\left|\frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\sum_{|k|<m}\left(1-\frac{|k|}{m}\right)b_k\rho(k)^p - \frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\xi^p\int_{-\pi}^{\pi}\frac{g(\lambda,\varphi^0)^{\star,p}}{g_\delta(\lambda,\varphi)}\mathrm{d}\lambda\right| \\
&= \left|\int_{-\pi}^{\pi}\frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\xi^p\left(q_m(\lambda,\varphi) - g_\delta(\lambda,\varphi)^{-1}\right)g(\lambda,\varphi^0)^{\star,p}\mathrm{d}\lambda\right| \\
&\leq \frac{\varepsilon}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}|\rho(0)^p| \leq \frac{c^2\varepsilon}{2\pi\kappa^2},
\end{aligned}$$

which concludes the proof. □

Lemma 6. *Let (Y_t) be a stationary process having an absolutely summable autocovariance function γ. Let f be its spectral density then for any positive integer m,*

$$\gamma(k)^m = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f^{\star,m}(\lambda)\mathrm{d}\lambda.$$

where $f^{\star,m}(\lambda)$ denotes the mth convolution product.

Proof of Lemma 6. By Theorem 4.3.2 of Brockwell & Davis (1991),

$$\gamma(k) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f(\lambda)\mathrm{d}\lambda.$$

Assume that,

$$\gamma(k)^{(m-1)} = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f^{\star,m-1}(\lambda)\mathrm{d}\lambda,$$

then

$$\begin{aligned}\gamma(k)^m &= \left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f^{\star,m-1}(\lambda)\mathrm{d}\lambda\right)\gamma(k)\\ &= \left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f^{\star,m-1}(\lambda)\mathrm{d}\lambda\right)\left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k} f(\nu)\mathrm{d}\nu\right).\end{aligned}$$

Observe that

$$f^{\star,m}(\lambda) = \int_{-\pi}^{\pi} f^{\star,m-1}(t) f(\lambda - t)\mathrm{d}t,$$

thus

$$\begin{aligned}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k\lambda} f^{\star,m}(\lambda)\mathrm{d}\lambda &= \int_{-\pi}^{\pi}\left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}k(\lambda-t)} f(\lambda-t)\mathrm{d}\lambda\right)\mathrm{e}^{\mathrm{i}kt} f^{\star,m-1}(t)\mathrm{d}t\\ &= \left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}ku} f(u)\mathrm{d}u\right)\left(\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}kt} f^{\star,m-1}(t)\mathrm{d}t\right),\end{aligned}$$

where the last equality comes from the fact that $\mathrm{e}^{\mathrm{i}kx} f(x)$ is 2π-periodic. This concludes the proof of the lemma. □

Proof of Lemma 3. By Propositions 1 and 4 and Example 1 of Wu (2007) the assumptions of Theorem 1 of Wu (2007) hold. Thus,

$$\sqrt{\frac{N}{2}}(F(c) - F(-c))\hat{\boldsymbol{\beta}}_N(\lambda_j) = \sqrt{\frac{2}{N}}\sum_{r=1}^{N}\psi(X_r)\begin{pmatrix}\cos(r\lambda_j)\\ \sin(r\lambda_j)\end{pmatrix} + o_P(1),\ \text{as } N\to\infty.$$

Moreover, by using Corollary 1 of Wu (2007), $o_p(1) = O_p(\log n/\sqrt{n})$ and does not depend on j, which gives the result. □

Proof of Lemma 4. By Propositions 1 and 4 and Example 1 of Wu (2007) the assumptions of Theorem 1 of Wu (2007) hold. Thus, $\hat{\boldsymbol{\beta}}_N(\lambda_j)$ defined by (A.25) satisfies the following convergence in distribution:

$$\sqrt{\frac{N}{2}}(F(c) - F(-c))\hat{\boldsymbol{\beta}}_N(\lambda_j) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}^{(j)}\right),\ N\to\infty,$$

with

$$\boldsymbol{\Delta}^{(j)} = \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(X_1)\psi(X_{k+1})\}\boldsymbol{\Delta}_k^{(j)},$$

where

$$\boldsymbol{\Delta}_k^{(j)} = \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\cos(\ell\lambda_j)\\ \sin(\ell\lambda_j)\end{pmatrix}\left(\cos((\ell+k)\lambda_j)\ \sin((\ell+k)\lambda_j)\right).$$

Observe that

$$\begin{aligned}\boldsymbol{\Delta}_k^{(j)} &= \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\frac{\cos(k\lambda_j)+\cos((2\ell+k)\lambda_j)}{2} & \frac{\sin(k\lambda_j)+\sin((2\ell+k)\lambda_j)}{2}\\ \frac{-\sin(k\lambda_j)+\sin((2\ell+k)\lambda_j)}{2} & \frac{\cos(k\lambda_j)-\cos((2\ell+k)\lambda_j)}{2}\end{pmatrix}\\ &= \begin{pmatrix}\cos(k\lambda_j) & \sin(k\lambda_j)\\ -\sin(k\lambda_j) & \cos(k\lambda_j)\end{pmatrix} + \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\frac{\cos((2\ell+k)\lambda_j)}{2} & \frac{\sin((2\ell+k)\lambda_j)}{2}\\ \frac{\sin((2\ell+k)\lambda_j)}{2} & \frac{-\cos((2\ell+k)\lambda_j)}{2}\end{pmatrix}.\end{aligned}$$

By observing that

$$\frac{1}{N}\sum_{\ell=1}^{N-|k|}\cos((2\ell+k)\lambda_j) = \frac{\cos(k\lambda_j)}{N}\sum_{\ell=1}^{N-|k|}\cos(2\ell\lambda_j) + \frac{\sin(k\lambda_j)}{N}\sum_{\ell=1}^{N-|k|}\sin(2\ell\lambda_j)$$
$$= \frac{\cos(k\lambda_j)}{N}\cos(\lambda_j(N-|k|-1))\frac{\sin(\lambda_j(N-|k|))}{\sin(\lambda_j)}$$
$$+ \frac{\sin(k\lambda_j)}{N}\sin(\lambda_j(N-|k|-1))\frac{\sin(\lambda_j(N-|k|))}{\sin(\lambda_j)}$$

tends to zero as N tends to infinity and that the same holds for $N^{-1}\sum_{\ell=1}^{N-|k|}\sin(2\ell+k)$, this shows that

$$\mathbf{\Delta}_k^{(j)} = \begin{pmatrix} \cos(k\lambda_j) & \sin(k\lambda_j) \\ -\sin(k\lambda_j) & \cos(k\lambda_j) \end{pmatrix},$$

which gives the convergence in distribution of $I_N^M(\lambda_j)$.

Observe that

$$\sum_{k\in\mathbb{Z}}\mathbb{E}\{\psi(X_1)\psi(X_{(k+1)})\}\cos(k\lambda_j) \leq \sum_{k\in\mathbb{Z}}\left|\mathbb{E}\left\{\psi\left(v\frac{X_1}{v}\right)\psi\left(v\frac{X_{(k+1)}}{v}\right)\right\}\right|,$$

where $v = \sqrt{\mathrm{Var}(X_1)}$. By expanding $\psi = \psi(vx)$ on the Hermite polynomials basis, we get that

$$\mathbb{E}\left\{\psi\left(v\frac{X}{v}\right)\psi\left(v\frac{X_{(k+1)}}{v}\right)\right\} = \sum_{p,q\geq 1}\frac{c_p(\psi)}{p!}\frac{c_q(\psi)}{q!}\mathbb{E}\{H_p(X_1/v)H_q(X_{(k+1)}/v)\}$$
$$= \sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\rho(k)^p \leq |\rho(k)|\|\psi\|_2^2 \leq c^2|\rho(k)|, \qquad \text{(A.40)}$$

where the last equality comes from Equation (2.1) of Breuer & Major (1983) and $\rho(k) = \mathrm{Corr}(X_1, X_{(k+1)})$. The result follows from the fact that $\rho(k)$ is absolutely sommable. □

Proof of Proposition 5. Observe that

$$\frac{1}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\xi^p\int_{-\pi}^{\pi}\frac{g(\lambda,\varphi^0)^{\star,p}}{g(\lambda,\varphi)}\mathrm{d}\lambda \geq \frac{(2\pi)^{p-1}a^{p-1}}{2\pi\kappa^2}\sum_{p\geq 1}\frac{c_p(\psi)^2}{p!}\xi^p\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{g(\lambda,\varphi^0)}{g(\lambda,\varphi)}\mathrm{d}\lambda\right).$$

Using similar arguments as those used in Proposition 10.8.1 of Brockwell & Davis (1991),

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{g(\lambda,\varphi^0)}{g(\lambda,\varphi)}\mathrm{d}\lambda > 1,$$

for all $\varphi \in \bar{C}$ such that $\varphi \neq \varphi^0$, which concludes the proof. □