

Universidade Federal de Mato Grosso
Instituto de Ciências Exatas e da Terra
Programa de Pós-Graduação em Recursos Hídricos
Planejamento e Gestão de Recursos Hídricos

Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão
Utilizando o Método de Análise de Espectro Singular.

Cesar Augusto Medeiros Destro

Cuiabá/MT
2010

Livros Grátis

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

Cesar Augusto Medeiros Destro

Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão
Utilizando o Método de Análise de Espectro Singular.

Dissertação apresentada ao programa de Pós-graduação
em Recursos Hídricos da Universidade Federal de Mato
Grosso, como requisito parcial para a obtenção do título
de Mestre em Recursos Hídricos.

Orientador: Prof. Dr. Gilson Alberto Rosa Lima
Coorientadora: Prof^a. Dr^a. Eliana Beatriz Nunes Rondon Lima

Cuiabá/MT
2010

D477s

Destro, Cesar Augusto Medeiros.

Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão Utilizando o Método de Análise de Cesar Augusto Medeiros. Cuiabá: UFMT, 2010.
99 fls.

Dissertação – Mestrado em Recursos Hídricos

Orientador: Prof. Dr. Gilson Alberto Rosa Lima

Co-Orientadora: Profª. Dra. Eliana Beatriz Nunes Rondon Lima

1.Análise de Séries Temporais. 2.Análise de Espectro Singular.
3.Caterpillar-SSA. I.Título.

CDU 556

Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão
Utilizando o Método de Análise de Espectro Singular.

Cesar Augusto Medeiros Destro

Dissertação apresentada ao programa de Pós-graduação em Recursos Hídricos da Universidade Federal de Mato Grosso, como requisito parcial para a obtenção do título de Mestre em Recursos Hídricos.

Aprovada por:

Gilson Alberto Rosa Lima
Doutor em Química
Departamento de Engenharia Sanitária e Ambiental/UFMT

Peter Zeilhofer
Doutor em Engenharia Florestal
Departamento de Geografia/UFMT

Alexandre Kepler Soares
Doutor em Hidráulica e Saneamento
Departamento de Engenharia Civil/UFG

Cuiabá/MT
2010

*Para Célia Lima de Sá Medeiros e
José Francisco Destro
pelo apoio e carinho dedicados.*

AGRADECIMENTOS

Ao professor Dr. G.A.R. Lima, pela orientação, sugestões e conselhos, os quais foram de fundamental importância para o desenvolvimento deste trabalho, bem como pela amizade que surgiu ao longo deste tempo.

Aos meus pais, José Francisco Destro e Célia Lima de Sá Medeiros, pelo estímulo e pela ajuda incondicional prestada durante todos estes anos.

Ao Prof. Dr. João Batista Lima e a Prof^ª. Dr^ª. Eliana Beatriz Nunes R. Lima, pela atenção, apoio e amizade.

A Acelmo de Jesus Brito, pela ajuda prestada na interpretação de alguns conceitos matemáticos e pelas discussões acerca do tema.

A Fernando Sanches Santos, pela elaboração das figuras com a localização das estações pluviométricas e com os pontos de medição de vazão e coleta de amostras.

A todos os professores do mestrado em Recursos Hídricos, especialmente: Dr. Ricardo Santos Silva Amorim, Dr. Alexandre Silveira e Dr^a. Eliana Freire G. de Carvalho Soares.

A todos os alunos da turma 2007/2 do mestrado em Recursos Hídricos, especialmente: Décio Eloit Siebert, Vanderlei Bellato, Sandra Márcia Laet, Suzy Mara Klemp, Soely Ovídio de Miranda, Juliana Kawanishi Braga e Osmar da Cruz Nascimento.

Aos amigos Luis Fernando Oliveira (Feijão), Gabriel dos Santos, Diego Coronel Bejarano (Paraguayo), Michael Lee (Chinês) e Jonas F. G. Batista (Sombra) com os quais tive o prazer de morar e conviver durante 5 meses em São Carlos/SP.

Ao estudante de doutorado Jorge Luis Rodrigues Pantoja Filho, pelas longas conversas que tornaram minha estadia na USP/São Carlos muito mais agradável e proveitosa. Pelo convite a Blumenau/SC, viagem inesquecível.

*“A ciência tem nos ensinado que,
por termos a habilidade de enganar
a nós mesmos, a subjetividade não
pode imperar livremente”.*

Carl Edward Sagan, Pálido Ponto Azul

RESUMO

DESTRO, C. A. M. (2010). Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão Utilizando o Método de Análise de Espectro Singular. 81p. Dissertação (Mestrado) – Programa de pós-graduação em Recursos Hídricos, Universidade Federal de Mato Grosso, Cuiabá, 2010.

Embora seja aceito que os programas de monitoramento coletam dados e informações relevantes para o processo de gestão, existe uma grande lacuna entre a informação necessária na gestão e a informação obtida nos atuais programas de monitoramento. Apesar de todo o esforço e investimento feito, a atual situação dos programas de monitoramento demonstra que a relação custo-benefício de tais programas ainda é muito pequena. A síndrome da riqueza de dados e pobreza de informação que afeta quase todos os programas de monitoramento em escala mundial levou pesquisadores dos cinco continentes a reavaliar as metodologias utilizadas no monitoramento. Neste sentido, o objetivo deste estudo foi aplicar o método de Análise de Espectro Singular (*Singular Spectrum Analysis* - SSA) para analisar séries temporais completas visando extrair informação sobre tendência e componentes sazonais, preencher falhas de dados de séries incompletas e prever de valores futuros.

As séries de dados utilizadas são provenientes dos sistemas de monitoramento existentes no Estado de Mato Grosso e foram obtidos do sistema de informações hidrometeorológicas (*hidroweb*) da Agência Nacional de Águas (ANA). As variáveis analisadas foram: vazão média mensal, precipitação acumulada mensal e turbidez. As séries de vazão e precipitação foram analisadas com o *software* Caterpillar[®]. As séries das variáveis vazão, precipitação e turbidez, foram preenchidas com o *software* CatMV[®]. Os resultados mostram que o método é robusto e, na análise, foi capaz de extrair tanto componentes de tendência linear quanto não linear, bem como os harmônicos de todas as séries. As séries reconstruídas durante a análise apresentaram Erro Médio Relativo (EMR) entre 0,4% e 0,9%. As séries reconstruídas durante preenchimento de falhas de dados apresentaram EMR entre 0,7% e 3,5%. O método foi capaz de prever valores futuros para um período de 24 meses mantendo a estrutura e o comportamento de uma série original de vazão (EMR igual a 0,45%) e de uma série de precipitação (EMR igual a 3,5%).

Na análise comparativa entre os métodos de preenchimento de falhas de dados SSA e Regressão Linear Simples (RLS), os EMRs foram bastante semelhantes. A principal vantagem do método SSA para o preenchimento de falhas de dados em relação aos métodos de regressão linear simples, regressão linear múltipla, método da ponderação regional e krigagem, deve-se a sua natureza auto-adaptativa que dispensa a dependência de séries temporais de estações vizinhas. O preenchimento é realizado durante o processo de reconstrução usando exclusivamente os próprios dados da série.

Este estudo mostrou que o método SSA, pode ser uma ferramenta útil na recuperação de informações perdidas e pode ser aplicado como ferramenta de auxílio na gestão de programas de monitoramento minimizando o custo do monitoramento e maximizando a qualidade da informação produzida a partir dos dados coletados, especialmente no monitoramento da qualidade da água onde os dados das variáveis são provenientes de campanhas de campo e posterior análise em laboratório e cujo custo é bastante elevado.

Palavras-chave: Análise de Séries Temporais, Análise de Espectro Singular, Caterpillar-SSA

ABSTRACT

DESTRO, C. A. M. (2010). Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão Utilizando o Método de Análise de Espectro Singular. 81p. Dissertação (Mestrado) – Programa de pós-graduação em Recursos Hídricos, Universidade Federal de Mato Grosso, Cuiabá, 2010.

Although it is generally accepted that the monitoring programs produce data and information relevant to the management process, there is a large gap between the information needed in management and information produced in current monitoring programs. Despite all efforts and investment made on monitoring program the current status of existing programs shows that the accruing benefits are low. Most of monitoring practices do not fulfill what is expected of monitoring. The rich-data but poor-information syndrome in water quality worldwide monitoring programs, led researchers to focus more critically on methodological procedures of data collection and analysis and its relation to monitoring management process. In this sense, the objective of this study was to apply the Singular Spectrum Analysis (SSA) method to analyze complete time series to extract information on trend and seasonal components and filling data gaps of incomplete time series and forecasting future values.

All the data set used in the study comes from Mato Grosso State monitoring systems and were obtained from the hydro-meteorological information system (Hidroweb) at National Water Agency (ANA). The analyzed variables were: average monthly stream flow, monthly precipitation and turbidity. The time series of stream flow and precipitation were analyzed using Caterpillar[®] software. The gaps of monthly stream flow, monthly precipitation and turbidity time series, were filled with CatMV[®] software. The results show that the methods are robust; the time series analysis was able to extract linear and non-linear trend components as well as the harmonics of all series. The reconstructed time series showed Mean Absolute Percentage Error (MAPE) between 0.4% and 9%. The time series reconstructed while filling data gaps showed MAPE between 0.7% and 3.5%. The method was able to predict future values for a period of 24 months, maintaining the pattern and behavior of the original series with MAPE between 0.2% and 3.5%.

The comparative analysis between SSA and Simple Linear Regression (SLR) methods show very similar MAPE. The main advantage of SSA on SLR, Multiple Linear Regression (MLR) weighting and regional kriging method is due to self-adaptive nature of SSA method releasing the dependency of time series neighboring stations. The filling gap is done during the reconstruction process using only their own data set.

The results showed that the SSA method can be a useful tool in the recovery of lost information and can be applied as a tool to aid the monitoring programs management by minimizing the cost of monitoring and maximizing the quality of information derived from the data collected. Especially in the water quality monitoring where all water quality variable data comes from field campaigns and subsequent laboratory analysis which has high cost.

Palavras-chave: Time Series Analysis, Singular Spectrum Analysis, Caterpillar-SSA

SUMÁRIO

LISTA DE FIGURAS	iv
LISTA DE TABELAS	viii
LISTA DE ABREVIATURAS, SÍMBOLOS E SIGLAS	ix
INTRODUÇÃO	1
1 REVISÃO DA LITERATURA	3
1.1 Séries Temporais e Processos Estocásticos	4
1.2 Análise de Séries Temporais	5
1.2.1 Significado dos Termos Sinal e Ruídos.....	6
1.2.2 Função de Autocorrelação.....	7
1.3 Análise Espectral	7
1.4 Fundamentos de Álgebra Linear	8
1.4.1 Matrizes e Vetores.....	9
1.4.2 Matriz Inversa e Transposta.....	9
1.4.3 Matriz Diagonal, Matriz Triangular e Matriz Ortogonal.....	10
1.4.4 Fatoração de Matrizes.....	11
1.4.5 Autovalores e Autovetores.....	11
1.4.6 Diagonalização de Matrizes.....	12
1.4.7 Decomposição Espectral.....	12
1.5 Análise de Espectro Singular (SSA)	13
1.5.1 Matriz Trajetória e Matriz de Covariância Defasada.....	15
1.5.2 Espectro Singular.....	15
2 METODOLOGIA	17
2.1 Aquisição e Pré-processamento dos Dados	18
2.1.1 Vazão Média Mensal.....	19
2.1.2 Precipitação Acumulada Mensal.....	20
2.1.3 Variáveis de Qualidade da Água.....	20
2.2 Séries Temporais Analisadas	21

2.3 Softwares utilizados nas análises e preenchimento	21
2.4 Cálculo do Periodograma	22
2.5 Estimativa dos Erros	23
2.6 Método da Análise de Espectro Singular (SSA)	23
2.6.1 Decomposição.....	24
2.6.1.1 <i>Inserção da série no espaço vetorial</i>	24
2.6.1.2 <i>Decomposição em Valores Singulares (SVD)</i>	25
2.6.2 Reconstrução.....	25
2.6.2.1 <i>Agrupamento</i>	25
2.6.2.2 <i>Diagonalização</i>	25
3 RESULTADOS E DISCUSSÃO	27
3.1 Análise das Séries Temporais	28
3.1.1 Vazão Média Mensal - Estação Cuiabá (66260001).....	28
3.1.1.1 <i>Decomposição da Série Original</i>	30
3.1.1.2 <i>Separação do Sinal e Ruído</i>	32
3.1.1.3 <i>Reconstrução do Sinal da Série</i>	36
3.1.2 Precipitação Acumulada Mensal - Estação Vila Alegre (1359001).....	39
3.1.2.1 <i>Decomposição da Série Original</i>	39
3.1.2.2 <i>Separação do Sinal e Ruído</i>	41
3.1.2.3 <i>Reconstrução do Sinal da Série</i>	43
3.2 Preenchimento de Falhas de Dados em Séries Temporais	46
3.2.1 Vazão Média Mensal - Estação Porto Cercado (6634000).....	47
3.2.1.1 <i>Decomposição da Série Original</i>	49
3.2.1.2 <i>Separação do Sinal e Ruído</i>	49
3.2.1.3 <i>Reconstrução do Sinal da Série</i>	51
3.2.2 Precipitação Acumulada Mensal - Estação Santa Rosa (857000).....	53
3.2.2.1 <i>Separação do Sinal e Ruído</i>	55
3.2.2.2 <i>Reconstrução do Sinal da Série</i>	56
3.2.2.3 <i>Comparação entre os Métodos SSA e Regressão Linear Simples (RLS)</i>	57
3.2.3 Séries Temporais de Variáveis de Qualidade da Água.....	60
3.2.3.1 <i>Turbidez - Estação Acorizal (66255000)</i>	63
3.2.3.1.1 <i>Decomposição da Série Original</i>	65
3.2.3.1.2 <i>Separação do Sinal e Ruído</i>	66

3.2.3.1.3 <i>Reconstrução do Sinal da Série</i>	67
3.2.3.1.4 <i>Reconstrução da Série Utilizando Autovetores de Sinal e Ruído</i>	69
3.3 Previsão de dados em variáveis ambientais	72
3.3.1 Vazão Média Mensal - Estação Cuiabá (66260001).....	72
3.3.2 Precipitação Acumulada Mensal - Estação Vila Alegre (1359001).....	73
4 CONCLUSÕES	75
5 REFERÊNCIAS BIBLIOGRÁFICAS	78

LISTA DE FIGURAS

Figura 2.1	Bacia hidrográfica do Rio Cuiabá e pontos de amostragem.....	18
Figura 2.2	Estações pluviométricas localizadas no Estado de Mato Grosso cujos dados foram utilizados neste estudo.....	19
Figura 3.1	Série temporal das vazões médias no rio Cuiabá, na estação de Cuiabá, durante o período de janeiro de 1961 a dezembro de 2007.....	29
Figura 3.2	Covariâncias médias dos vetores defasados.....	31
Figura 3.3	Periodograma médio da série temporal.....	31
Figura 3.4	Valores singulares dos autovetores defasados.....	32
Figura 3.5	Dez primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=276$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 6; (f) componente da tendência da série; (g) e (h) componente de baixa frequência, correlacionada com a tendência da série; (i) e (j) componente harmônico de período aproximadamente igual a 12.....	34
Figura 3.6	Matriz de correlação entre os autovetores reconstruídos.....	35
Figura 3.7	Reconstrução da tendência da série original. (a) utilizando apenas os vetores 7 e 8 e (b) utilizando apenas os vetores 1 e 6.....	36
Figura 3.8	Reconstrução do sinal e dos ruídos da séries. (a) tendência geral da série, reconstruída utilizando os autovetores 1, 6, 7 e 8; (b) componente harmônico reconstruído utilizando os autovetores 2-5 e 9-10 e; (c) reconstrução dos ruídos utilizando os autovetores 11-276.....	37
Figura 3.9	Reconstrução do sinal da série (tendência e harmônicos), considerando os 10 primeiros autovetores extraídos na etapa da decomposição.....	38
Figura 3.10	Erros absolutos (resíduos) cometidos na etapa de reconstrução do sinal da série, considerando os autovetores 1 ao 10.....	38
Figura 3.11	Série temporal de precipitação acumulada, registrada na estação de Vila Alegre, durante o período de janeiro de 1993 a dezembro de 2006.....	40
Figura 3.12	Covariâncias médias dos vetores defasados.....	41

Figura 3.13	Periodograma médio da série temporal.....	41
Figura 3.14	Valores singulares dos autovetores defasados.....	42
Figura 3.15	Três primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=84$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12.....	42
Figura 3.16	Matriz de correlação entre os autovetores reconstruídos.....	43
Figura 3.17	Reconstrução do sinal e dos ruídos da série. (a) tendência geral da série, reconstruída utilizando o autovetor 1; (b) componente harmônico reconstruído utilizando os autovetores 2-3 e; (c) reconstrução dos ruídos utilizando os autovetores 4-84.....	44
Figura 3.18	Reconstrução do sinal da série (tendência e harmônicos), considerando os 3 primeiros autovetores extraídos na etapa da decomposição.....	45
Figura 3.19	Erros absolutos (resíduos) cometidos na etapa de reconstrução do sinal da série, considerando os autovetores 1 ao 3.....	45
Figura 3.20	Série temporal das vazões médias obtidas no rio Cuiabá, na estação de Porto Cercado, durante o período de dezembro de 1969 a junho de 2005.....	48
Figura 3.21	Valores singulares dos autovetores defasados.....	49
Figura 3.22	Cinco primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=216$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 18.....	50
Figura 3.23	Séries reconstruídas através do método da projeção. (a) reconstrução da tendência da série utilizando o autovetor 1, sobreposta a série original; (b) reconstrução do componente harmônico da série com o agrupamento dos autovetores 2, 3, 4 e 5.....	51
Figura 3.24	Sinal reconstruído da série e resíduos da reconstrução. (a) série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5), sobreposta a série temporal original; (b) erros absolutos (resíduos) cometidos na etapa da reconstrução.....	52
Figura 3.25	Série temporal de precipitação acumulada mensal, registrada na estação Santa Rosa, durante o período de janeiro de 1993 a dezembro de 2006.....	54
Figura 3.26	Valores singulares dos autovetores defasados.....	55
Figura 3.27	Três primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=84$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12.....	56

Figura 3.28	Sinal reconstruído da série e resíduos da reconstrução. (a) série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2 e 3), sobreposta a série temporal original e; (b) erros absolutos (resíduos) cometidos na etapa da reconstrução.....	57
Figura 3.29	Regressões lineares (a) entre a precipitação registrada na estação Novo Planeta - eixo das abcissas - e a registrada na estação Santa Rosa - eixo das ordenadas; (b) entre a precipitação registrada na estação Trivelato - eixo das abcissas - e a registrada na estação Santa Rosa - eixo das ordenadas.....	59
Figura 3.30	Série temporal de turbidez no rio Cuiabá, obtida na estação de Acorizal, durante o período de janeiro de 1993 a março de 2006. A linha vermelha representa o ponto de truncamento da série.....	64
Figura 3.31	Série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5).....	65
Figura 3.32	Valores singulares dos autovetores defasados.....	66
Figura 3.33	Cinco primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=60$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 4.....	67
Figura 3.34	Séries reconstruídas através do método da projeção. (a) reconstrução da tendência da série utilizando o autovetor 1, sobreposta a série original; (b) reconstrução do componente harmônico da série com o agrupamento dos autovetores 2, 3, 4 e 5.....	68
Figura 3.35	Sinal reconstruído da série temporal a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5), sobreposta a série temporal original.....	68
Figura 3.36	Erros absolutos (resíduos) cometidos na etapa da reconstrução do sinal da série temporal de turbidez.....	69
Figura 3.37	Seis primeiros autovetores reconstruídos provenientes de ruídos na série temporal.....	70
Figura 3.38	Série reconstruída a partir do agrupamento dos trinta primeiros autovetores, sobreposta a série temporal original.....	71
Figura 3.39	Reconstrução da série temporal truncada e previsão dos valores retirados da série, sobreposta a série temporal original.....	72
Figura 3.40	Reconstrução da série temporal original e previsão para dois anos hidrológicos, sobreposta a série original.....	73
Figura 3.41	Reconstrução da série temporal truncada e previsão dos valores retirados da série, sobreposta a série temporal original.....	73

Figura 3.42	Reconstrução da série temporal original e previsão para dois anos hidrológicos, sobreposta a série original.....	74
--------------------	--	-----------

LISTA DE TABELAS

Tabela 2.1	Séries Temporais analisadas com o método SSA.....	21
Tabela 3.1	Estações pluviométricas mais próximas a estação Santa Rosa.....	57
Tabela 3.2	Coefficientes de correlação entre as séries temporais de precipitação acumulada mensal.....	58
Tabela 3.3	Erros Médios Absolutos e os Erros Médios Relativos do preenchimento das falhas de dados.....	59
Tabela 3.4	Características das séries temporais das variáveis de qualidade da água de algumas estações de amostragem da bacia do Rio Cuiabá.....	62

LISTA DE ABREVIATURAS, SÍMBOLOS E SIGLAS

ANA	Agência Nacional das Águas
EMA	Erro Médio Absoluto
EMR	Erro Médio Relativo
RLS	Regressão Linear Simples
SEMA-MT	Secretaria de Estado de Meio Ambiente de Mato Grosso
SSA	<i>Singular Spectrum Analysis</i> (Análise de Espectro Singular)
SVD	<i>Singular Value Decomposition</i> (Decomposição em Valores Singulares)
UNT	Unidade Nefelométrica de Turbidez
A	Matriz de dimensões $m \times n$
A^T	Transposta da matriz A
A⁻¹	Inversa da matriz A , também denotada por B
Λ	Matriz diagonal cujas entradas não negativas são iguais a λ
B	Inversa da matriz A , também denotada por A⁻¹
D	Matriz diagonal
D^T	Transposta da matriz D
E	Matriz cujos elementos são os autovetores x
E⁻¹	Inversa da matriz E
I	Matriz identidade
L	Matriz triangular inferior
L^T	Transposta da matriz L
Q	Matriz ortogonal
Q^T	Transposta da matriz Q
Q⁻¹	Inversa da matriz Q
ℝ^L	Espaço vetorial real de dimensão L
S	Matriz de covariância defasada
S^T	Transposta da matriz S
U	Matriz triangular superior
U'	Matriz U fatorada

\mathbf{x}	Autovetores de uma matriz qualquer
λ	Autovalores de uma matriz qualquer
\mathbf{X}	Matriz trajetória
\mathbf{X}^T	Transposta da matriz \mathbf{X}
F_t	Valor observado numa série temporal
K	Autovetores defasados resultantes da etapa de decomposição
N	Tamanho da série temporal
L	Tamanho da janela
Y_t	Valor estimado para um dado F_t numa série temporal
μ	Média dos vetores X_i
σ^2	Variância dos vetores X_i
ρ	Coefficiente de correlação linear
t	Valor do teste t de student
df	Graus de liberdade do resíduo
p -value	Valor da probabilidade

INTRODUÇÃO

Uma série temporal de dados contém informação a respeito do sistema físico, bioquímico, biológico ou socioeconômico que a produziu (GHIL *et al.*, 2002). A característica mais importante das séries temporais é que observações vizinhas são dependentes e, geralmente, se está interessado em analisar e modelar esta dependência. Como a maior parte dos procedimentos estatísticos foram desenvolvidos para analisar observações independentes, a análise de séries temporais requer o uso de técnicas específicas (EHLERS, 2009). Recentemente, vários métodos para análise uni e multivariada de séries temporais foram desenvolvidos e implementados com o objetivo de extrair informações das séries temporais (GHIL *et al.*, 2002).

A maior parte dos conjuntos de dados utilizados nas ciências ambientais são obtidos à partir de observações em campo, sob condições adversas, de maneira que as séries temporais apresentam falhas de dados, ou seja, períodos em que não foram realizadas observações, seja porque o valor não foi realmente medido ou por falhas nos equipamentos de medição (UYSAL, 2007). As falhas nos conjuntos de dados são prejudiciais na medida em que tem um impacto prático no tamanho da amostra disponível para análise (HAIR *et al.*, 2005). Consequentemente, também prejudicam a estatística descritiva, subestimando ou superestimando os parâmetros estatísticos e dificultando a extração de informação sobre efeitos sazonais da série temporal. Além disso, grande parte dos modelos para análise de séries temporais partem do pressuposto de que as observações que formam a série temporal foram realizadas com frequência constante, o que, geralmente, não ocorre em função das falhas de dados. Tendo em vista que as falhas geram várias dificuldades na análise das séries temporais, o preenchimento destas falhas é necessário para que toda a informação presente na série temporal possa ser extraída.

Segundo Hocke & Kämpfer (2009) existem vários métodos para contornar o problema das falhas de dados, tais como os métodos de interpolação linear e cúbica, triangulação, expansão harmônica, assimilação de dados a partir de observações em modelos gerais,

regressão linear, dentre outros. Muitos destes métodos necessitam do conhecimento prévio de uma função matemática que descreva o comportamento dos dados observados antes do preenchimento das falhas ou são dependentes da função de distribuição amostral dos dados. Como as variáveis ambientais possuem oscilações periódicas devido à dinâmica solar, atmosférica, oceânica e do solo, os métodos baseados em análise espectral, por serem métodos auto-adaptativos, não-paramétricos e que não necessitam do conhecimento prévio de uma função matemática, são considerados eficientes para o preenchimento de falhas de dados. Recentemente, a Análise de Espectro Singular (SSA) foi implementada em *softwares* visando a análise e o preenchimento de falhas de dados em séries temporais (SCHOELLHAMER, 2001; KONDRASHOV *et al.*, 2005; GOLYANDINA & OSIPOV, 2007; KONDRASHOV & GHIL, 2006). Baseados na SSA padrão, Ghil *et al.* (2002) desenvolveram o *software* kSpectra Toolkit[®], enquanto Golyandina *et al.* (2001) desenvolveram os *softwares* Caterpillar-SSA[®] e CatMV[®].

No presente estudo, foram utilizadas séries temporais de vazão média mensal, precipitação acumulada mensal e uma variável de qualidade da água – turbidez. As séries temporais destas variáveis permitem visualizar a variabilidade natural que ocorre sob influência do ciclo hidrológico. Porém, a maioria das séries temporais destas variáveis apresentam falhas de dados dificultando a extração da informação contida nestas séries além de dificultar a tomada de decisão quanto à racionalização do uso da água e à preservação dos recursos hídricos. Neste contexto, o presente estudo visa analisar, preencher falhas de dados e fazer previsão de dados destas séries temporais. Os dados utilizados neste estudo são provenientes do programa de monitoramento do Estado de Mato Grosso, cujos dados estão disponíveis no sistema de informações hidrometeorológicas da Agência Nacional das Águas (ANA), *Hidroweb*.

Capítulo 1

Revisão da Literatura

Este capítulo apresenta a revisão da literatura realizada tanto em artigos científicos publicados recentemente com a descrição ou com a aplicação do método de Análise de Espectro Singular (SSA), como em livros-texto sobre análise de séries temporais e processos estocásticos. Inicialmente são apresentados, de maneira geral, os objetivos da análise de séries temporais e alguns conceitos e definições importantes. Segue uma breve revisão da álgebra linear visando o entendimento de conceitos e definições utilizados no formalismo matemático do método SSA e, por fim, é apresentado o método implementado no algoritmo Caterpillar[®].

1.1 Séries Temporais e Processos Estocásticos

A maioria dos processos físicos envolve um elemento aleatório em sua estrutura, comumente denominado de processo estocástico. Os processos estocásticos podem ser definidos como fenômenos estatísticos que evoluem no tempo, de acordo com leis probabilísticas. Matematicamente, um processo estocástico pode ser definido como um conjunto de variáveis aleatórias ordenadas temporalmente (CHATFIELD, 2004). Similarmente, uma série temporal é definida como um conjunto de observações quantitativas dispostas em ordem cronológica (KIRCHGÄSSNER & WOLTERS, 2007; CHATFIELD, 2004) ou, em outras palavras, é um conjunto de observações, onde cada uma das observações está relacionada à um instante de tempo específico.

As séries temporais, ou os processos estocásticos, são ditas discretas quando as observações são feitas em intervalos de tempo fixos, ou contínuas quando as observações são registradas ininterruptamente durante um intervalo de tempo (BROCKWELL & DAVIS, 2002; CHATFIELD, 2004). Entretanto, é importante frisar que um processo, na verdade, não pode ser medido efetivamente de forma contínua, uma vez que isso implicaria num número infinito de observações em um intervalo de tempo bastante curto. A distinção entre as séries discretas e as séries contínuas é feita essencialmente para se determinar que tipo de modelo será aplicado. Os modelos de tempo contínuo podem ser aplicados para todos os dados, mas são mais complexos e podem não ser apropriados caso mudanças estruturais ocorram em momentos distintos da série, enquanto os modelos de tempo discreto requerem que as observações sejam realizadas em intervalo de tempo constante, a menos que possa ser utilizado um mecanismo para o preenchimento das falhas de dados (LINDSEY, 2004).

As séries temporais discretas são também chamadas de funções amostrais, uma vez que se originam a partir de amostragens realizadas em intervalos de tempos fixos e sucessivos. Séries mensais de precipitação constituem um tipo diferente de série temporal discreta, uma vez que a variável não possui uma medida instantânea, mas sim valores acumulados em intervalos de tempos iguais (CHATFIELD, 2004). De maneira geral, dados obtidos à partir de observações sequenciais no tempo são extremamente comuns em muitas áreas. A análise de tais séries ajuda a determinar algumas propriedades do sistema, através da quantificação de determinadas características que ela porventura possua. Estas propriedades podem ajudar no entendimento e previsão do comportamento futuro do sistema (GHIL *et al.*,

2002; KIRCHGÄSSNER & WOLTERS, 2007; CRYER & CHAN, 2008). As principais características presentes em uma série temporal são a tendência e as variações sazonais que podem ser modeladas deterministicamente com funções matemáticas temporais. Outra característica importante da maior parte das séries temporais é que observações próximas no plano temporal, tendem a ser correlacionadas. Grande parte da metodologia de análise de séries temporais visa explicar essa correlação e as características principais dos dados utilizando modelos estatísticos apropriados e métodos descritivos (COWPERTWAIT & METCALFE, 2009).

1.2 Análise de Séries Temporais

Muitos problemas estatísticos estão voltados a estimar propriedades de uma população a partir de uma amostra. Na análise de séries temporais, entretanto, há uma situação bastante diferente uma vez que a ordem das observações é determinada pelo tempo e geralmente é impossível fazer mais do que uma observação em um determinado momento (CHATFIELD, 2004). De maneira geral, pode-se dizer que os objetivos da análise de uma série temporal são compreender o mecanismo ou modelo estocástico que dá origem a série observada, prever ou prever os valores futuros de uma série baseado nos dados históricos desta série e, possivelmente, de outras séries ou fatores relacionados (CRYER & CHAN, 2008), descrever o comportamento da série e procurar por periodicidades relevantes (MORETTIN & TOLOI, 2006).

A análise de séries temporais passa, inicialmente, pela representação gráfica dos dados dispostos sequencialmente ao longo do tempo. Estas representações podem revelar padrões de comportamento importantes, tais como as tendências de crescimento/decrescimento, padrões cíclicos, alterações estruturais, observações discrepantes, dentre outras. Uma suposição bastante frequente que se faz a respeito de uma série temporal é que ela é estacionária, ou seja, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável (MORETTIN & TOLOI, 2006). Ainda segundo Morettin & ToloI (2006), a maior parte das séries que se encontra na prática apresentam alguma forma de não-estacionariedade, ou seja, de maneira geral as séries apresentam tendências, sendo o caso mais simples aquele em que a série flutua ao redor de uma reta com inclinação positiva ou negativa.

De maneira geral, a ideia básica por trás da análise de séries temporais é que é possível

decompor a série em um número finito de componentes independentes, mas não diretamente observáveis, que se propagam regularmente e podem, assim, ser calculados antecipadamente (KIRCHGÄSSNER & WOLTERS, 2007). Os métodos tradicionais de análise de séries temporais preocupam-se principalmente com a decomposição da série em componentes que representam a tendência e a variação sazonal ou outras mudanças cíclicas. Demais variações são atribuídas a flutuações irregulares normalmente denominadas ruídos.

1.2.1 Significado dos Termos Sinal e Ruídos

A tendência de uma série temporal pode ser entendida como uma mudança de longo prazo no nível médio da série, e independe se a mudança é no sentido de aumento ou diminuição dos valores médios. Além disso, o modelo explicativo da tendência pode ser linear ou não linear. A tendência pode ser representada por um polinômio determinístico (SAHU, 2003) e, caso a tendência seja monotonicamente crescente ou decrescente, uma função linear ou quadrática pode ser apropriada para representá-la. Caso contrário, pode-se ajustar polinômios de ordem mais alta. O problema mais sério que se encontra ao estimar a tendência através de um polinômio é que, embora ele possa ajustar-se bem ao conjunto de valores observados, extrapolações futuras podem não ser bem ajustadas (MORETTIN & TOLOI, 2006).

A sazonalidade de uma série pode ser entendida por comportamentos que tendem a se repetir a cada período de tempo, ou seja, considera-se sazonal fenômenos que ocorrem regularmente. Uma importante característica das séries sazonais é que ocorrem relações entre observações para meses sucessivos de um ano particular, além de ocorrerem relações entre as observações para o mesmo mês em anos sucessivos (MORETTIN & TOLOI, 2006).

A tendência e a sazonalidade são normalmente denominados de “sinal” da série temporal, e são resultados da dinâmica do sistema. O sinal pode ser resumido como o resultado de um conjunto de processos físicos e interações entre os diferentes componentes do sistema climático que atuam nas mais variadas escalas temporais e espaciais. O ruído pode ter um significado físico ou é, então, gerado pelos instrumentos de medição e compreende todos os outros aspectos considerados irrelevantes para o sinal.

Existem duas maneiras distintas de análise de séries temporais. Por um lado tem-se os métodos de domínio temporal, que têm sua origem na teoria clássica da correlação, e levam

inevitavelmente a construção de modelos paramétricos (MORETTIN & TOLOI, 2006). Muitos dos métodos utilizados para estimar os parâmetros destes modelos podem ser vistos como variantes sofisticadas dos métodos de regressão linear (POLLOCK, 1999). Por outro lado, tem-se os métodos de domínio das frequências, ou de análise espectral, que são uma extensão dos métodos de análise de Fourier, que consideram que ao longo de um intervalo de tempo finito, qualquer função analítica pode ser aproximada, em qualquer grau de precisão desejado, através de funções seno e cosseno (POLLOCK, 1999). A análise espectral fornece uma maneira alternativa para a visualização de um processo, através da decomposição da série em uma soma de componentes periódicos não correlacionados e de frequências diferentes (FAN & YAO, 2003).

1.2.2 Função de Autocorrelação

Uma característica importante das séries temporais é que as observações adjacentes são, geralmente, mais estritamente relacionadas do que aquelas mais separadas no tempo. Uma ferramenta bastante útil para descobrir dependências em dados de séries temporais é a função de autocorrelação, que contém informação sobre a correlação entre as observações da série em vários intervalos de tempo distintos, denominados defasagens (LINDSEY, 2004).

Considerando um vetor X_i , com média μ , sua autocorrelação $R(h)$ é definida como:

$$R(h) = \frac{(X_i - \mu)(X_{i+h} - \mu)}{\sigma^2} \quad (1.1)$$

em que h é o deslocamento no tempo (defasagem) e σ^2 é a variância do vetor X_i .

O comportamento periódico das funções de autocorrelação caracteriza uma série temporal como periódica, sendo possível determinar qual o principal componente sazonal da série através desta função. A autocovariância pode ser calculada pela expressão 1.1, retirando-se o termo referente a variância do vetor X_i . A autocovariância descreve o quanto o vetor X_i varia em conjunto com um vetor X_{i+h} .

1.3 Análise Espectral

Qualquer série temporal que tem um padrão de repetição é considerada periódica. O sinal periódico fundamental é a onda senoidal e, como demonstrou Joseph Fourier, a soma de ondas senoidais é uma boa aproximação para descrever a maioria dos sinais periódicos

(LINDSEY, 2004). No seu sentido mais restrito, a análise de Fourier, ou análise harmônica de uma série temporal, corresponde a decomposição da série em uma soma de componentes senoidais. No entanto, o termo pode ser usado num sentido mais amplo para descrever qualquer processo de análise de dados que descreva as medidas das flutuações em uma série histórica, comparando-os com senoides (BLOOMFIELD, 2000).

Uma importante técnica de análise de dados no domínio das frequências é o denominado periodograma, que corresponde a transformada de Fourier das funções de autocovariância. O periodograma foi uma das primeiras ferramentas estatísticas para o estudo de variações periódicas em séries temporais e, antes do seu desenvolvimento, a análise periódica era bastante difícil e geralmente viável apenas quando o período de interesse abrangia toda a série temporal. A obtenção do periodograma é descrita no item 2.4.

1.4 Fundamentos de Álgebra Linear

A álgebra linear diz respeito a simplificação de matrizes. Resumidamente, existem duas maneiras de simplificar uma matriz. Elas podem ser simplificadas através de um método chamado de eliminação, ou podem ser simplificadas através de um processo chamado de decomposição espectral. A simplificação através da eliminação produz matrizes diagonais e triangulares, e é utilizada extensivamente na solução de sistemas lineares de forma

$$\mathbf{Ax} = \mathbf{b} \quad (1.2)$$

em que \mathbf{A} é a matriz de coeficientes, \mathbf{x} é a uma matriz coluna que contém as incógnitas e \mathbf{b} é uma matriz coluna que contém as constantes do sistema.

Enquanto a simplificação através da eliminação preserva o conjunto de todas as equações lineares possíveis nos vetores linha da matriz, a simplificação por decomposição espectral preserva os autovetores. A decomposição espectral é consideravelmente útil no estudo da dinâmica dos sistemas e é de fundamental importância para o entendimento da SSA. A maior parte deste item foi compilado de Elsner & Tsonis (1996), demais bibliografias consultadas são citadas no decorrer do texto.

1.4.1 Matrizes e Vetores

Uma matriz \mathbf{A} de dimensões $m \times n$ é uma tabela de mn elementos, dispostos em m linhas e n colunas:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad (1.3)$$

Pode-se utilizar também a notação $\mathbf{A} = (a_{ij})_{m \times n}$ e diz-se que a_{ij} é o elemento de posição i, j da matriz \mathbf{A} . Se $m = n$ então a matriz \mathbf{A} é uma matriz quadrada de ordem n e os elementos $a_{11}, a_{22}, \dots, a_{nn}$ formam a diagonal principal de \mathbf{A} . Uma matriz com apenas uma linha é chamada de matriz linha ou vetor linha, e uma matriz com apenas uma coluna é chamada de matriz coluna ou vetor coluna (LIPSCHUTZ & LIPSON, 2001).

A matriz $n \times n$ cujos elementos na diagonal principal são todos iguais a 1 enquanto todos os outros elementos são iguais a 0, é chamada de matriz identidade \mathbf{I} . A matriz identidade \mathbf{I} é semelhante ao escalar 1 pois, para qualquer matriz \mathbf{A} quadrada de ordem n ,

$$\mathbf{A} \mathbf{I} = \mathbf{I} \mathbf{A} = \mathbf{A} \quad (1.4)$$

a matriz identidade \mathbf{I} pode ser escrita como (LIPSCHUTZ & LIPSON, 2001):

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad (1.5)$$

1.4.2 Matriz Inversa e Transposta

Uma matriz \mathbf{A} é dita inversível (não singular) se existe uma matriz \mathbf{B} de maneira que

$$\mathbf{A} \mathbf{B} = \mathbf{B} \mathbf{A} = \mathbf{I} \quad (1.6)$$

onde \mathbf{I} é a matriz identidade. Denotando a matriz inversa de \mathbf{A} por \mathbf{A}^{-1} , então

$$\mathbf{A} \mathbf{A}^{-1} = \mathbf{A}^{-1} \mathbf{A} = \mathbf{I} \quad (1.7)$$

Existem vários métodos diferentes para o cálculo da matriz \mathbf{A}^{-1} . Na prática, frequentemente a maneira mais direta de se obter \mathbf{A}^{-1} é utilizando métodos tais como o de Crout's, uma vez que o número de operações necessárias são minimizadas.

A matriz transposta de \mathbf{A} , denotada por \mathbf{A}^T , pode ser obtida diretamente através da troca das colunas pelas linhas da matriz, ou seja, a i -ésima linha de \mathbf{A} é a i -ésima coluna de \mathbf{A}^T . Assim, se \mathbf{A} possui dimensão n por m , então \mathbf{A}^T possui dimensão m por n . Se a matriz \mathbf{A} for igual a sua transposta, ou seja, se $\mathbf{A} = \mathbf{A}^T$, então \mathbf{A} é dita matriz simétrica. Matrizes simétricas surgem naturalmente de várias maneiras. Especificamente no caso da SSA, a matriz simétrica surge uma vez que, estatisticamente, a correlação entre os eventos i e j é igual a correlação entre os eventos j e i , de maneira que se os valores contidos na matriz \mathbf{A} forem as correlações existentes entre os eventos, então $a_{ij} = a_{ji}$ para todo i e j , então \mathbf{A} é simétrica.

1.4.3 Matriz Diagonal, Matriz Triangular e Matriz Ortogonal

Uma matriz quadrada $\mathbf{D} = [d_{ij}]$ é dita diagonal se todos os elementos fora da sua diagonal principal são nulos. Tal matriz é, as vezes, denotada por (LIPSCHUTZ & LIPSON, 2001):

$$\mathbf{D} = \text{diag}(d_{11}, d_{22}, \dots, d_{nn}) \quad (1.8)$$

Uma matriz quadrada $\mathbf{A} = [a_{ij}]$ é triangular se todas as entradas abaixo ou acima da diagonal principal são iguais a 0. Uma matriz é dita triangular superior, denotada por \mathbf{U} , se $a_{ij} = 0$ para todo $i > j$. Uma matriz é dita triangular inferior, denotada por \mathbf{L} , se $a_{ij} = 0$ para todo $i < j$.

Uma matriz real é ortogonal se $\mathbf{A}^T = \mathbf{A}^{-1}$, isto é, se $\mathbf{A} \mathbf{A}^T = \mathbf{A}^T \mathbf{A} = \mathbf{I}$. Deste modo, \mathbf{A} deve ser necessariamente quadrada e inversível. Se \mathbf{A} é uma matriz ortogonal 3×3 com linhas

$$u_1 = (a_1, a_2, a_3), \quad u_2 = (b_1, b_2, b_3), \quad u_3 = (c_1, c_2, c_3) \quad (1.9)$$

então

$$\mathbf{A} \mathbf{A}^T = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I} \quad (1.10)$$

Multiplicando \mathbf{A} por \mathbf{A}^T e igualando as entradas correspondentes, se obtém nove equações:

$$\begin{aligned} a_1^2 + a_2^2 + a_3^2 &= 1, & a_1 b_1 + a_2 b_2 + a_3 b_3 &= 0, & a_1 c_1 + a_2 c_2 + a_3 c_3 &= 0 \\ b_1 a_1 + b_2 a_2 + b_3 a_3 &= 0, & b_1^2 + b_2^2 + b_3^2 &= 1, & b_1 c_1 + b_2 c_2 + b_3 c_3 &= 0 \\ c_1 a_1 + c_2 a_2 + c_3 a_3 &= 0, & c_1 b_1 + c_2 b_2 + c_3 b_3 &= 0, & c_1^2 + c_2^2 + c_3^2 &= 1 \end{aligned} \quad (1.11)$$

ou seja $u_1 \times u_1 = 1$, $u_2 \times u_2 = 1$, $u_3 \times u_3 = 1$ e $u_i \times u_j = 0$, se $i \neq j$. Assim, diz-se que as linhas u_1 , u_2 e u_3 são vetores unitários e ortogonais. Generalizando, os vetores u_1, u_2, \dots, u_m de \mathbb{R}^n formam um conjunto ortonormal de vetores se cada um desses vetores são unitários e ortogonais aos demais, isto é

$$u_i \times u_j = \begin{cases} 0 & \text{se } i \neq j \\ 1 & \text{se } i = j \end{cases} \quad (1.12)$$

Em outras palavras, $u_i \times u_j = \delta_{ij}$ onde δ_{ij} é a função delta de Kronecker.

1.4.4 Fatoração de Matrizes

Muitas vezes pode ser útil simplificar matrizes, ou seja, decompor uma matriz em um produto de matrizes mais simples. Este processo é chamado de fatoração, a qual pode ser alcançada aplicando o método chamado de eliminação Gaussiana. A eliminação Gaussiana fatora a matriz \mathbf{A} em um produto $\mathbf{L} \times \mathbf{U}$, onde \mathbf{L} é uma matriz triangular inferior e \mathbf{U} uma matriz triangular superior. Qualquer matriz \mathbf{A} cuja diagonal principal é constituída por valores diferentes de zero possui uma única fatoração $\mathbf{L} \times \mathbf{U}$ e, portanto, pode-se substituir \mathbf{A} pela fatoração. Entretanto, é mais conveniente se a matriz \mathbf{U} , assim como seu complemento \mathbf{L} , possuir valores iguais a 1 na sua diagonal principal. Isto pode ser obtido através da fatoração da matriz \mathbf{U} até a obtenção de uma matriz diagonal \mathbf{D} . Assim, uma matriz \mathbf{A} pode ser escrita como $\mathbf{A} = \mathbf{L} \mathbf{D} \mathbf{U}'$.

É importante notar que, se \mathbf{A} é simétrica, e se ela sofrer fatoração, então a matriz triangular superior \mathbf{U} é igual à transposta da matriz triangular inferior \mathbf{L} . Assim, pode-se escrever

$$\mathbf{A} = \mathbf{L} \mathbf{D} \mathbf{L}^T \quad (1.13)$$

1.4.5 Autovalores e Autovetores

Para a SSA, se está interessado em uma família de matrizes $\mathbf{A} - \lambda \mathbf{I}$. Em particular se

deseja conhecer os valores de λ , tal que $\mathbf{A} - \lambda \mathbf{I}$ seja singular¹. Se a determinante da matriz \mathbf{A} (denotada por $\det \mathbf{A}$ ou $|\mathbf{A}|$) é igual a zero, então \mathbf{A} é singular, e se $\det \mathbf{A} \neq 0$, então \mathbf{A} é inversível. Os valores de λ que tornam $\mathbf{A} - \lambda \mathbf{I}$ singular são chamados de autovalores. Para cada um dos autovalores encontrados há um conjunto de vetores que satisfazem a equação

$$(\mathbf{A} - \lambda \mathbf{I}) \mathbf{x} = 0, \quad (1.14)$$

ou

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad (1.15)$$

onde \mathbf{x} são vetores conhecidos como autovetores de \mathbf{A} . Geometricamente, os autovetores podem ser considerados eixos de um novo sistema de coordenadas. Então, os autovetores são vetores que, quando multiplicados por \mathbf{A} , não têm sua direção alterada, apenas a sua escala. É importante ressaltar que, em função da grande quantidade de operações necessárias para calcular a determinante de grandes matrizes, na prática, não se utiliza o conceito de determinantes para o cálculo dos autovalores e dos autovetores.

1.4.6 Diagonalização de Matrizes

Se \mathbf{A} é uma matriz de $n \times n$ dimensões e que possui n autovetores independentes linearmente \mathbf{e}_i , $i = 1, \dots, n$, pode-se construir uma matriz \mathbf{E} na qual cada coluna é um autovetor de \mathbf{A} , então

$$\mathbf{E}^{-1} \mathbf{A} \mathbf{E} = \mathbf{\Lambda}, \quad (1.16)$$

onde $\mathbf{\Lambda}$ é uma matriz diagonal onde as entradas não negativas correspondem aos autovalores de \mathbf{A} . O produto $\mathbf{E}^{-1} \mathbf{A} \mathbf{E}$ é chamado de forma diagonal da matriz \mathbf{A} .

1.4.7 Decomposição Espectral

Se \mathbf{A} é uma matriz real e simétrica, então cada um dos autovalores de \mathbf{A} também são reais e se todos os autovalores são distintos, então os autovetores correspondentes são ortogonais. Ou seja, se \mathbf{x}_i são os autovetores, então $\mathbf{x}_i^T \mathbf{x}_j = 0$ para todo $i \neq j$. A matriz \mathbf{A} pode ser diagonalizada por uma matriz ortogonal \mathbf{Q} cujas colunas são autovetores ortonormais² de \mathbf{A} . Então, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ e, além disso, $\mathbf{Q}^T = \mathbf{Q}^{-1}$. A matriz \mathbf{Q} possui colunas ortonormais com

$$\mathbf{Q}^{-1} \mathbf{A} \mathbf{Q} = \mathbf{\Lambda}, \quad (1.17)$$

então

1 Uma matriz é dita singular quando não admite uma inversa.

2 Dois autovetores são ditos ortonormais se ambos são unitários e ortogonais entre si.

$$\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T, \quad (1.18)$$

ou

$$\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T + \dots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T \quad (1.19)$$

que é a chamada decomposição espectral da matriz \mathbf{A} , que expressa a matriz \mathbf{A} como um resumo de projeções uni-dimensionais $\mathbf{e}_i \mathbf{e}_i^T$. A decomposição de matrizes é fundamental na teoria da álgebra linear possuindo muitas aplicações aos problemas existentes nas ciências naturais e ciências correlatas.

1.5 Análise de Espectro Singular (SSA)

O método SSA aplicado a análise de séries temporais foi introduzida por Broomhead & King (1986a, 1986b). Atualmente, existe uma vasta literatura a respeito do método e seu desenvolvimento desde então, como por exemplo Buchstaber (1994), Elsner & Tsonis (1996), Danilov (1997) e Golyandina *et al.* (2001). Várias modificações, aplicáveis às mais diversas situações, foram propostas como a *Multichannel SSA* que estende a SSA padrão para a análise multivariada de séries (BROOMHEAD & KING, 1986b), a SSA para análise de séries contínuas, a SSA sequencial para detecção de mudança estrutural (MOSKVINA & ZHIGLJAVSKY, 2000), dentre outras. O termo “espectro singular” provém da decomposição espectral (que conserva os autovalores) de uma matriz \mathbf{A} em um conjunto (espectro) de autovalores. Estes autovalores, λ , são os valores que tornam a matriz $\mathbf{A} - \lambda \mathbf{I}$ singular. O termo “Análise de Espectro Singular” não é o mais adequado, pois a decomposição tradicional em autovalores envolvendo dados multivariados é também uma análise de espectro singular. Mais propriamente, a análise de espectro singular poderia ser chamada de análise de séries temporais utilizando espectro singular (ELSNER & TSONIS, 1996).

Dois grupos principais desenvolveram algoritmos para a análise de séries temporais baseados na SSA padrão. Ghil *et al.* (2002) perceberam que o método poderia ser utilizado tanto como um método de domínio das frequências quanto de domínio temporal, para a análise de séries temporais e investigaram as propriedades e aplicações do método. Baseado na metodologia deste grupo, foi desenvolvido o software kSpectra Toolkit[®]. Outra abordagem bastante conhecida é o método Caterpillar, desenvolvida por Golyandina *et al.* (2001) e que deu origem aos softwares Caterpillar-SSA[®] e CatMV[®]. O primeiro faz a análise espectral de séries temporais completas, enquanto o segundo permite a entrada de séries incompletas e é capaz de analisar e, simultaneamente, preencher as falhas de dados presentes.

A SSA incorpora elementos da análise clássica de séries temporais, estatísticas multivariadas, geometria multivariada, dinâmica de sistemas e processamento de sinal. Apesar de muitos elementos probabilísticos e estatísticos estarem presentes na base da SSA, ela não é um método estatístico, em termos de estatística clássica. Particularmente, não se faz suposições estatísticas relativas ao sinal ou ao ruído durante a realização da análise da série e da investigação da propriedade dos algoritmos implementados (GOLYANDINA *et al.*, 2001). O método pode ser entendido como uma generalização da Análise de Componentes Principais (ACP), no qual ao invés de se ter variáveis obtidas simultaneamente em diferentes locais no espaço, tem-se várias versões de uma única série temporal, cada uma com uma defasagem temporal diferente. Para efeitos da SSA, defasagem é o procedimento que transforma uma série temporal simples em um conjunto multivariado de observações ou, matematicamente, a defasagem transforma a série temporal $F_N = \{f_0, \dots, f_{N-1}\}$, em múltiplas séries X_1, \dots, X_K com vetores $X_i = (f_{i-1}, \dots, f_{i+L-1})^T \in \mathbb{R}^L$, onde $K = N - L + 1$. Os vetores X_i são chamados de vetores defasados e K corresponde à quantidade de vetores defasados formados na etapa de inserção a partir da fixação do tamanho da janela L .

O único parâmetro de entrada para que a análise da série temporal seja feita através da SSA, é o tamanho da janela, L . A escolha deste parâmetro, que é essencial para a etapa de decomposição da série, depende das propriedades da série e da finalidade da análise. Entretanto, segundo Hassani (2007), alguns resultados teóricos indicam que o valor de L deve ser grande o suficiente, mas não superior a $N / 2$. Se L for suficientemente grande, pode-se considerar cada vetor defasado X_i como uma série isolada e pode-se investigar a dinâmica e as características de cada um deles. O exemplo mais simples de aplicação desta técnica é o bem conhecido método da média móvel, no qual calcula-se a média de cada vetor defasado obtido.

Caso a série temporal possua um componente periódico conhecido (por exemplo, um componente sazonal), uma melhor separabilidade da componente periódica é obtida fixando-se um tamanho de janela proporcional a esse período. É importante notar, que a SSA não utiliza nenhum modelo paramétrico para estimar os componentes de tendência e sazonalidade. Assim, ao analisar uma série temporal real com a ajuda de SSA dificilmente se pode esperar obter componentes harmônicos exatos ou uma tendência exatamente linear na etapa de decomposição, mesmo que estes componentes estejam presentes nas séries. Isto ocorre devido a influência de ruídos e é uma consequência da natureza não-paramétrica do método. Normalmente, no entanto, pode-se obter uma boa aproximação para as séries

(GOLYANDINA *et al.*, 2001).

1.5.1 Matriz Trajetória e Matriz de Covariância Defasada

A implementação da SSA parte da construção da chamada matriz trajetória, obtida a partir da fixação do tamanho da janela, L . Cada um dos K vetores defasados são, então, dispostos nas linhas de uma matriz, denominada matriz trajetória, \mathbf{X} , que possui dimensão

$K \times L$. A matriz de covariância defasada, \mathbf{S} , é calculada através do produto entre a matriz trajetória \mathbf{X} e a sua transposta \mathbf{X}^T . \mathbf{S} é uma matriz quadrada de dimensões $L \times L$ cujos elementos são proporcionais aos coeficientes de correlação linear entre os pares de vetores defasados utilizados na construção da matriz trajetória. Todos os elementos de \mathbf{S} são números reais e a matriz \mathbf{S} é simétrica, ou seja, $\mathbf{S} = \mathbf{S}^T$ (ELSNER & TSONIS, 1996).

1.5.2 Espectro Singular

Uma vez que \mathbf{S} é uma matriz real e simétrica, há uma matriz \mathbf{E} cujas colunas são ortonormais e uma matriz $\mathbf{\Lambda}$ tal que

$$\mathbf{S} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^T \quad (1.20)$$

ou

$$\mathbf{S} \mathbf{E} = \mathbf{E} \mathbf{\Lambda} \quad (1.21)$$

a qual é chamada de decomposição espectral da matriz \mathbf{S} . Lembrando que $\mathbf{\Lambda}$ é uma matriz diagonal na qual as entradas não negativas são os autovalores de \mathbf{S} e, usando a definição de \mathbf{S} , tem-se

$$\mathbf{X}^T \mathbf{X} \mathbf{E} = \mathbf{E} \mathbf{\Lambda}, \quad (1.22)$$

$$\mathbf{E}^T \mathbf{X}^T \mathbf{X} \mathbf{E} = \mathbf{\Lambda}, \quad (1.23)$$

ou

$$(\mathbf{X} \mathbf{E})^T (\mathbf{X} \mathbf{E}) = \mathbf{\Lambda} \quad (1.24)$$

A matriz $\mathbf{X} \mathbf{E}$ é a matriz trajetória projetada na base \mathbf{E} . A matriz diagonal $\mathbf{\Lambda}$ consiste em valores ordenados $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$. A raiz quadrada destes valores são chamados de

valores singulares de \mathbf{X} . Estes valores singulares ordenados são chamados, conjuntamente, de espectro singular. Observe que autovetores de \mathbf{S} e a raiz quadrada dos autovalores de \mathbf{S} são os vetores singulares e os valores singulares de \mathbf{X} . Assim, \mathbf{X} pode ser escrita como (ELSNER & TSONIS, 1996):

$$\mathbf{X} = \mathbf{D} \mathbf{L} \mathbf{E}^T \quad (1.25)$$

então, a partir da definição de \mathbf{S} , pode-se escrever

$$\mathbf{S} = \mathbf{X}^T \mathbf{X} = (\mathbf{D} \mathbf{L} \mathbf{E}^T)^T (\mathbf{D} \mathbf{L} \mathbf{E}^T) = \mathbf{E} \mathbf{L} \mathbf{D}^T \mathbf{D} \mathbf{L} \mathbf{E}^T \quad (1.26)$$

como $\mathbf{D}^T \mathbf{D} = \mathbf{I}$, então

$$\mathbf{S} = \mathbf{E} \mathbf{L}^2 \mathbf{E}^T \quad (1.27)$$

segue que

$$\mathbf{L}^2 = \mathbf{\Lambda} \quad (1.28)$$

assim, os autovetores de \mathbf{S} são os vetores singulares a direita de \mathbf{X} . Similarmente, pode-se demonstrar que os vetores singulares a esquerda de \mathbf{X} são os autovetores de $\mathbf{X}\mathbf{X}^T$.

Os aspectos de álgebra linear apresentados neste capítulo constituem a base matemática da análise espectral utilizado no método da SSA. O formalismo matemático do algoritmo implementado nos softwares CatMV[®] e Caterpillar[®] é descrito no próximo capítulo.

Capítulo 2

Metodologia

Este capítulo descreve o método de obtenção e o pré-processamento dos dados de vazão média mensal, precipitação acumulada mensal e dos dados de qualidade da água. Cita os *softwares* utilizados nas análises, no armazenamento dos dados, confecção de gráficos e análises estatísticas. Aborda o método de obtenção do periodograma, dos erros médios absolutos (EMA), dos erros médios relativos (EMR) e descreve matematicamente os dois estágios de decomposição e reconstrução dos algoritmos implementados nos softwares CatMV[®] e Caterpillar[®].

2.1 Aquisição e Pré-processamento dos Dados

Os dados utilizados neste estudo foram obtidos do sistema *HidroWeb*³. Os dados de vazão média mensal e de qualidade da água são provenientes das estações de monitoramento localizadas na bacia hidrográfica do rio Cuiabá, enquanto os dados de precipitação são provenientes das estações pluviométricas distribuídas pelo estado de Mato Grosso. Foram obtidos dados de nove estações de medição de vazão, doze estações de análise de qualidade da água e de quarenta estações pluviométricas. A Figura 2.1 mostra a localização dos pontos de amostragem de qualidade da água e de medição de vazão existentes na área da bacia do Rio Cuiabá, destacando as estações utilizadas neste estudo.

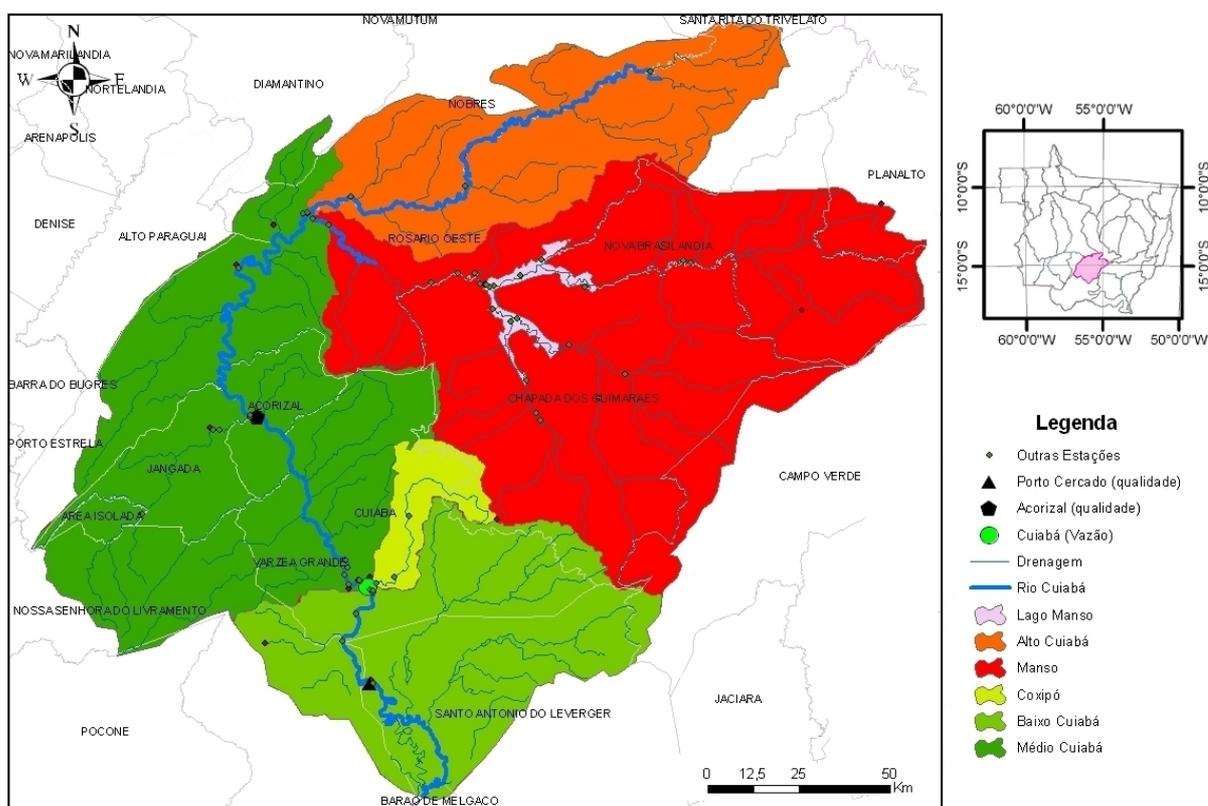


Figura 2.1 – Bacia hidrográfica do Rio Cuiabá e pontos de amostragem.

A localização geográfica das estações pluviométricas cujos dados foram obtidos para este estudo, e que localizam-se no Estado de Mato Grosso, é mostrada na Figura 2.2.

3 <http://hidroweb.ana.gov.br/>

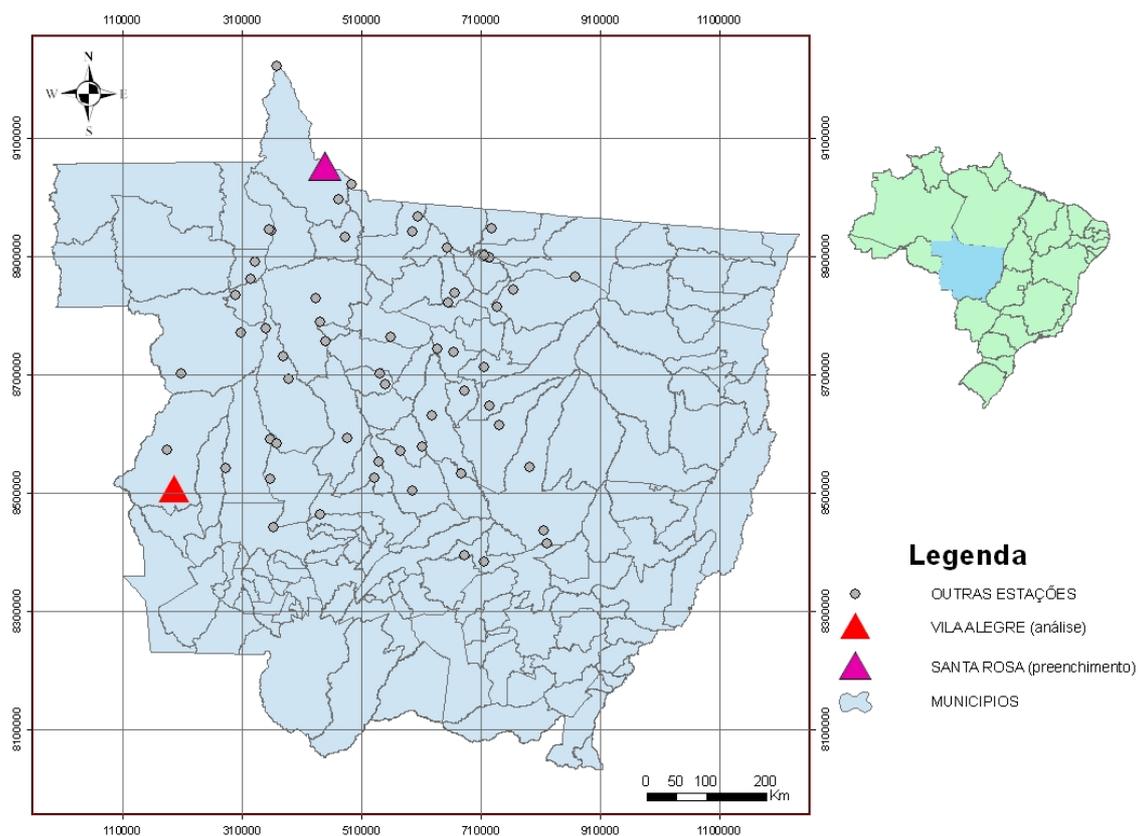


Figura 2.2 – Estações pluviométricas localizadas no Estado de Mato Grosso cujos dados foram utilizados neste estudo.

2.1.1 Vazão Média Mensal

O sistema de informações hidrometeorológicas da ANA disponibiliza os dados de vazão máxima, vazão média e vazão mínima das estações de monitoramento. Neste estudo foram utilizados as séries de vazão média das estações. Os dados obtidos foram os referentes as estações Marzagão (66140000), Rosário Oeste (66250001), Acorizal (66255000), Cuiabá (66260001), Barão de Melgaço (66280000), Porto Cercado (66340000), Quilombo (66175000), Casca (66200000) e Manso (66231000).

Destas estações, apenas a estação Cuiabá e a estação Acorizal não apresentaram falhas no conjunto de dados e, conseqüentemente, poderiam ser utilizadas para análise sem que nenhum outro procedimento fosse necessário. A estação Acorizal possui 504 observações, correspondentes ao período de janeiro de 1966 a dezembro de 2007, enquanto a estação Cuiabá possui 552 observações, correspondentes ao período de dezembro de 1962 a setembro de 2007. Tendo em vista a maior quantidade de dados presentes na série temporal da estação Cuiabá, bem como a importância dos dados desta estação no contexto da gestão de recursos

hídricos, a análise de séries temporais nos dados de vazão média mensal foi feita nesta estação. Das estações que apresentaram falhas de dados se considerou mais importante a análise e preenchimento simultâneo das falhas da estação Porto Cercado, uma vez que esta estação localiza-se na área do Pantanal matogrossense, onde muitos estudos ambientais são realizados e, as falhas presentes na série temporal desta estação podem prejudicar estudos que dependam destes dados. A estação Porto Cercado possui uma série temporal que compreende o período de dezembro de 1969 a junho de 2006. Neste período foram feitas 375 observações e ocorreram 65 falhas de dados.

2.1.2 Precipitação Acumulada Mensal

Foram obtidos dados de precipitação acumulada mensal de 40 estações pertencentes a rede de monitoramento do Estado de Mato Grosso. Destas, apenas as estações Vila Alegre (1359001) e Novo Planeta (957001) não apresentaram falhas de dados em suas séries temporais. Das estações com falhas, a que apresentou a maior quantidade foi a estação Fazenda Santa Emília (1053001), com 15,6% de falhas (23 valores faltantes). Duas estações, Fontanilhas (1158001) e Deciolândia (1457003), apresentaram apenas uma falha (0,6%). As demais estações apresentaram de 2 a 20 falhas (de 1,2% a 12% de falhas).

Para a análise das séries temporais optou-se pela estação Vila Alegre, tendo em vista a maior quantidade de observações mensais ($N=168$), correspondentes ao período de janeiro de 1993 a dezembro de 2006. Para o preenchimento de falhas, optou-se pela série da estação Santa Rosa (857000). Esta estação possui 159 observações e 9 falhas de dados (cerca de 5% de dados faltantes). Para a comparação com o método da Regressão Linear Simples, obteve-se, ainda, dados de 6 estações consideradas próximas à estação Santa Rosa, sendo que 2 delas – Barra do São Manuel (758000) e Mineração São Benedito (957000) – pertencem aos estados do Amazonas e Pará, respectivamente.

2.1.3 Variáveis de Qualidade da Água

Os dados de qualidade da água referem-se às estações Acorizal, Barão de Melgaço, Córrego do Barbado, Nobres, Passagem da Conceição, Porto Cercado, São Gonçalo, Mané Pinto, Ribeirão dos Cocais, Marzagão, Rosário Oeste e Praia do Poço. Os dados destas estações foram analisados inicialmente por Nascimento (2009). Todas as variáveis, de todas as estações apresentaram falhas de dados. Para o preenchimento das falhas, foi escolhida a

variável Turbidez, em função das características sazonais apresentadas. Este trabalho mostra apenas o preenchimento para a estação Acorizal, tendo em vista que é a estação que apresenta maior número de falhas de dados para a variável turbidez (39,0%). A qualidade dos dados de qualidade da água é discutida no ítem 3.2.3.

2.2 Séries Temporais Analisadas

Para a análise de séries temporais foi selecionada uma estação de medição de vazão (estação Cuiabá) e uma estação de precipitação (estação Vila Alegre) que não possuíam falhas de dados para que se pudesse aplicar o método SSA. Das estações com falhas de dados, foram escolhidas uma de vazão (estação Porto Cercado) uma de precipitação (estação Santa Rosa), e uma de monitoramento da qualidade da água (estação Acorizal) para se efetuar a análise com preenchimento simultâneo das falhas. A Tabela 2.1 apresenta resumidamente as características dos dados das estações escolhidas.

Tabela 2.1 – Séries temporais analisadas com o método SSA.

Informações	Cuiabá	Porto Cercado	Vila Alegre	Santa Rosa	Acorizal
Código	66260001	66340000	1359001	857000	66255000
Variável	Vazão média	Vazão média	Precipitação	Precipitação	Turbidez
Período de Monitoramento	jan/62 a dez/07	dez/69 a jun/06	jan/93 a dez/06	jan/93 a dez/06	jan/93 a mar/06
N	552	439	168	168	159
Observações	552	374	168	159	97
% Falhas	0,00%	14,81%	0,00%	5,36%	39,00%

A previsão de dados foi feita tanto para a série temporal de vazão média mensal da estação Cuiabá, quanto para a série de precipitação acumulada mensal da estação Vila Alegre. As séries foram, inicialmente, truncadas para que se pudesse calcular os erros da previsão. A seguir, foi feita a previsão para dois anos a partir das séries completas.

2.3 Softwares utilizados nas análises e preenchimento

As análises das séries temporais completas foram feitas utilizando os algoritmos implementados no *software* Caterpillar® e as análises com preenchimento simultâneo das falhas foram feitas utilizando os algoritmos implementados no *software* CatMV®. Foram utilizados, ainda, os *softwares* Gnumeric®, para o armazenamento dos dados e confecção dos gráficos, e o pacote estatístico R®, para a realização de algumas análises estatísticas.

2.4 Cálculo do Periodograma

Para descrever séries estacionárias finitas, porém com um número razoavelmente grande de observações, é conveniente usar a expansão de Fourier da série inicial. A expansão é dada por:

$$f_n = c_0 + \sum_{k=1}^{\lfloor N/2 \rfloor} (c_k \cos(2\pi n k / N) + s_k \sin(2\pi n k / N)) \quad (2.1)$$

onde N é o número de observações, $0 \leq n < N$, e $s_{N/2} = 0$ para N par. O termo c_0 é igual a média da série e, se a série é centralizada, $c_0 = 0$. O periodograma da série é análogo a medida

espectral. Por definição, o periodograma $\prod_f^N(\omega)$ da série $F = (f_0, \dots, f_{N-1})$ é:

$$\prod_f^N \frac{1}{N} \left| \sum_{n=0}^{N-1} e^{-i2\pi\omega n} f_n \right|^2, \quad \omega \in (-1/2, 1/2] \quad (2.2)$$

uma vez que os elementos da série F são números reais, $\prod_f^N(-\omega) = \prod_f^N(\omega)$ para $|\omega| < 1/2$ e, além disso, pode-se considerar apenas o intervalo $[0, 1/2]$ para ω . Se a série pode ser representada pela expansão de Fourier, pode-se demonstrar que:

$$\prod_f^N(k/N) = \frac{N}{2} \begin{cases} 2c_0^2 & \text{para } k=0, \\ c_k^2 + s_k^2 & \text{para } 0 < k < N/2, \\ 2c_{N/2}^2 & \text{para } k = N/2. \end{cases} \quad (2.3)$$

o último caso só é possível caso N seja ímpar. Considerando duas séries que podem ser representadas através da equação 2.1 o periodograma pode ser descrito pela da equação:

$$\|F\|^2 = \sum_{k=0}^{\lfloor N/2 \rfloor} \prod_f^N(k/N) \quad (2.4)$$

A equação 2.4 implica que os valores do periodograma no ponto k/N descrevem a influência dos componentes harmônicos com frequência $\omega = k/N$ na série. O periodograma pode ser utilizado para determinar quais componentes harmônicos estão presentes na série (GOLYANDINA *et al.*, 2001).

2.5 Estimativa dos Erros

O Erro Médio Absoluto (EMA) é uma grandeza utilizada para medir o quão próximo um valor estimado está do valor observado. O EMA é calculado pela seguinte expressão:

$$EMA = \frac{1}{N} \sum_{t=1}^N |F_t - Y_t| = \frac{1}{N} \sum_{t=1}^N |e_t| \quad (2.5)$$

onde F_t é o valor observado, Y_t é o valor estimado e N é o número de observações. Como o nome sugere, o EMA corresponde a média dos erros absolutos e pode ser utilizado na determinação do erro de previsão de um valor na análise de séries temporais.

O Erro Médio Relativo (EMR) é uma medida da precisão de um valor ajustado em uma série temporal. É expresso em termos de percentagem, e é definido pela seguinte expressão:

$$EMR = \frac{1}{N} \sum_{t=1}^N \frac{|F_t - Y_t|}{F_t} \quad (2.6)$$

Embora o conceito do EMR seja claro e convincente, existe pelo menos um inconveniente na sua utilização: se existirem valores iguais a zero na série temporal, ocorrerá uma divisão por zero. Nestes casos, todas as medidas iguais a zero devem ser descartadas ao se calcular o EMR.

2.6 Método da Análise de Espectro Singular (SSA)

O método da Análise de Espectro Singular consiste em dois estágios complementares. O primeiro estágio decompõe a série original, enquanto o segundo estágio reconstrói a série. Cada estágio, por sua vez, possui dois passos. O primeiro estágio inclui a inserção da série num espaço vetorial (*embedding*) e a Decomposição em Valores Singulares (*Singular Value Decomposition*, SVD), enquanto o segundo estágio compreende o agrupamento e a diagonalização (GOLYANDINA *et al.*, 2001; HASSANI, 2007).

Na primeira etapa (denominada de inserção), um série unidimensional é representada como uma série multidimensional cuja dimensão é chamada de tamanho da janela, L . A série temporal multidimensional (que é uma sequência de vetores) forma a matriz trajetória. O único (e mais importante) parâmetro de entrada é o tamanho da janela. A segunda etapa, SVD, é a decomposição em valores singulares da matriz trajetória em uma soma de matrizes bi-

ortogonais de posto⁴ 1. A etapa de agrupamento corresponde a separação das matrizes, calculadas da etapa anterior, em vários grupos e a soma das matrizes dentro de cada grupo. O resultado desta etapa é a representação da matriz trajetória em uma soma de várias matrizes resultantes. A última etapa transforma cada matriz resultante em uma nova série temporal, as quais representam os componentes presentes na série inicial. Isto é feito através de uma operação chamada de diagonalização, que é uma operação linear aplicada na matriz trajetória e que transforma ela própria na série inicial. Assim, a série inicial é decomposta em vários componentes independentes (GOLYANDINA *et al.*, 2001) e que podem ser agrupados e reconstruídos numa nova série de tamanho N .

2.6.1 Decomposição

2.6.1.1 Inserção da série no espaço vetorial

Considerando uma série temporal $F_N = (f_0, \dots, f_{N-1})$ de tamanho N , fixando o tamanho da janela, L , sendo $1 < L < N$, a etapa de inserção da série num espaço vetorial, transforma a série original em uma sequência de vetores defasados de dimensão L $\{X_i\}_{i=1}^K$, onde $K = N - L + 1$, ou seja:

$$X_i = (f_{i-1}, \dots, f_{i+L-2})^T, \quad (2.7)$$

sendo $1 \leq i \leq K$. A matriz trajetória da série F_N , é formada pelos vetores defasados:

$$\mathbf{X} = [X_1 : \dots : X_K] \quad (2.8)$$

Em outras palavras, a matriz trajetória é:

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix} \quad (2.9)$$

Obviamente, $x_{ij} = f_{i+j-2}$ e a matriz \mathbf{X} tem valores iguais nas diagonais, $i + j$ é constante. Matrizes deste tipo são chamadas Matrizes de Hankel.

4 Chama-se de posto de uma matriz o número de linhas ou colunas linearmente independentes. Em outras palavras, dois vetores, u e v , são linearmente independentes quando não são proporcionais, ou seja, não existe nenhum número real β que verifique $u = \beta \times v$.

2.6.1.2 Decomposição em Valores Singulares (SVD)

O resultado desta etapa é a decomposição da matriz trajetória \mathbf{X} , em seus valores singulares. Considere a matriz $\mathbf{S} = \mathbf{X} \mathbf{X}^T$ de dimensão $L \times L$. Denotando por $\lambda_1, \dots, \lambda_L$ os autovalores de \mathbf{S} , em ordem decrescente de magnitude ($\lambda_1 \geq \dots \geq \lambda_L \geq 0$), por U_1, \dots, U_L os autovetores correspondentes. Deixando $d = \max \{i, \text{tal que } \lambda_i > 0\}$. Se denotarmos $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ ($i = 1, \dots, d$), então, o resultado da decomposição em valores singulares da matriz trajetória \mathbf{X} pode ser expressa como:

$$\mathbf{X} = \mathbf{X}_1 + \dots + \mathbf{X}_d, \quad (2.10)$$

onde d é a quantidade de autovalores diferentes de zero (da matriz \mathbf{S}) e $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$. As matrizes \mathbf{X}_i possuem posto 1. Portanto, elas são matrizes elementares. Os termos $(\sqrt{\lambda_i}, U_i, V_i)$ são chamados de *eigen triples* da SVD.

2.6.2 Reconstrução

2.6.2.1 Agrupamento

Uma vez que a expansão foi obtida, o procedimento de agrupamento particiona o conjunto de índices $\{1, \dots, d\}$ em m subconjuntos I_1, \dots, I_m . Fazendo $I = \{i_1, \dots, i_p\}$, então a matriz resultante \mathbf{X}_I correspondente ao grupo I é definida como $\mathbf{X}_I = \mathbf{X}_{i_1} + \dots + \mathbf{X}_{i_p}$. Estas matrizes são calculadas para $I = I_1 + \dots + I_m$ e a expansão leva à decomposição:

$$\mathbf{X} = \mathbf{X}_{I_1} + \dots + \mathbf{X}_{I_m}. \quad (2.11)$$

O procedimento de escolha do conjunto de $I = I_1, \dots, I_m$ é chamado de agrupamento.

2.6.2.2 Diagonalização

O último passo da SSA transforma cada matriz resultante da decomposição e que fora agrupada, em uma nova série de tamanho N . Chamando de \mathbf{Y} a matriz $L \times K$ com elementos y_{ij} , $1 \leq i \leq L$, $1 \leq j \leq K$. Definindo $L^* = \min(L, K)$, $K^* = \max(L, K)$ e $N = L + K - 1$. Fazendo $y_{ij}^* = y_{ij}$ se $L < K$ e $y_{ij}^* = y_{ji}$ caso contrário. Então, a diagonalização transforma a matriz \mathbf{Y} em uma série g_0, \dots, g_{N-1} através da expressão:

$$g_k = \begin{cases} \frac{1}{k+1} \sum_{m=1}^{k+1} y_{m, k-m+2}^* & \text{para } 0 \leq k \leq L^* - 1 \\ \frac{1}{L^*} \sum_{m=1}^{L^*} y_{m, k-m+2}^* & \text{para } L^* - 1 \leq k \leq K^* \\ \frac{1}{N-k} \sum_{m=k-K^*+2}^{N-K^*+1} y_{m, k-m+2}^* & \text{para } K^* \leq k \leq N \end{cases} \quad (2.12)$$

A expressão acima corresponde aos elementos das diagonais da matriz $i + j = k + 2$. A escolha $k = 0$ leva a $g_0 = y_{11}$, para $k = 1$ tem-se $g_1 = (y_{12} + y_{21})/2$ e assim por diante. Note que se a matriz \mathbf{Y} é a matriz trajetória de uma série (h_0, \dots, h_{N-1}) , em outras palavras, se \mathbf{Y} é uma matriz de Hankel, então $g_i = h_i$ para todo i . A diagonalização aplicada a matriz resultante \mathbf{X}_{lk} produz a série $\tilde{F}^{(k)} = (\tilde{F}_0^{(k)}, \dots, \tilde{F}_{N-1}^{(k)})$ e, portanto, a série inicial f_0, \dots, f_{N-1} é decomposta em uma soma de m séries:

$$f_n = \sum_{k=1}^m \tilde{f}_n^{(k)} \quad (2.13)$$

As quatro etapas descritas, estão implementadas nos *softwares* Caterpillar[®], utilizado na análise das séries temporais, e CatMV[®], utilizado para a análise e preenchimento simultâneo de falhas, cujos resultados são apresentados e discutidos no próximo capítulo.

Capítulo 3

Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos com a aplicação do método SSA para analisar o comportamento de séries temporais completas e preencher falhas de dados de séries temporais incompletas. O método foi aplicado em séries de dados de vazão média mensal, precipitação acumulada mensal e em uma variável de qualidade da água (turbidez). Os dados de vazão média mensal são provenientes das estações Cuiabá (66260001) e Porto Cercado (66340000), os dados de precipitação são provenientes das estações Vila Alegre (1359001) e Santa Rosa (857000) e os dados de turbidez são provenientes da estação Acorizal (66255000).

Na primeira seção são mostrados os resultados da análise espectral de séries temporais completas das variáveis vazão média mensal e precipitação acumulada mensal. Na segunda seção são expostos os resultados da análise espectral e do preenchimento simultâneo das falhas de dados das séries incompletas de vazão média mensal, precipitação acumulada mensal e de turbidez. Com o intuito de demonstrar a robustez do método SSA, uma análise comparativa entre os métodos SSA e Regressão Linear Simples (RLS) no preenchimento de falhas de dados de série de precipitação é apresentada. E, por fim, são apresentados os resultados da previsão de dados para as séries temporais de vazão média mensal da estação Cuiabá e de precipitação acumulada mensal da estação Vila Alegre.

3.1 Análise das Séries Temporais

Para demonstrar a capacidade do método SSA na extração de informações sobre o comportamento de séries temporais foi utilizado o *software* Caterpillar[®]. A seguir, são mostrados os resultados obtidos na análise da variável vazão média mensal e, a seguir, os resultados da análise da variável precipitação acumulada mensal.

3.1.1 Vazão Média Mensal - Estação Cuiabá (66260001)

Os resultados são mostrados sob duas óticas distintas. Primeiro através da análise visual do gráfico da série temporal e, a seguir, através do método SSA. O gráfico da série temporal de vazão média mensal da estação Cuiabá pode ser observado na Figura 3.1. A série possui 552 observações ($N=552$) mensais, correspondentes ao período de janeiro de 1962 a dezembro de 2007.

Pela análise visual da série, nota-se um comportamento periódico, apresentando sazonalidade com período de 12 meses. Os maiores valores das vazões ocorrem preferencialmente nos meses de janeiro, fevereiro e março, enquanto os valores mais baixos ocorrem preferencialmente nos meses de julho, agosto e setembro. A maior vazão observada no período de monitoramento foi de 2.232,7 m³/s, em fevereiro de 1994 (ano hidrológico 93/94), enquanto a vazão mínima observada foi de 46,0 m³/s, em setembro de 1964 (ano hidrológico 63/64). A Figura 3.1 mostra picos acentuados nos meses de janeiro de 1979 e 1994, e em fevereiro de 1980, e picos intermediários nos meses de março de 1974 e 1980, e fevereiro 1982, 1992, 1996 e 2003.

De maneira geral, a análise visual das séries temporais permite revelar os padrões de comportamento da série tais como tendências de crescimento ou decréscimo, comportamento cíclico (sazonal) e *outliers* (valores discrepantes). Entretanto, para que se possa extrair informações que não podem ser facilmente observadas a partir do gráfico cartesiano é necessário a aplicação de métodos mais específicos para que se possa fazer uma descrição mais detalhada sobre o comportamento da série

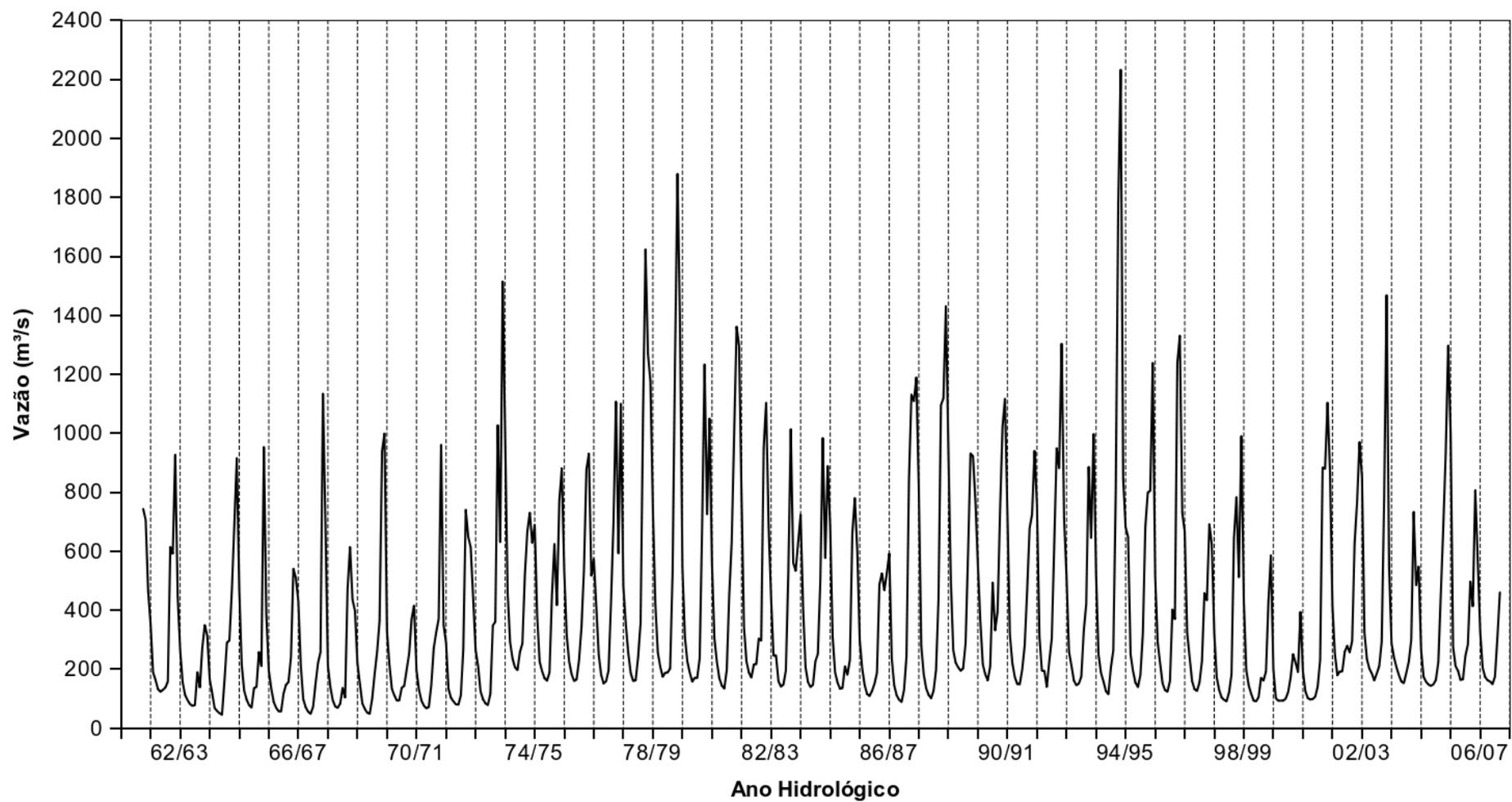


Figura 3.1 – Série temporal das vazões médias no rio Cuiabá, na estação de Cuiabá, durante o período de janeiro de 1961 a dezembro de 2007.

Conforme o exposto, o algoritmo implementado no software Caterpillar® é composto por duas etapas de cálculo. A primeira realiza a decomposição da série em autovetores defasados e a segunda realiza a reconstrução da série a partir dos autovetores defasados escolhidos.

3.1.1.1 Decomposição da Série Original

À partir de conhecimento prévio, bem como da análise visual da série temporal, o tamanho da janela (L) foi fixado em 276 para que a decomposição da série pudesse ser realizada. Foi escolhido um valor para L divisível por 12, tendo em vista que a série possui periodicidade anual, e o mais próximo possível da metade da série ($N/2$), conforme sugerido por Hassani (2007) e Golyandina & Osipov (2007). Especificamente neste caso, o valor para L foi igual à metade da série. Baseado no tamanho da janela fixado, na etapa de decomposição obteve-se um conjunto de 277 vetores defasados ($K=277$).

O principal componente harmônico da série pode ser visualizado a partir do comportamento das funções de covariância dos autovetores. A covariância pode ser entendida como uma medida da variação conjunta de duas variáveis aleatórias ao longo do tempo, e fornece uma medida não padronizada do grau no qual as variáveis se movem juntas. No caso das séries temporais, onde apenas uma variável está envolvida na análise, a covariância deve ser entendida como uma medida da variação conjunta de dois vetores ao longo do tempo. Segundo Golyandina *et al.* (2001), funções de covariância de séries temporais periódicas e de séries aperiódicas ou caóticas, possuem comportamentos consideravelmente diferentes. No caso das séries aperiódicas ou caóticas, as funções de covariância tendem a zero, enquanto quase todas as séries periódicas são caracterizadas pelo comportamento periódico das suas funções de covariância.

No gráfico das covariâncias médias entre os vetores, mostrado na Figura 3.2, nota-se o comportamento periódico das funções de covariância da série original. A distância, no eixo das abcissas, entre maiores valores de covariância, assim como entre os menores, é igual a 12 e representa a principal periodicidade da série. O gráfico das covariâncias confirma o resultado inicial da periodicidade mostrado na análise visual da Figura 3.1.

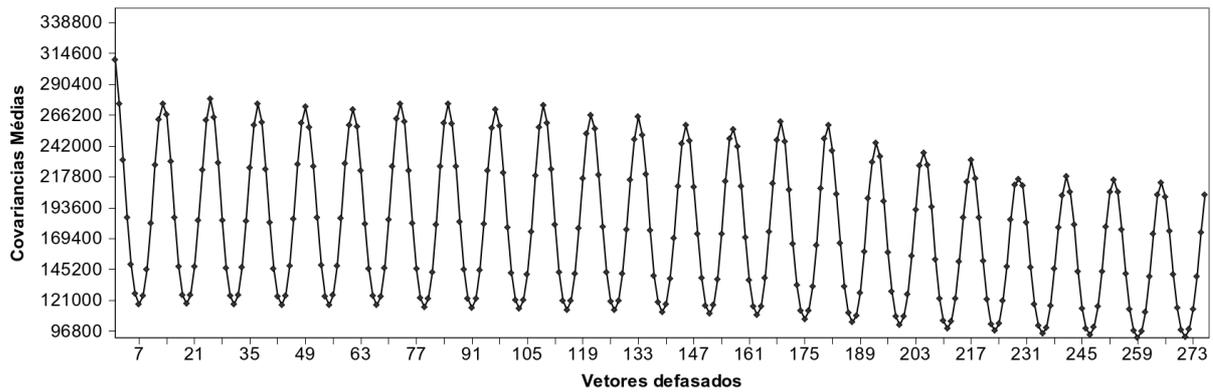


Figura 3.2 – Covariâncias médias dos vetores defasados.

Uma informação adicional a respeito da periodicidade é dada a partir da análise do periodograma, onde pode-se observar os componentes sazonais extraídos da série temporal. Dois componentes harmônicos de frequências iguais a 6 e 12 são distinguíveis no periodograma representado na Figura 3.3.

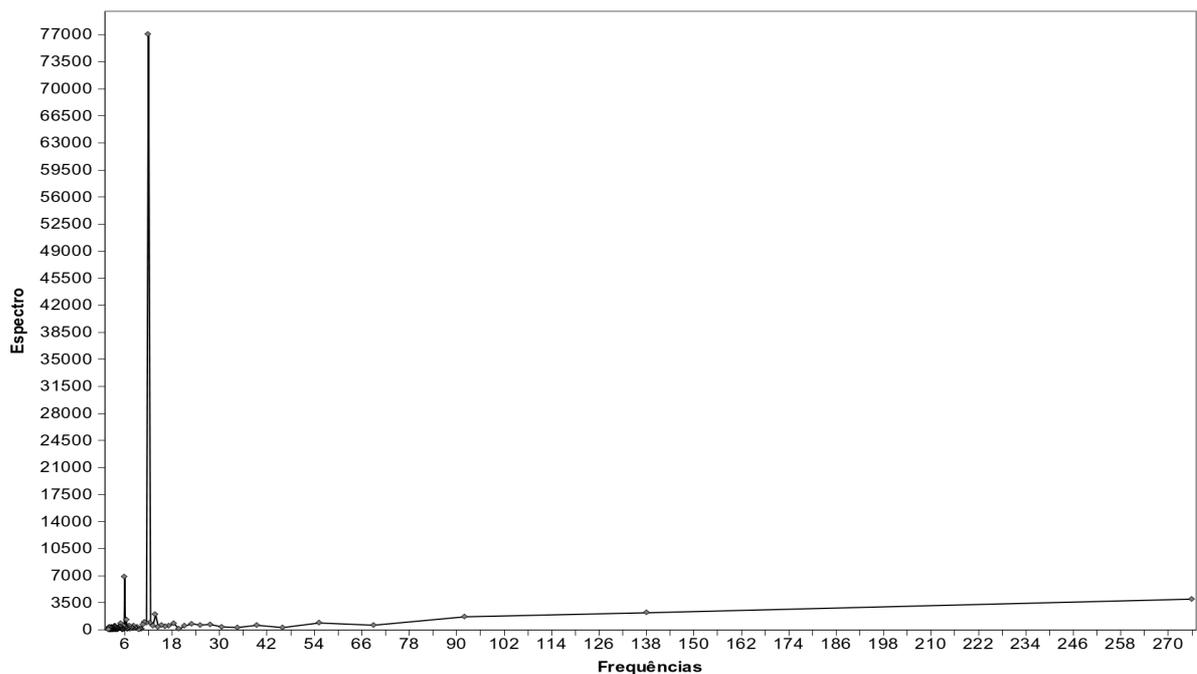


Figura 3.3 – Periodograma médio da série temporal.

Um dos componentes diz respeito a periodicidade anual, já mencionada anteriormente e conhecida através da análise visual. O outro componente harmônico possui periodicidade igual a 6, e não é facilmente identificável na análise visual da série temporal (Figura 3.1), embora esteja razoavelmente bem destacado no periodograma. Enquanto a periodicidade anual pode ser atribuída ao ciclo cheia/estiagem, a periodicidade igual a 6 meses é menos clara, mas representa, provavelmente, a variação que ocorre dentro de cada período do ciclo

cheia/estiagem. As Figuras 3.2 e 3.3 contém a informação de que a série original possui dois componentes harmônicos. Um de período igual a 12 e outro de período igual a 6 meses. A identificação dos principais componentes harmônicos extraídos da série original é necessário pois, no processo de reconstrução da série a partir dos autovetores, deve-se escolher aqueles que representam estes componentes harmônicos, assim como a tendência da série. A escolha é feita através da separação entre o sinal e os ruídos, descrita a seguir.

3.1.1.2 Separação do Sinal e Ruído

Como destacado por Hassani (2007) e Golyandina *et al.* (2001), cada componente harmônico da série resulta em um par de autovetores de mesmo período e valores singulares próximos, enquanto componentes de tendência resultam em um autovetor isolado. Além disso, ruídos produzem uma sequência de autovetores com valores singulares próximos, que formam um leve e constante decréscimo quando *plotados* em um gráfico. Assim, a separação entre o sinal e os ruídos da série podem ser observados no gráfico dos valores singulares dos autovetores (Figura 3.4). Para que a visualização dos pares de autovetores não ficasse muito prejudicada, o eixo das abscissas mostra apenas os valores dos 70 primeiros autovetores.

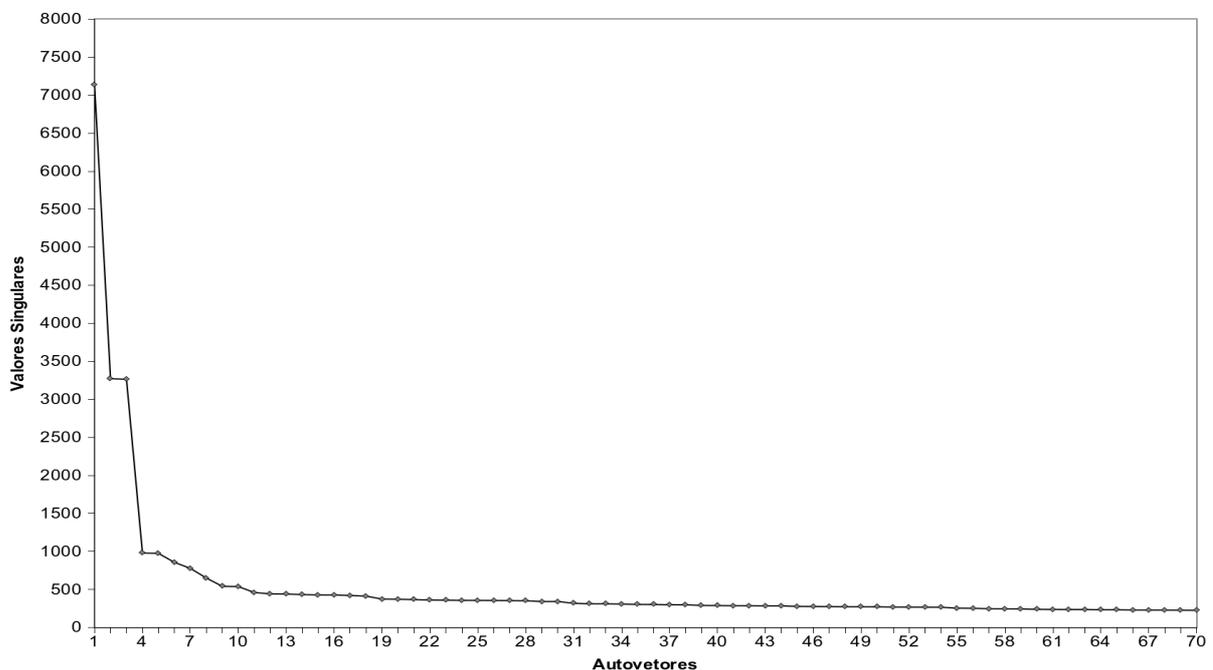


Figura 3.4 – Valores singulares dos autovetores defasados.

Na Figura 3.4, pode-se observar que o valor singular do autovetor 1 é consideravelmente maior que os valores singulares dos outros autovetores. Assim, pode-se dizer que o autovetor 1 está bem separado dos demais autovetores e representa a tendência

principal da série como pode ser observado na Figura 3.5 (a). Os autovetores 2 e 3 possuem valores singulares bastante próximos, formando um platô no gráfico e, portanto, representam um componente harmônico. Pode-se notar que este par de autovetores possui período igual a 12, observando a Figura 3.5 (b) e (c). O platô formado pelos autovetores 4 e 5 também representa um componente harmônico, neste caso de período igual a 6 como pode ser observado na Figura 3.5 (d) e (e). A extração destes componentes está de acordo com a informação contida no periodograma da série, mostrado na Figura 3.3.

A partir do vetor 5, há uma diminuição nos valores singulares correspondentes aos autovetores 6, 7, e 8. Observando a Figura 3.5 (f), nota-se que o autovetor 6 é, provavelmente, um autovetor de tendência. A interpretação dos autovetores 7 e 8 não é tão clara. Observando o comportamento destes autovetores, na Figura 3.5 (g) e (h), pode-se dizer que estão relacionados a um componente sazonal de baixa frequência ou de tendência da série. Os autovetores 9 e 10 formam outro platô no gráfico e, portanto, devem representar um componente harmônico. Observando a Figura 3.5 (i) e (j) e analisando os valores dos picos e dos mínimos, notou-se que este par de autovetor representa um componente harmônico de período um pouco maior que 12, provavelmente resultado da mistura entre o componente harmônico de período igual a 12 com outros componentes da série. A partir do autovetor 11, os valores singulares diminuem suave e constantemente, e foram interpretados como sendo provenientes de ruídos na série e, por isso, não foram utilizados na etapa de reconstrução.

Hassani (2007) e Golyandina *et al.* (2001) destacam que a etapa da decomposição é bem sucedida apenas se conseguir decompor a série em componentes independentes entre si. Uma maneira de se medir a independência entre os componentes é considerar o coeficiente de correlação entre eles. O coeficiente de correlação, ρ , entre dois autovetores indica a força de relacionamento entre eles. Quanto maior for o módulo do coeficiente de correlação, ou seja, quanto mais próximo de 1, maior é a relação e, conseqüentemente, a dependência entre os componentes. Dois autovetores resultantes de um mesmo componente harmônico devem estar altamente correlacionados, e não se correlacionar significativamente com nenhum outro componente, evidenciando a separação do sinal.

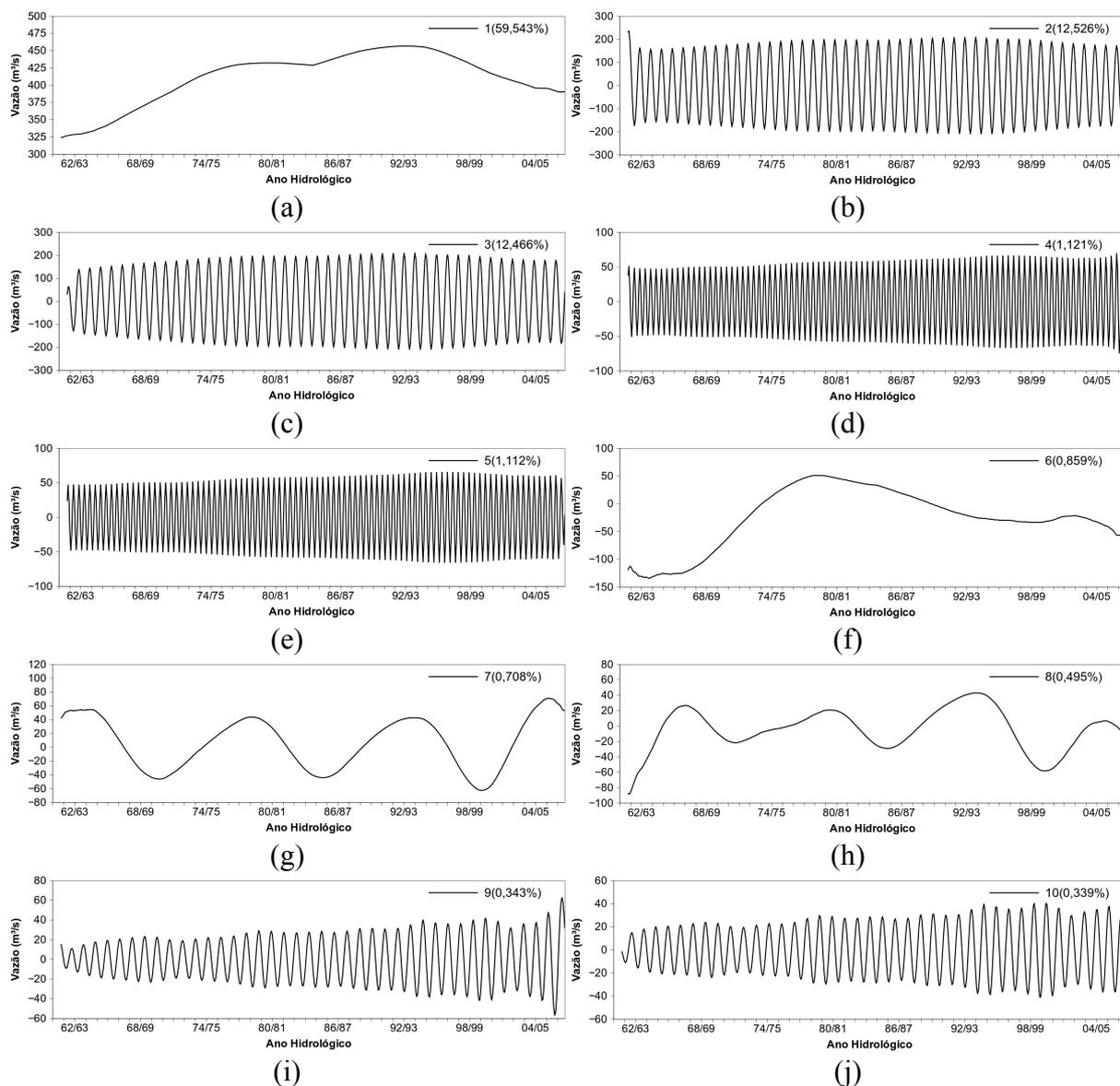


Figura 3.5 – Dez primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=276$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 6; (f) componente da tendência da série; (g) e (h) componente de baixa frequência, correlacionada com a tendência da série; (i) e (j) componente harmônico de período aproximadamente igual a 12.

Em função da mistura entre os componentes harmônicos, de tendência e os ruídos, foi difícil decidir quais autovetores deveriam ser incluídos na reconstrução da série. Do exposto, a inclusão dos autovetores 6, 7, 8, 9 e 10, bem como a definição se os autovetores 7 e 8 seriam representativos para a tendência ou se representariam um componente harmônico, se baseou também na visualização da matriz de correlação entre os componentes reconstruídos. Analisando a matriz de correlação, Figura 3.6, pode-se notar que os pares de autovetores originados de um mesmo componente harmônico estão altamente correlacionados, como os

autovetores 2 e 3 ($\rho = 1,000$) provenientes do componente harmônico de período igual a 12, os autovetores 4 e 5 ($\rho = 1,000$) provenientes do componente harmônico de período igual a 6 e os autovetores 9 e 10 ($\rho = 0,998$). O autovetor 1, que representa a tendência principal da série, não se correlaciona com nenhum outro autovetor o que também demonstra que a separação sinal-ruído foi realizada de forma eficiente. O autovetor 6, como dito anteriormente, também é um autovetor de tendência, e sua correlação com o autovetor 7 é bastante baixa ($\rho = 0,0770$). Os autovetores 7 e 8 estão correlacionados entre si, embora esta correlação não seja tão alta ($\rho = 0,8200$) quanto a correlação entre os outros autovetores harmônicos.

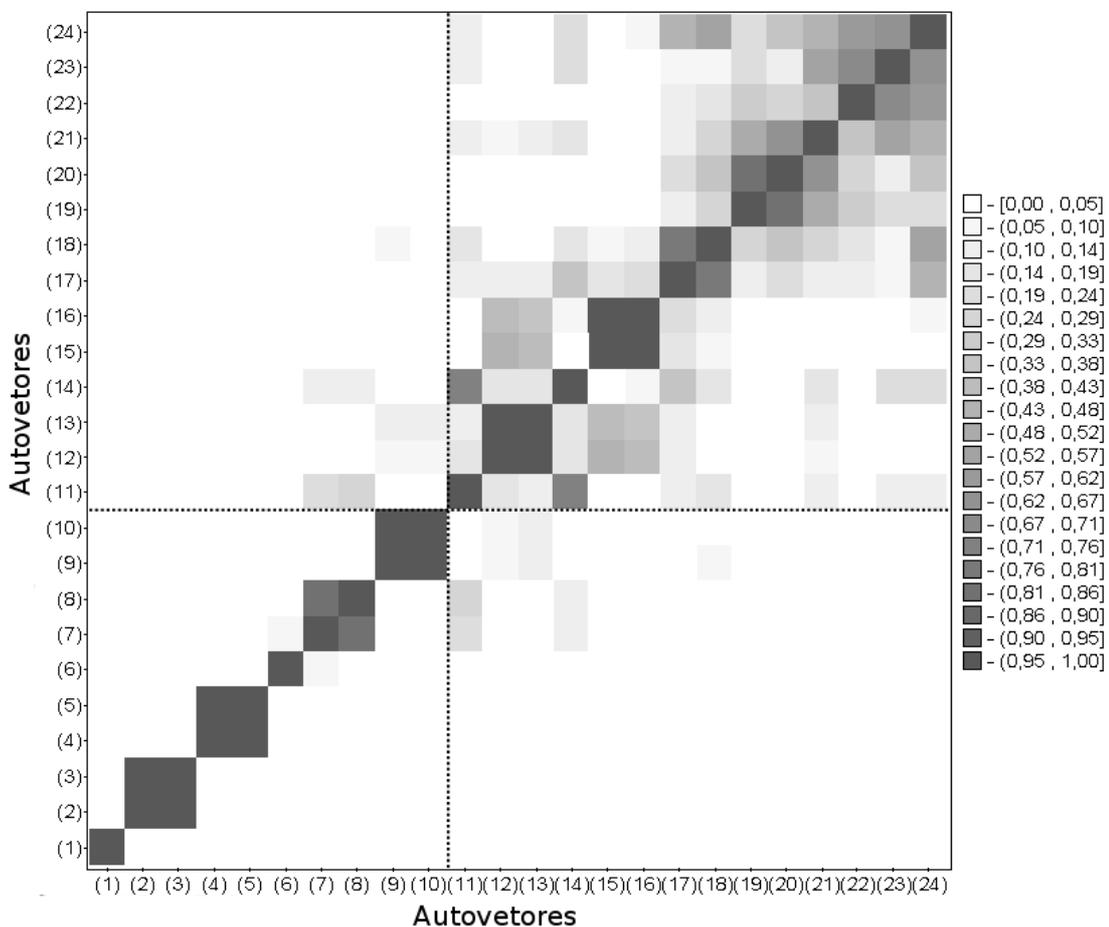


Figura 3.6 – Matriz de correlação entre os autovetores reconstruídos.

Como os valores singulares dos autovetores 7 e 8 não são muito próximos, se optou por incluir estes autovetores para a reconstrução da tendência da série. O autovetor 11 bem como os autovetores posteriores possuem correlações com vários outros autovetores, o que indica mistura entre os autovetores. Estes autovetores foram interpretados como sendo provenientes de ruídos na série temporal.

3.1.1.3 Reconstrução do Sinal da Série

O agrupamento dos autovetores para a reconstrução do sinal pode ser feito de várias maneiras. Se forem considerados a tendência (autovetores 1, 6, 7, e 8), os componentes harmônicos de frequência igual a 12 (autovetores 2, 3, 9 e 10), igual a 6 (autovetores 4 e 5) e os ruído (autovetores 11 a 277), se obtém quatro grupos de autovetores ($m = 4$). Se for considerado cada par de autovetores, se formam seis grupos ($m = 6$): 1, 2-3, 4-5, 6, 7-8, 9-10. Para a reconstrução final da série foram agrupados todos os autovetores referentes ao sinal e, além disso se reconstruiu cada elemento do sinal (harmônico e tendência) separadamente. A Figura 3.7 (a) demonstra a reconstrução da série utilizando apenas os autovetores 7 e 8, e tem como resultado uma série de baixa variação o que confirma a suposição de que estes autovetores são representativos para a tendência. Além disso, a tendência da série não é bem representada quando é reconstruída apenas com os autovetores 1 e 6, Figura 3.7 (b).

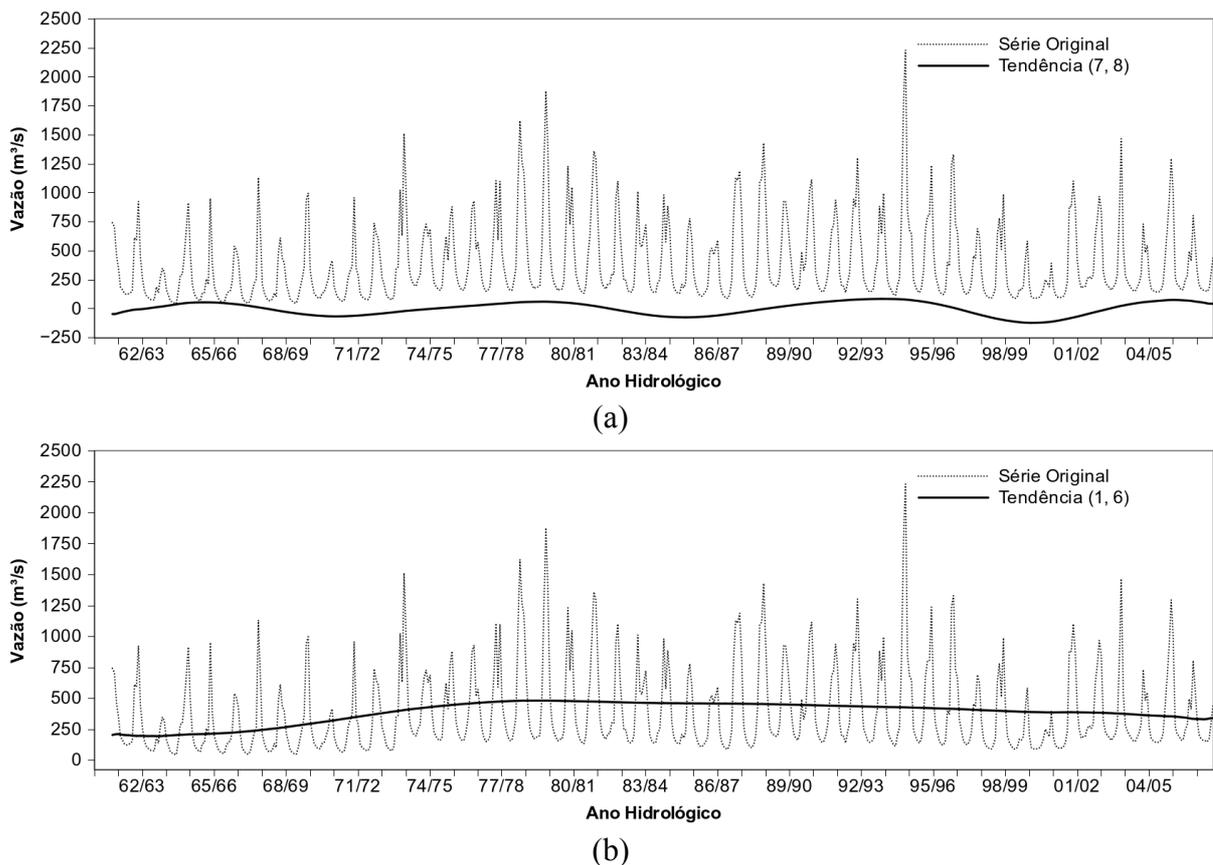
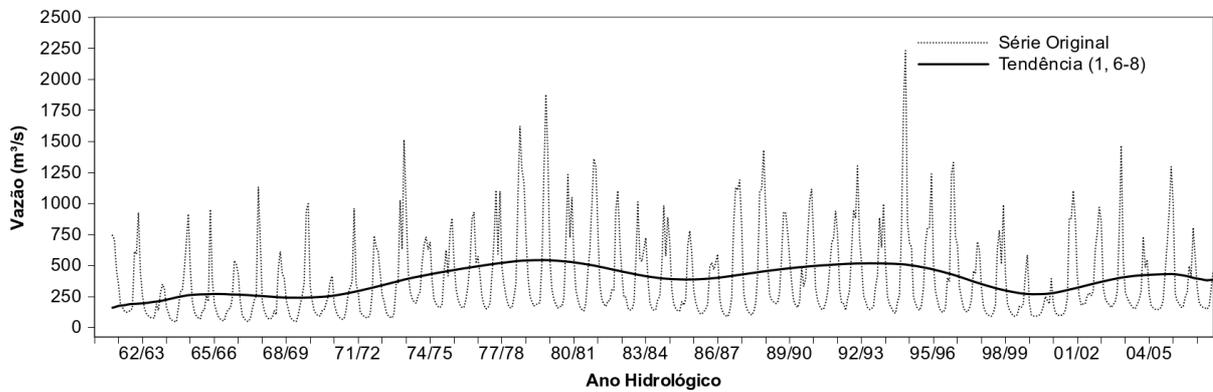


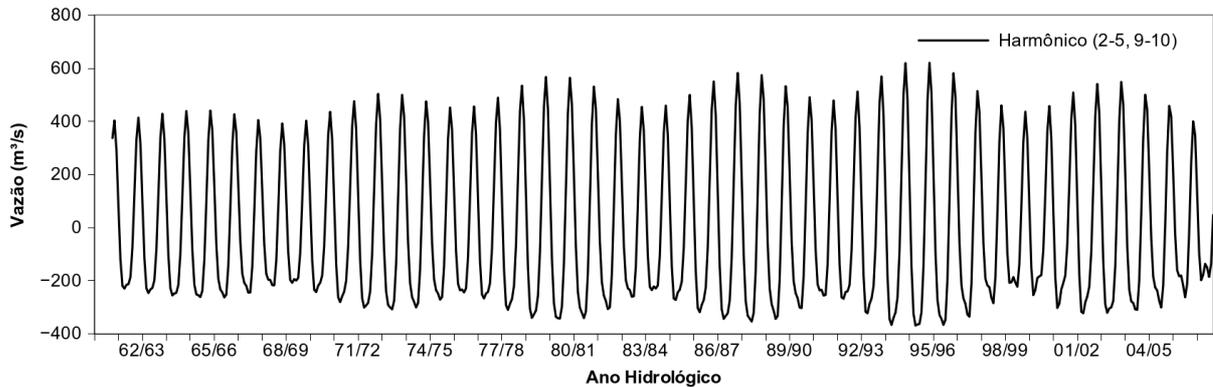
Figura 3.7 – Reconstrução da tendência da série original. (a) utilizando apenas os vetores 7 e 8 e (b) utilizando apenas os vetores 1 e 6.

Deste modo, a reconstrução do sinal da série considerou apenas três grupos ($m = 3$), sendo os autovetores 7 e 8 agrupados com os autovetores 1 e 6, ou seja, foram considerados autovetores representativos para a tendência da série. Além disso, os componentes harmônicos

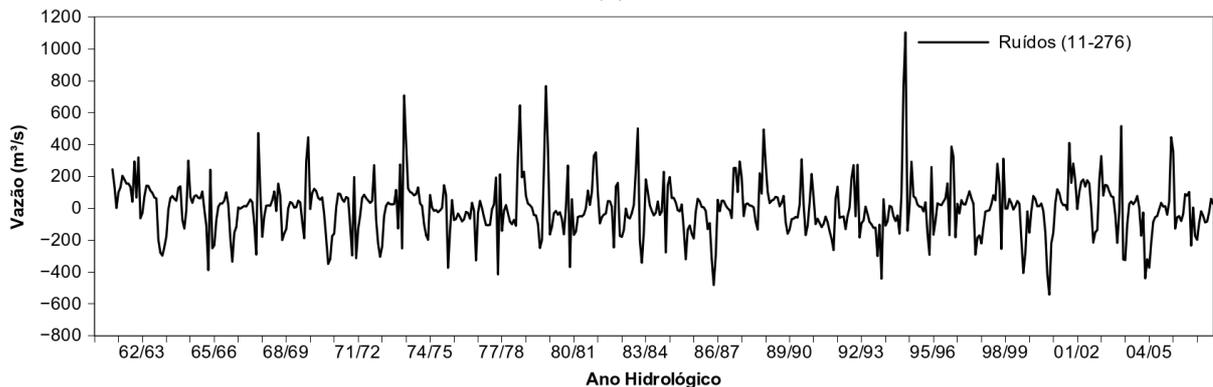
(autovetores 2, 3, 4, 6, 9 e 10) formaram um único grupo e os ruídos (autovetores 11 ao 277) formaram o último grupo. Assim, se pôde reconstruir os componentes de tendência, harmônicos e ruídos da série temporal, como pode ser observado nas Figuras 3.8 (a), (b) e (c), respectivamente.



(a)



(b)



(c)

Figura 3.8 – Reconstrução do sinal e dos ruídos da séries. (a) tendência geral da série, reconstruída utilizando os autovetores 1, 6, 7 e 8; (b) componente harmônico reconstruído utilizando os autovetores 2-5 e 9-10 e; (c) reconstrução dos ruídos utilizando os autovetores 11-276.

A tendência da série possui um comportamento não linear, e a reconstrução deste sinal considerando os autovetores 1, 6-8, Figura 3.8 (a), se mostrou bastante satisfatória, e melhor

ajustada à série, se comparado com as tendências obtidas a partir dos autovetores 1;6 e 7-8 separadamente. A reconstrução da tendência mostra uma variação a cada 12 anos, aproximadamente, no comportamento da série, embora seja necessário um número maior de dados para que se possa afirmar que esta variação é característica da série. Agrupando a tendência e a sazonalidade da série temporal (1-10), é possível reconstruir o sinal da série, como pode-se ver na Figura 3.9.

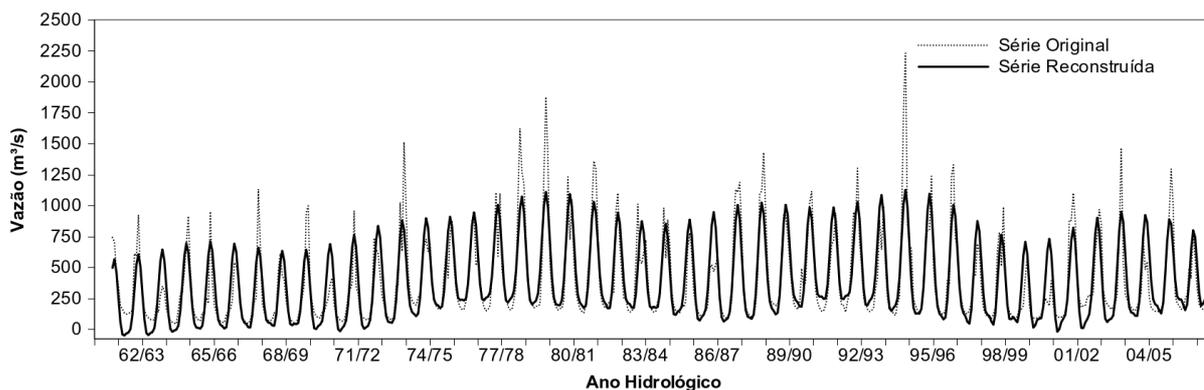


Figura 3.9 – Reconstrução do sinal da série (tendência e harmônicos), considerando os 10 primeiros autovetores extraídos na etapa da decomposição.

O maior resíduo encontrado na etapa de reconstrução foi 1.104,24 m³/s, correspondente ao mês de fevereiro de 1994. Embora boa parte do erro cometido possa ser atribuído aos ruídos presentes na série original, pode-se dizer que o erro na reconstrução também pode ser devido a eventos ambientais extremos e não previsíveis facilmente. O Erro Médio Absoluto (EMA) e o Erro Médio Relativo (EMR) foram iguais a 121,1 m³/s e 0,4% respectivamente e, além disso, a série original e a série reconstruída estão significativamente correlacionadas ($\rho=0,8564$; $t=38,8962$, $df=550$, $p\text{-value}<2,2\times 10^{-16}$). Os erros absolutos cometidos na etapa da reconstrução podem ser observados na Figura 3.10.

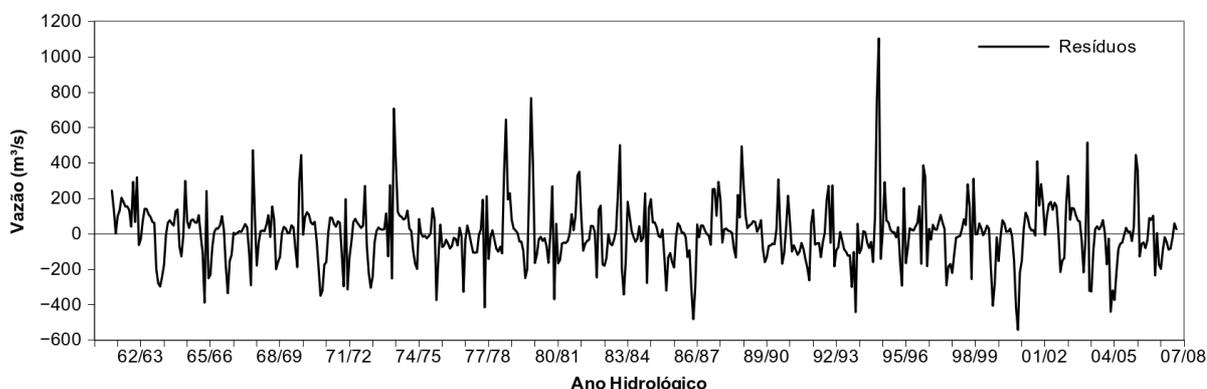


Figura 3.10 – Erros absolutos (resíduos) cometidos na etapa de reconstrução do sinal da série, considerando os autovetores 1 ao 10.

Levando em conta a magnitude dos erros e do coeficiente de correlação, pode-se concluir que a série reconstruída com os 10 primeiros autovetores extraídos representam bem o comportamento da série original. Além de revelar as variações sazonais, outro fato relevante no que diz respeito a robustez do método, é sua capacidade em revelar tendências cujo comportamento é não linear. Os resultados mostram que a análise espectral é uma técnica útil quando se quer extrair informação sobre o comportamento de séries temporais, revelando sua tendência e efeitos sazonais. A seguir é mostrado os resultados da análise espectral para uma série de dados de precipitação.

3.1.2 Precipitação Acumulada Mensal - Estação Vila Alegre (1359001)

A estação Vila Alegre possui 168 observações ($N=168$) mensais, correspondentes ao período de janeiro de 1993 a dezembro de 2006. O gráfico cartesiano dos valores da precipitação acumulada mensal (mm) da estação é mostrado na Figura 3.11. Analisando visualmente a série, nota-se um comportamento periódico, apresentando sazonalidade com período de 12 meses. Como nas séries de vazão média, os picos ocorrem aproximadamente nos finais e os valores mais baixos por volta do início de cada ano hidrológico. Nos períodos de maior precipitação, nota-se um comportamento irregular, e há a ocorrência de valores de precipitação iguais a zero, geralmente nos meses de junho, julho e/ou agosto caracterizando o período de seca na região. As maiores precipitações observadas na série foram 486 mm, em janeiro de 1995, 480,6 mm, em fevereiro de 2006, 440,8 mm, em fevereiro de 1994 e 432,3, em março de 1993 e fevereiro de 2001.

3.1.2.1 Decomposição da Série Original

Levando em consideração o que já foi discutido para a decomposição da série de vazão, e considerando o tamanho desta série temporal, $N=168$, a decomposição foi feita fixando o tamanho da janela (L) em 84. Com isso, na etapa de decomposição obteve-se 85 vetores defasados ($K=85$). Novamente, o gráfico das covariâncias médias entre os vetores defasados (Figura 3.12), confirma a principal periodicidade da série proveniente do ciclo sazonal de 12 meses.

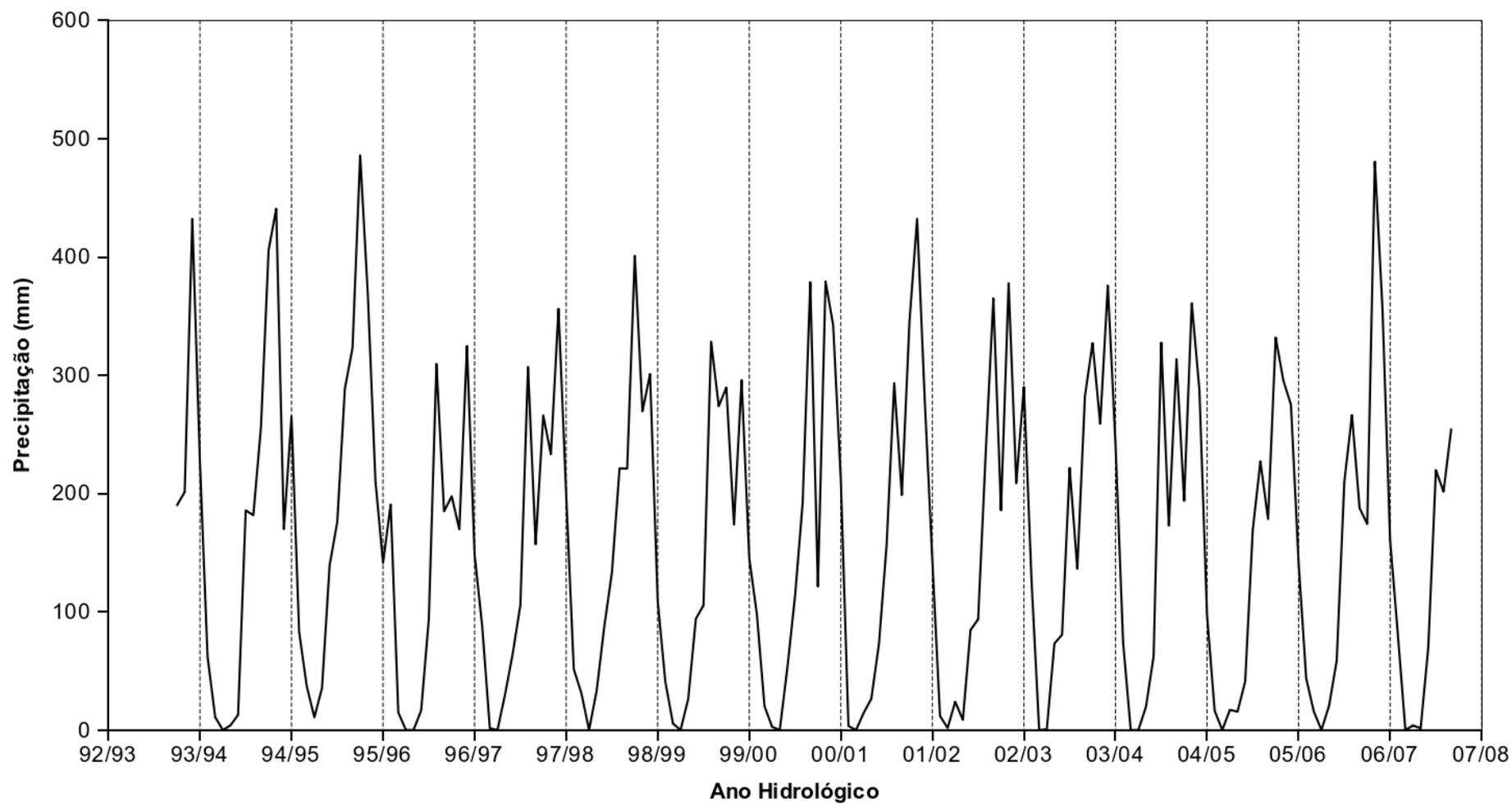


Figura 3.11 – Série temporal de precipitação, registrada na estação de Vila Alegre, durante o período de janeiro de 1993 a dezembro de 2006.

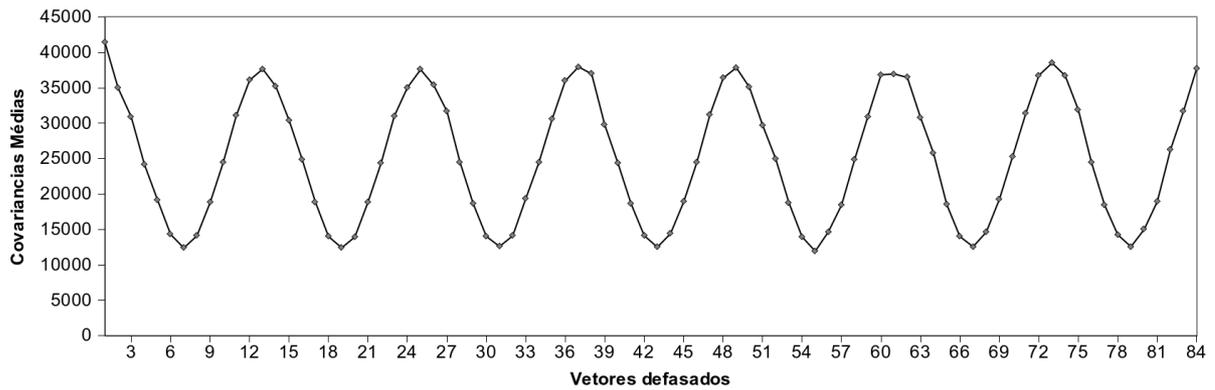


Figura 3.12 – Covariâncias médias dos vetores defasados.

O periodograma médio da série, mostrado na Figura 3.13, mostra nitidamente a separação do componente harmônico de período igual a 12 meses. Outros componentes harmônicos, de periodicidade igual a 4 e 6 meses, também podem ser observados, embora a separação não seja boa, uma vez que o poder espectral é baixo.

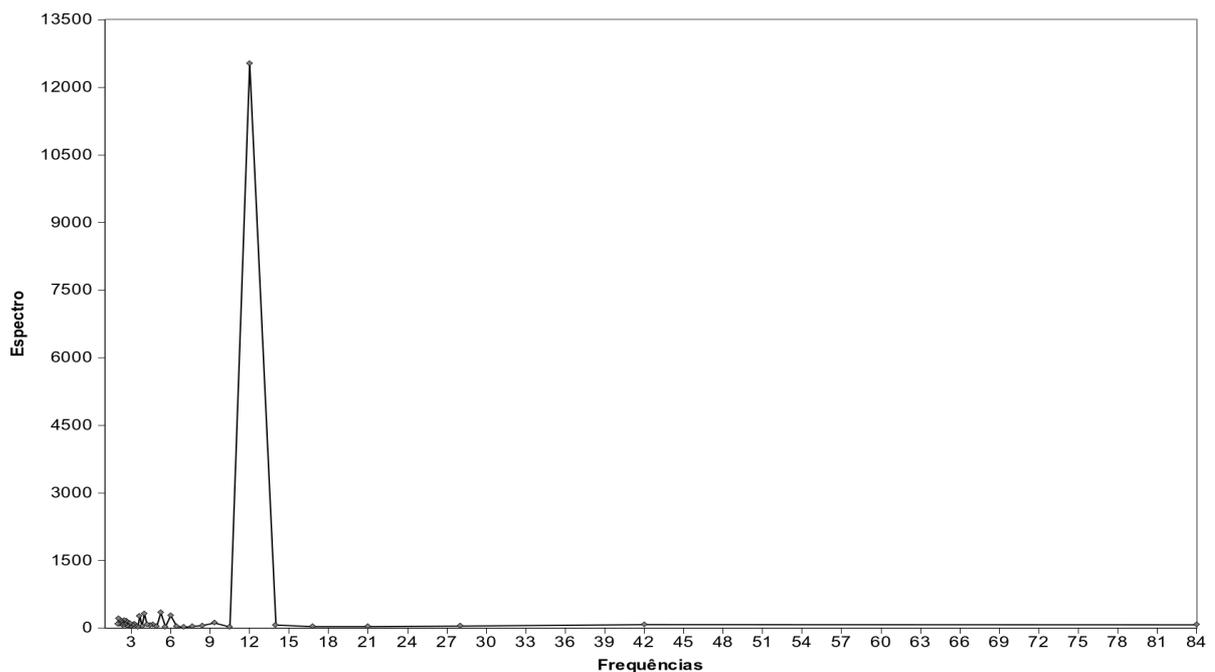


Figura 3.13 – Periodograma médio da série temporal.

3.1.2.2 Separação do Sinal e Ruído

Pode-se observar na Figura 3.14 os valores singulares dos autovetores. O autovetor 1, mais separado dos demais autovetores, representa a tendência principal da série. O componente harmônico anual resultou em um par de autovetores com valores singulares próximos (autovetores 2 e 3), formando o primeiro platô no gráfico. A partir daí, foi difícil identificar autovetores correspondentes ao sinal da série, uma vez que os seus valores

singulares estão bastante próximos.

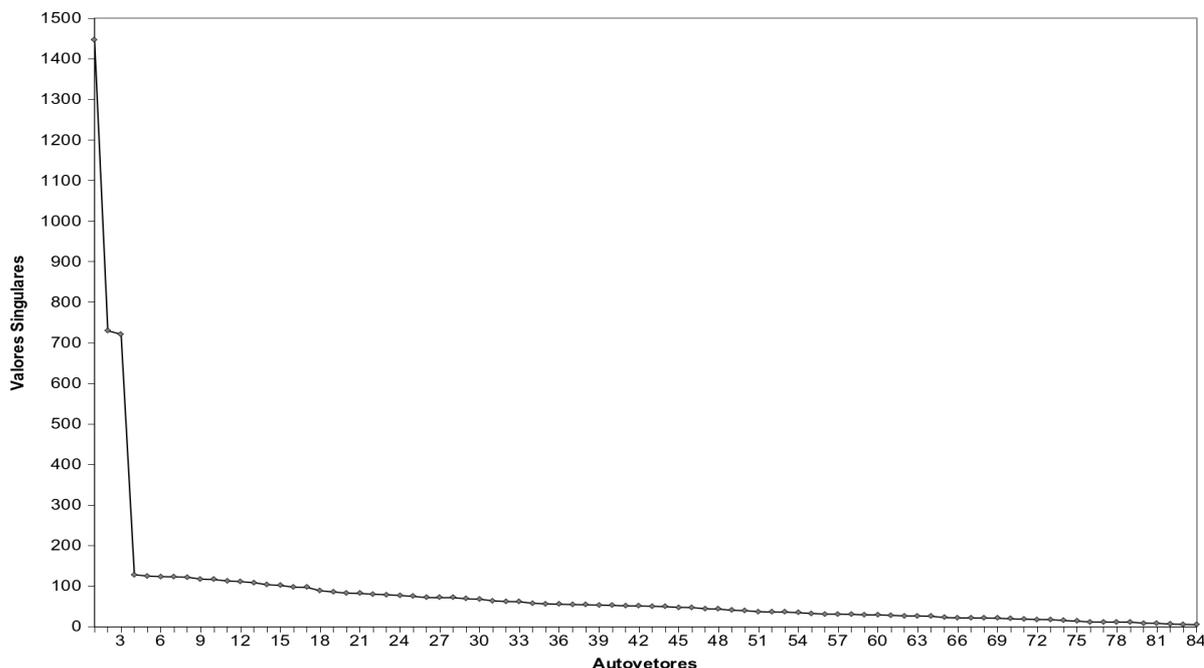


Figura 3.14 – Valores singulares dos autovetores defasados.

Observando a representação 2D dos autovetores na Figura 3.15, se pode dizer que o autovetor 1 representa a tendência principal da série, enquanto os autovetores 2 e 3 representam a sazonalidade anual da série (período igual a 12).

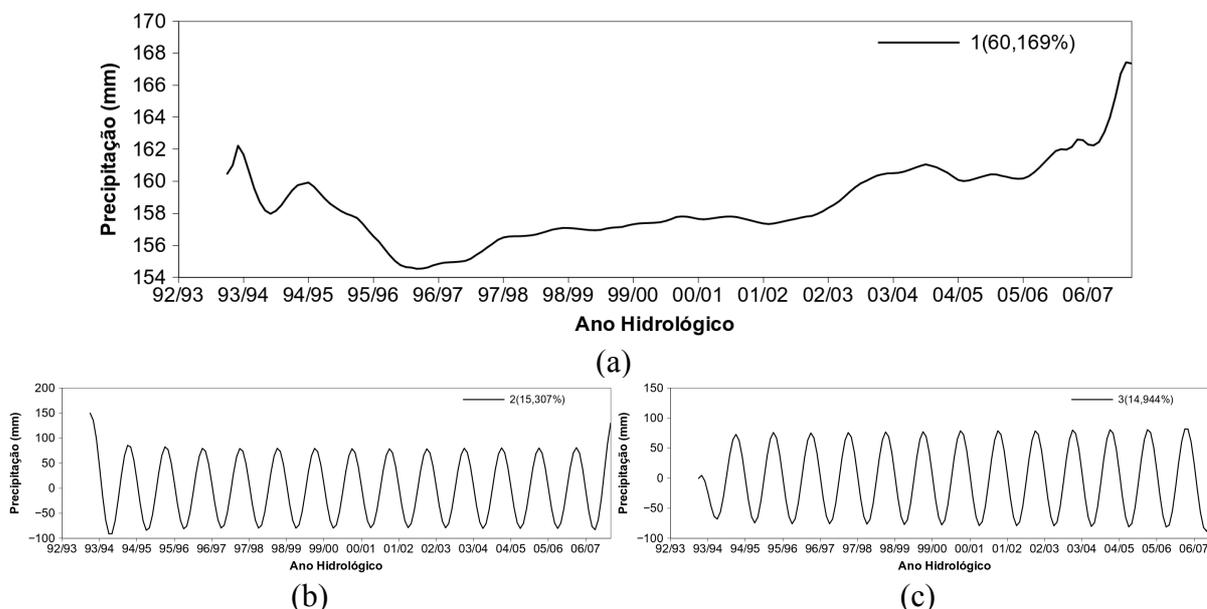


Figura 3.15 – Três primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=84$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12.

Analisando a matriz de correlação, Figura 3.16, pode-se notar que os autovetores 2 e 3, originados do componente harmônico de período igual a 12, apresentaram alta correlação

($\rho=1,000$), e não se correlacionam com nenhum outro autovetor, o que evidencia sua boa extração. O autovetor 1, que representa a tendência principal da série, não se correlaciona com nenhum outro autovetor, também evidenciando a boa extração da tendência da série. O autovetor 4 bem como os autovetores posteriores possuem correlações com vários outros autovetores não subsequentes e foram considerados ruídos.

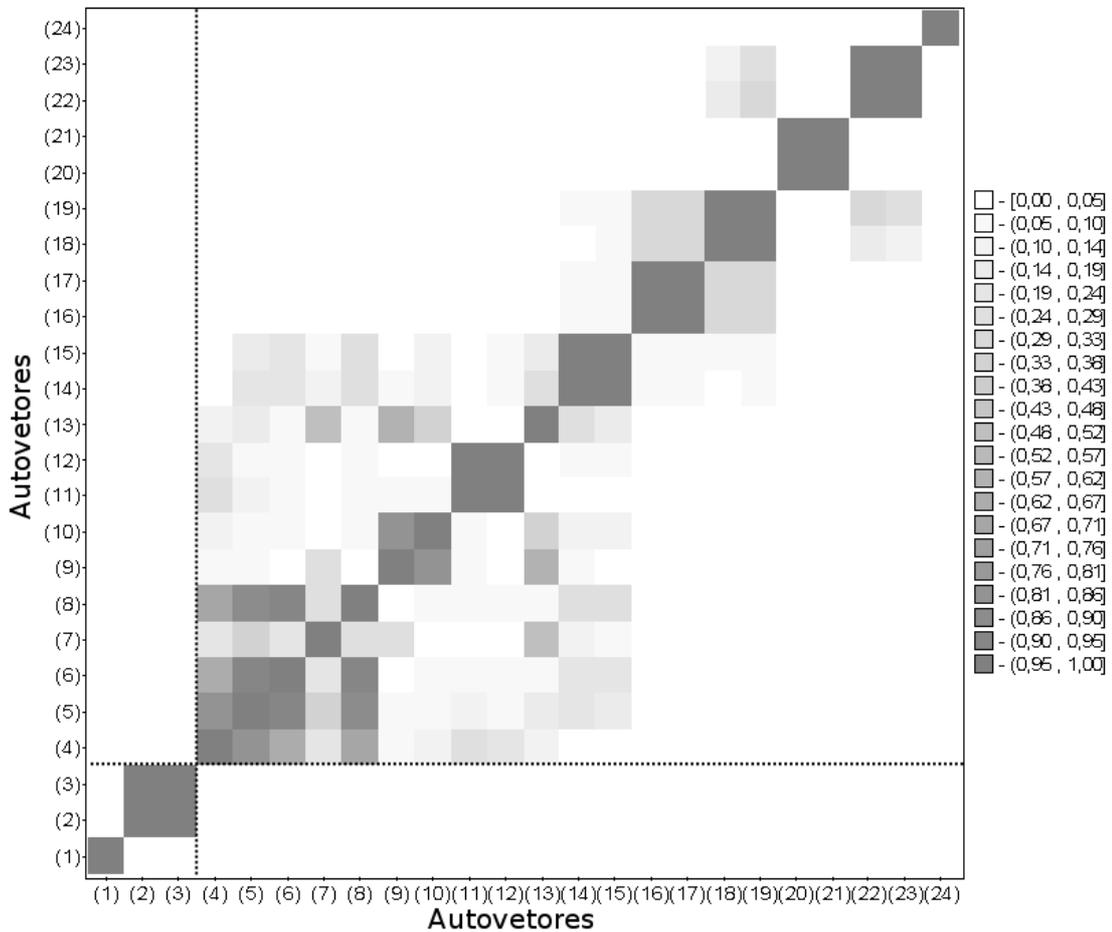
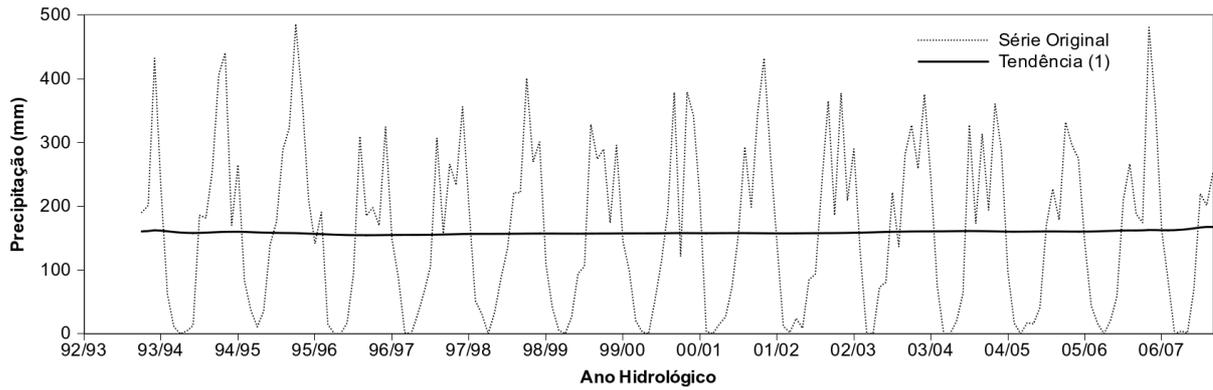


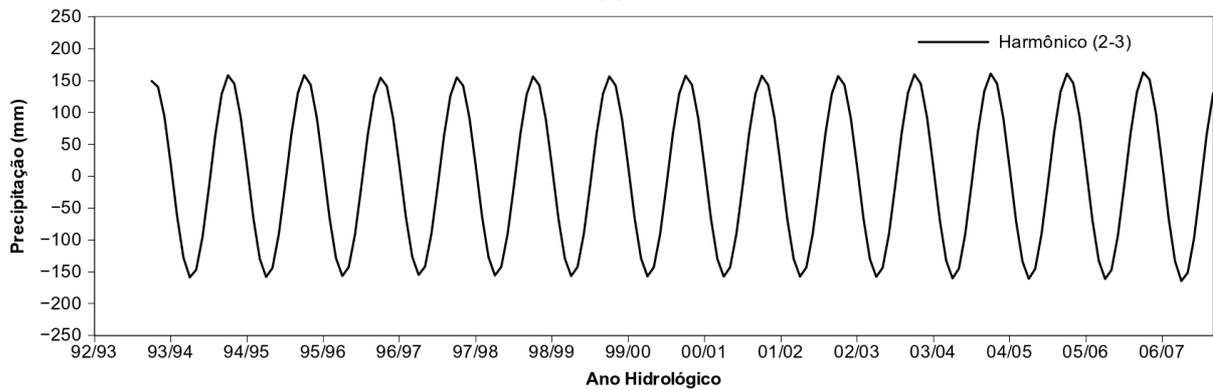
Figura 3.16 – Matriz de correlação entre os autovetores reconstruídos.

3.1.2.3 Reconstrução do Sinal da Série

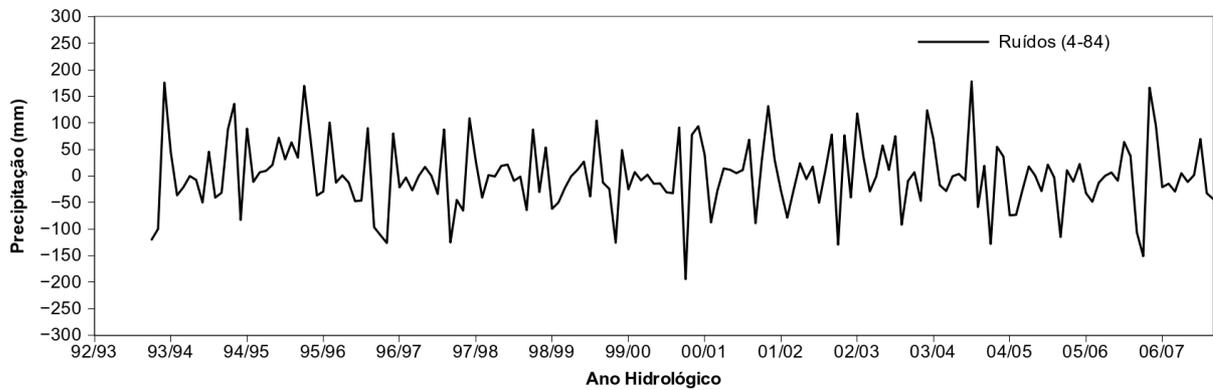
Neste caso, como apenas três autovetores foram representativos para o sinal da série, o agrupamento pode ser feito apenas de duas maneiras. O autovetor 1 pode ficar junto ou separado dos autovetores 2 e 3 na etapa da reconstrução. A tendência só pode ser reconstruída a partir do autovetor 1, e a sazonalidade a partir dos autovetores 2 e 3, enquanto os ruídos podem ser reconstruídos com os demais autovetores. A reconstrução da tendência da série pode ser vista na Figura 3.17 (a), enquanto a sazonalidade e os ruídos nas Figuras 3.17 (b) e (c), respectivamente.



(a)



(b)



(c)

Figura 3.17 – Reconstrução do sinal e dos ruídos da série. (a) tendência geral da série, reconstruída utilizando o autovetor 1; (b) componente harmônico reconstruído utilizando os autovetores 2-3 e; (c) reconstrução dos ruídos utilizando os autovetores 4-84.

A tendência da série é linear e praticamente estacionária. Apenas é possível observar que a precipitação possui uma leve tendência de aumento a partir de $N=36$ (janeiro de 1996), como mostra a Figura 3.15 (a). A variação de amplitude no harmônico da série também é pequena. Agrupando a tendência e a sazonalidade da série temporal (autovetores 1, 2 e 3), foi possível reconstruir o sinal da série, como pode-se ver na Figura 3.18.

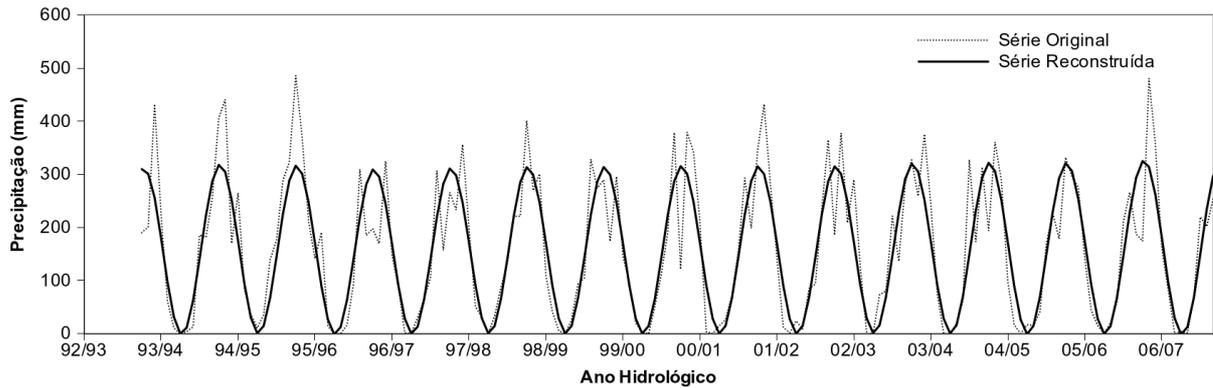


Figura 3.18 – Reconstrução do sinal da série (tendência e harmônicos), considerando os 3 primeiros autovetores extraídos na etapa da decomposição.

Para os dados de precipitação da estação Vila Alegre, os resíduos da reconstrução variaram de -194 a +178,2 mm. O EMA e o EMR foram iguais a 47,8 mm e 0,9% respectivamente e, além disso, a série original e a série reconstruída estão significativamente correlacionadas ($\rho = 0,8672$; $t = 22,4349$, $df = 166$, $p\text{-value} < 2,2 \times 10^{-16}$). Os resíduos da etapa de reconstrução podem ser observados na Figura 3.19.

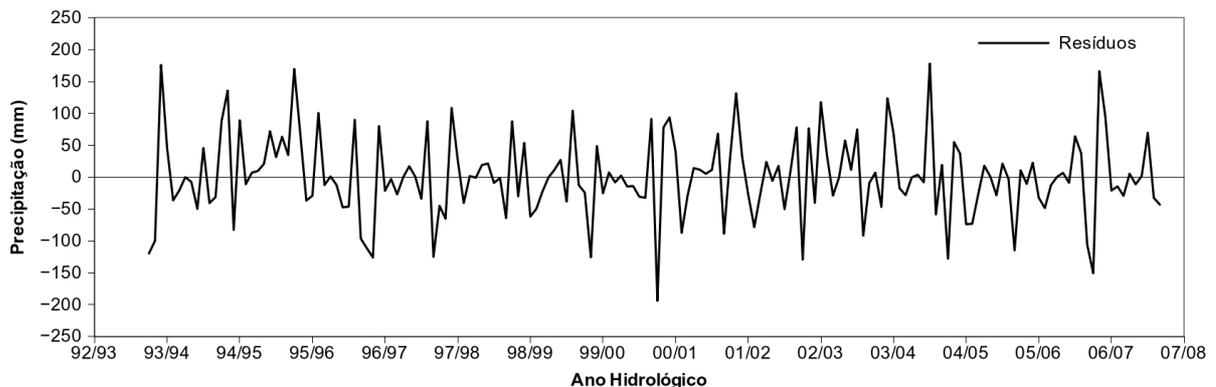


Figura 3.19 – Erros absolutos (resíduos) cometidos na etapa de reconstrução do sinal da série, considerando os autovetores 1 ao 3.

Assim como para a variável vazão média, a reconstrução da série utilizando os três primeiros autovetores foi bastante satisfatória, considerando os baixos erros relativos e a correlação entre as séries. Também foi possível extrair satisfatoriamente a tendência principal e os componentes sazonais da série. Os resultados apresentados até aqui demonstram que o método de análise espectral pode ser utilizado para análise e interpretação de séries temporais das variáveis vazão média mensal e precipitação acumulada mensal. A principal vantagem do método é sua capacidade de extrair informações mesmo quando a série apresenta comportamento não linear. Além disso o método é capaz de revelar os comportamentos sazonais da série. Marques *et al.* (2006), aplicaram o método SSA para previsão em séries

temporais anuais de precipitação, e também conseguiram uma boa extração de componentes, embora as séries possuíssem comportamento irregular.

Os resultados mostrados a seguir enfocam outra aplicação do método. Além da análise das séries, o método também pode ser utilizado para realizar o preenchimento das falhas de dados presentes nas séries.

3.2 Preenchimento de Falhas de Dados em Séries Temporais

Há um grande interesse da comunidade científica em observações meteorológicas para o desenvolvimento de estudos ambientais, de variabilidade climática e de previsão do tempo. Entretanto, é grande a dificuldade para obtenção de dados meteorológicos de séries temporais longas e sem falhas de dados. Os motivos para que as falhas ocorram dependem da variável e do sistema de monitoramento em questão mas, como destacado por Broersen *et al.* (2004) e Broersen (2006), geralmente estão associadas a quebra de equipamentos, presença de *outliers* ou entrada errada de dados e, especificamente no caso da meteorologia, dados astronômicos, ou observações via satélite, as condições do tempo podem dificultar a amostragem, gerando as falhas. Além destes fatores, pode-se citar, ainda, a falta de recursos para a medição, como responsável pelo surgimento das falhas nos conjuntos de dados.

Os modelos matemáticos de análise de séries temporais partem do pressuposto de que as observações foram realizadas sempre com a mesma frequência, o que não ocorre em função das falhas existentes nas séries. Estas falhas de dados, ou valores faltantes, impedem e/ou prejudicam a análise estatística, pois reduzem a precisão dos resultados gerados. Por estes motivos, vários métodos foram propostos para que o preenchimento das falhas pudesse ser feito e, especificamente no caso do método SSA, a partir da análise de séries estacionárias, Schoellhamer (2001) descreveu um método de preenchimento. Outro método, delineado por Kondrashov *et al.* (2005), pode ser aplicado no preenchimento de séries temporais com presença de ruído vermelho⁵. O método implementado no *software* CatMV[®], é uma extensão do método Caterpillar-SSA, descrito detalhadamente por Golyandina *et al.* (2001), e utilizado nos itens anteriores para a análise das séries temporais.

A seguir, é mostrada a capacidade do método de SSA no preenchimento de falhas de dados de séries temporais. São mostrados os resultados do preenchimento de uma série

⁵ Também chamado de ruído marrom ou browniano, refere-se a qualquer sistema onde a densidade de potência diminui com o aumento da frequência.

temporal da variável vazão média mensal e da variável precipitação acumulada mensal.

3.2.1 Vazão Média Mensal - Estação Porto Cercado (6634000)

Para demonstrar a capacidade do método de SSA para o preenchimento de falhas em séries de vazão média mensal, foi utilizado os dados da estação Porto Cercado. A estação Porto Cercado possui cerca de 15% de dados faltantes. O número de observações, N , que deveriam estar presentes na série temporal caso não tivesse ocorrido nenhuma falha no monitoramento, é 439. Entretanto, existem apenas 374 valores observados.

A Figura 3.20 mostra a série histórica de vazão média mensal da estação Porto Cercado, de dezembro de 1969 a junho de 2005, bem como as falhas de dados, representadas pelas descontinuidades no gráfico. O ano hidrológico 88/89 possui apenas dois valores de vazão observada, e nos anos hidrológicos 89/90 e 90/91 não foi realizada nenhuma leitura de vazão, gerando a maior lacuna da série temporal. A série possui, ainda, outras falhas de dados nos anos hidrológicos 91/92, 92/93, 93/94, 94/95, 96/97, 97/98, 99/00, 01/02, 02/03 e 03/04. Assim como a série da estação Cuiabá, a série da estação Porto Cercado apresenta um comportamento periódico com valores elevados no final, e valores menores no início de cada ano hidrológico em função dos períodos de cheia (outubro a março) e estiagem (abril a setembro), respectivamente. O máximo valor registrado para a vazão foi de $693 \text{ m}^3/\text{s}$ no ano hidrológico 78/79. É importante notar que o maior pico de vazão desta série coincide com o segundo maior pico da série da estação Cuiabá. Além disso, a falha no final do ano hidrológico 93/94 coincide com a maior vazão observada na série da estação Cuiabá (ver Figura 3.1). O menor valor observado ocorreu no ano hidrológico 70/71, onde se registrou uma vazão média de $77,4 \text{ m}^3/\text{s}$, valor consideravelmente maior que a menor vazão registrada na série da estação Cuiabá. De maneira geral, as mínimas vazões registradas na estação Porto Cercado são maiores que as mínimas registradas, no mesmo mês, na estação Cuiabá. Porém, em alguns casos, como nos anos hidrológicos 73/74, 74/75, 75/76, 76/77 dentre outros, as vazões mínimas registradas na estação Porto Cercado foram menores que as registradas na estação Cuiabá.

A seguir é mostrado os resultados da decomposição da série original em autovetores defasados. Será feita a escolha dos autovetores que contém as informações sobre o comportamento da série e posteriormente a reconstrução da série temporal com preenchimento simultâneo das falhas de dados.

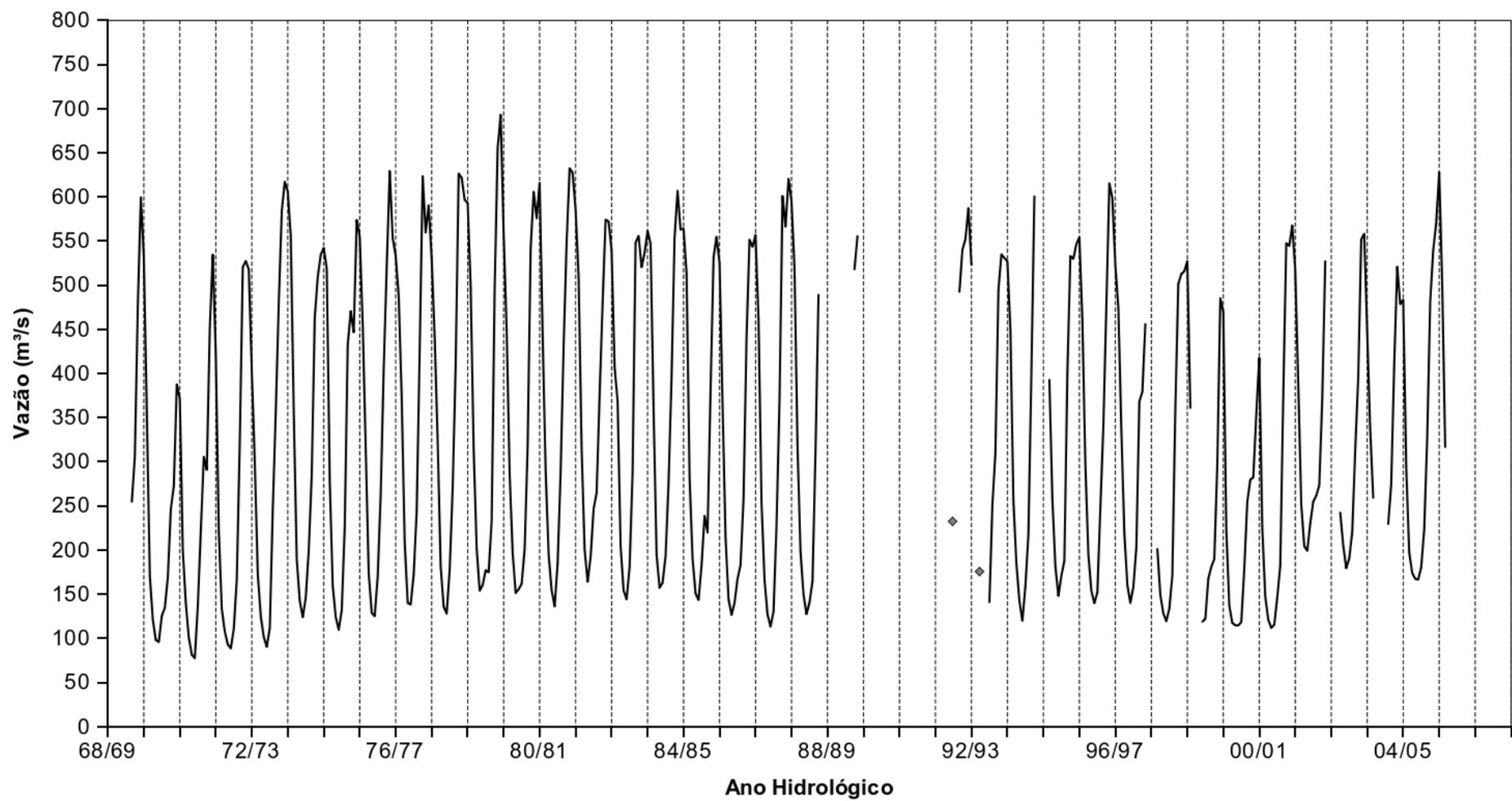


Figura 3.20 – Série temporal das vazões médias obtidas no rio Cuiabá, na estação de Porto Cercado, durante o período de dezembro de 1969 a junho de 2005.

3.2.1.1 Decomposição da Série Original

A periodicidade principal desta série também foi considerada igual a 12 em função da variação anual decorrente dos períodos sazonais já mencionados. A tendência principal da série não pôde ser facilmente observada através do gráfico da cartesiano. Como a série apresenta uma periodicidade anual e $N=439$, o tamanho da janela (L) foi fixado em 216. Optou-se pelo maior L possível e o mais próximo da metade da série ($N/2$), uma vez que a quantidade de falhas é relativamente pequena. O limiar escolhido foi coincidente com o número de falhas, o que resultou na maior quantidade possível de vetores defasados completos para o tamanho de janela dado. Assim, obteve-se 224 vetores defasados (K) completos na etapa de decomposição.

3.2.1.2 Separação do Sinal e Ruído

A interpretação do gráfico dos valores singulares dos autovetores, Figura 3.21, é semelhante ao já descrito na análise das séries históricas da estação Cuiabá e Vila Alegre. O autovetor 1, mais separado dos demais autovetores, representa a tendência principal da série. A confirmação desta suposição foi feita observando a representação 2D do autovetor 1, Figura 3.22 (a).



Figura 3.21 – Valores singulares dos autovetores defasados.

Ainda pela Figura 3.21, o componente harmônico anual resultou em um par de autovetores com valores singulares próximos (autovetores 2 e 3), formando o primeiro platô

no gráfico. Embora os autovetores 4 e 5 não estejam bem separados dos demais autovetores, também formam um platô no gráfico e, provavelmente, representam um componente harmônico da série. A partir do autovetor 6, foi difícil identificar autovetores correspondentes ao sinal da série, uma vez que os seus valores singulares estão bastante próximos.

Observando a representação 2D dos autovetores 2 e 3, Figura 3.22 (b) e (c) se pode dizer que estes autovetores representam a sazonalidade anual da série (período igual a 12). Os autovetores 4 e 5 representam o componente harmônico de periodicidade igual a 18 meses. Porém, há uma nítida diminuição na amplitude dos autovetores, no início da série - Figura 3.22 (d) e (e).

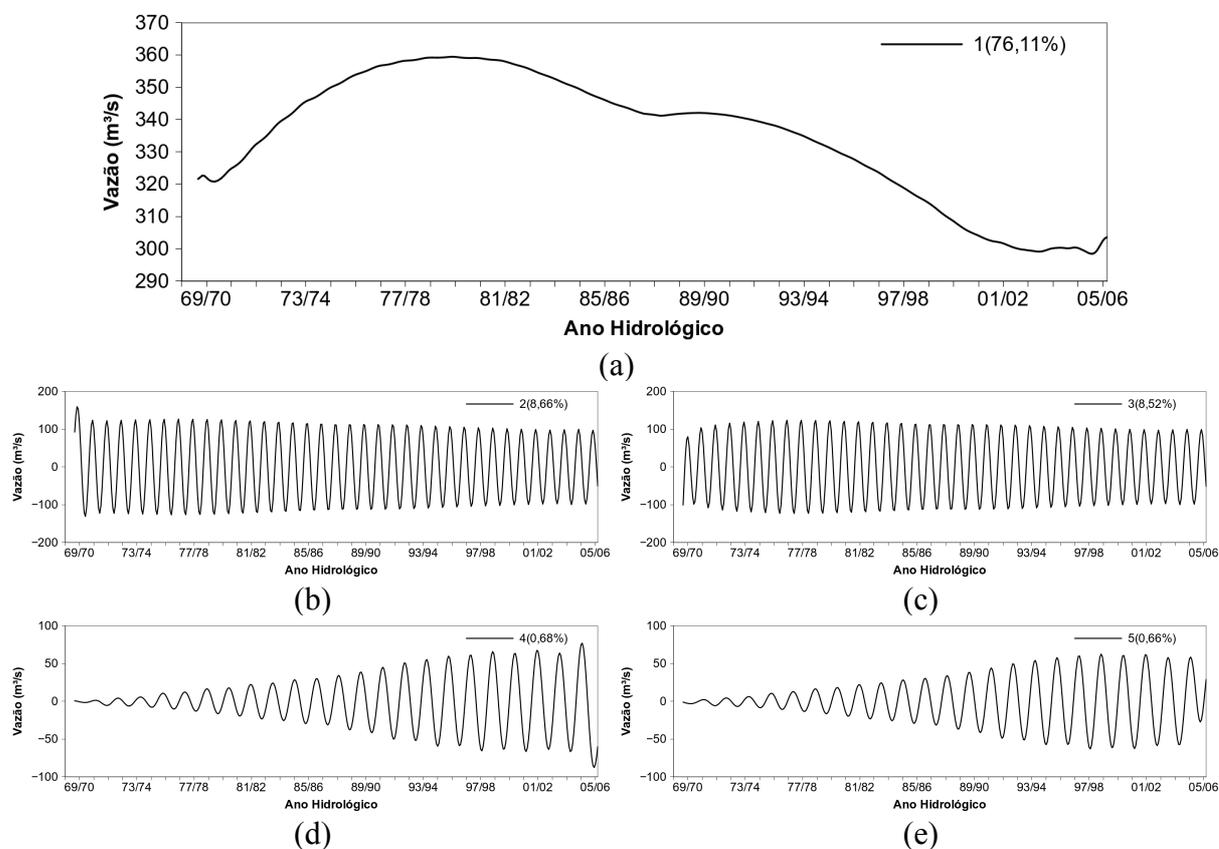


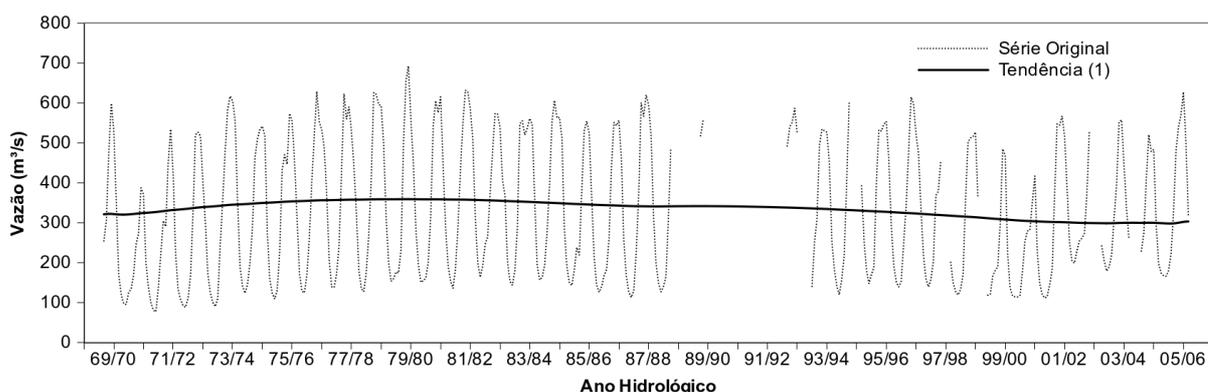
Figura 3.22 – Cinco primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=216$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 18.

Na série temporal da estação Cuiabá, foi extraído um componente harmônico de período igual a 6, o que não ocorreu nesta série. Não obstante, o componente harmônico de período igual a 18 não foi extraído na série da estação Cuiabá, o que pode indicar comportamentos diferenciados da vazão nestas duas estações ou problemas no processo de medição da variável. De qualquer maneira, ressalta-se que o regime de escoamento na região de Cuiabá é diferente do regime de escoamento da região pantaneira, onde situa-se a estação

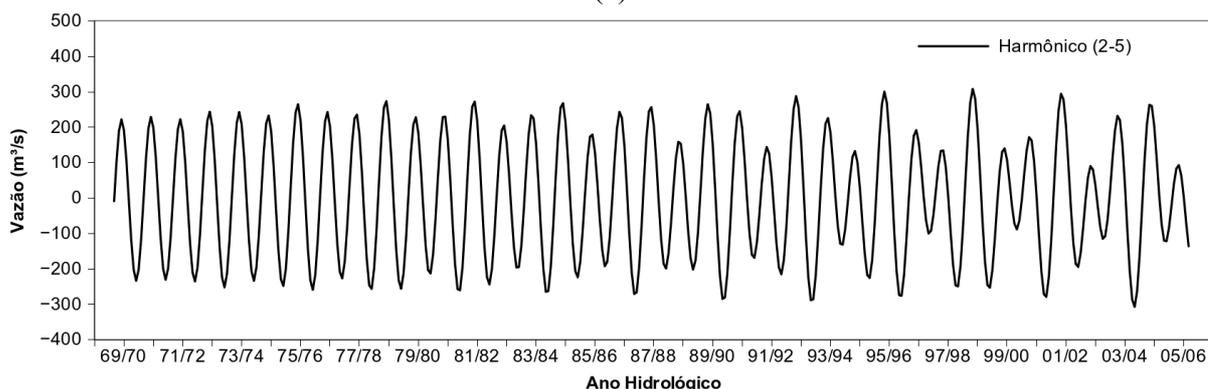
Porto Cercado. Enquanto o componente harmônico de 12 meses pode ser facilmente atribuído ao ciclo de estiagem e cheia decorrente das características climáticas da região, o componente harmônico de 18 meses requer análises mais específicas para sua devida interpretação.

3.2.1.3 Reconstrução do Sinal da Série

Para a reconstrução desta série pode-se considerar a tendência (autovetor 1), os componentes harmônicos de frequência igual a 12 (autovetores 2 e 3), igual a 18 (autovetores 4 e 5) e os ruídos (autovetores 6 a 224), obtendo-se quatro grupos de autovetores ($m = 4$). Se os autovetores harmônicos forem todos agrupados, se obtém três grupos ($m = 3$): 1, 2-5, 6-224. A Figura 3.23 (a), mostra a reconstrução da série utilizando apenas o autovetor 1, ou seja, mostra a tendência principal da série temporal, enquanto a Figura 3.23 (b) mostra a série reconstruída agrupando os autovetores 2, 3, 4 e 5, que representam os harmônicos da série.



(a)



(b)

Figura 3.23 – Séries reconstruídas através do método da projeção. (a) reconstrução da tendência da série utilizando o autovetor 1, sobreposta a série original; (b) reconstrução do componente harmônico da série com o agrupamento dos autovetores 2, 3, 4 e 5.

Tanto a tendência quanto o componente harmônico da série foram reconstruídos mesmo onde a série temporal apresentava falhas, demonstrando a capacidade do método em estimar os dados faltantes. A partir do agrupamento de todos os autovetores representativos do

sinal (tendência e harmônicos) foi possível reconstruir o sinal da série temporal original com preenchimento simultâneo dos dados faltantes, como pode ser observado na Figura 3.24 (a). Os maiores resíduos da etapa de reconstrução foram $-248,60 \text{ m}^3/\text{s}$ e $217,84 \text{ m}^3/\text{s}$ cometidos na reconstrução dos valores referentes a fevereiro de 1971 e abril de 2005, respectivamente. O resíduos da etapa de reconstrução são mostrados na Figura 3.24 (b).

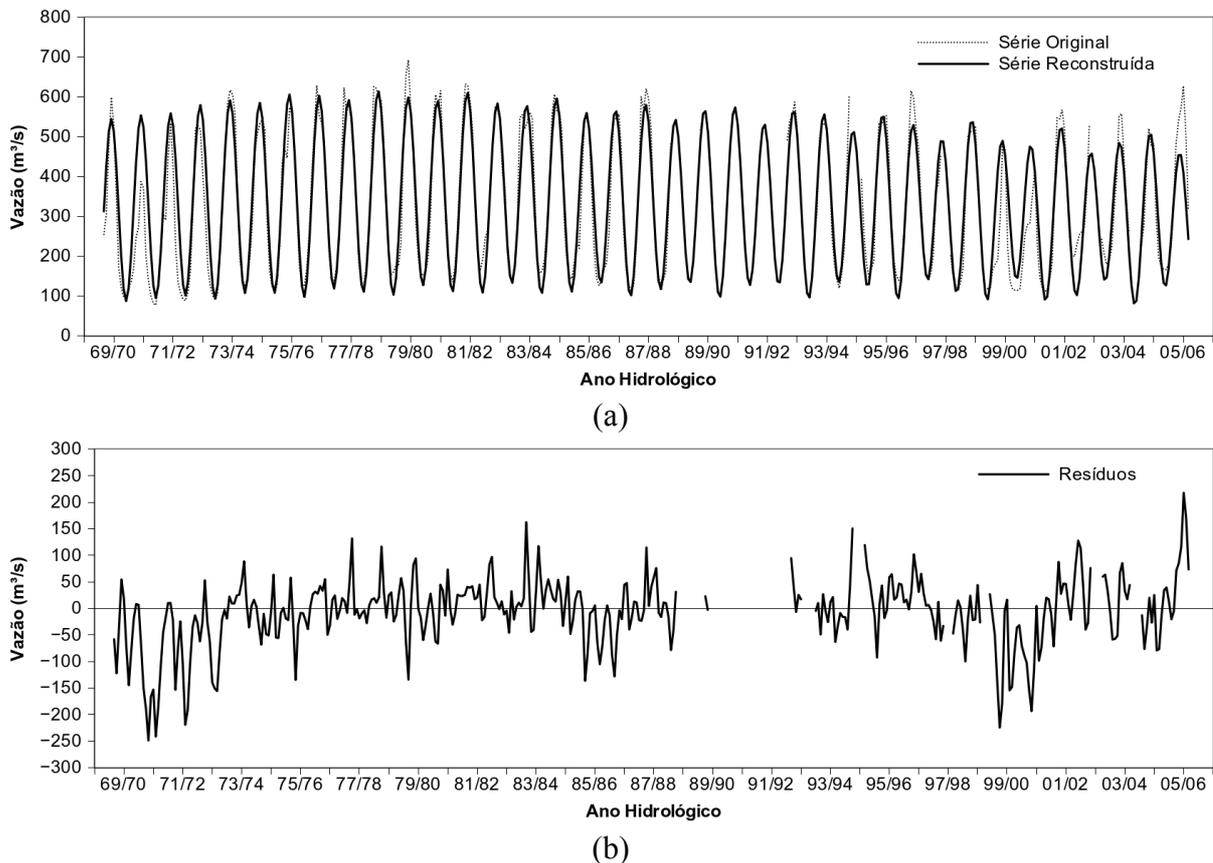


Figura 3.24 – Sinal reconstruído da série e resíduos da reconstrução. (a) série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5), sobreposta a série temporal original; (b) erros absolutos (resíduos) cometidos na etapa da reconstrução.

Neste caso, o EMA e o EMR foram iguais a $46,7 \text{ m}^3/\text{s}$ e $0,2\%$ respectivamente. A série original e a série reconstruída também apresentaram uma alta correlação ($\rho = 0,9238$; $t = 46,5405$, $df = 372$, $p\text{-value} < 2,2 \times 10^{-16}$) e, com base nisso, pode-se dizer que a reconstrução da série apenas com os 5 primeiros autovetores extraídos foi bastante satisfatória, e representativa para a série original, além do preenchimento das falhas ter sido coerente com as características de sazonalidade apresentadas pela série original.

3.2.2 Precipitação Acumulada Mensal - Estação Santa Rosa (857000)

No caso da estação Santa Rosa, a quantidade de observações que deveriam estar presentes no conjunto de dados seriam 168 ($N=168$), correspondentes ao período de janeiro de 1993 a dezembro de 2006. Entretanto, a série possui apenas 159 valores observados, com falhas, nos anos hidrológicos de 93/94 (3 falhas), 94/95 (1 falha) e no ano hidrológico 02/03 (5 falhas). A Figura 3.25 mostra a série temporal da variável precipitação da estação Santa Rosa de janeiro de 1993 a dezembro de 2006. As falhas de dados são representadas pelas descontinuidades da linha que representa as observações.

Assim como observado nas séries de vazão, a série de precipitação apresenta um comportamento periódico, com valores mais elevados no final, e valores menores no início de cada ano hidrológico, em função dos períodos de cheia e estiagem, respectivamente. Observa-se valores acumulados acima de 500mm em fevereiro de 1994 (ano hidrológico 93/94), em dezembro de 2001 (ano hidrológico 01/02) e em janeiro de 2003 (ano hidrológico 02/03). Verifica-se a existência de meses em que não foi registrada precipitação, geralmente nos meses de junho, julho e/ou agosto de cada ano. Embora seja possível inferir uma sazonalidade anual para a série a partir do gráfico da série temporal, a tendência não é facilmente observada.

Como a série de precipitação também apresenta uma periodicidade anual, o tamanho da janela escolhido foi um valor divisível por 12 e o mais próximo da metade da série ($L=84$). O limiar escolhido foi coincidente com o número de falhas presentes (9), e resultou na maior quantidade possível de vetores defasados completos para o tamanho de janela escolhido ($K=85$).

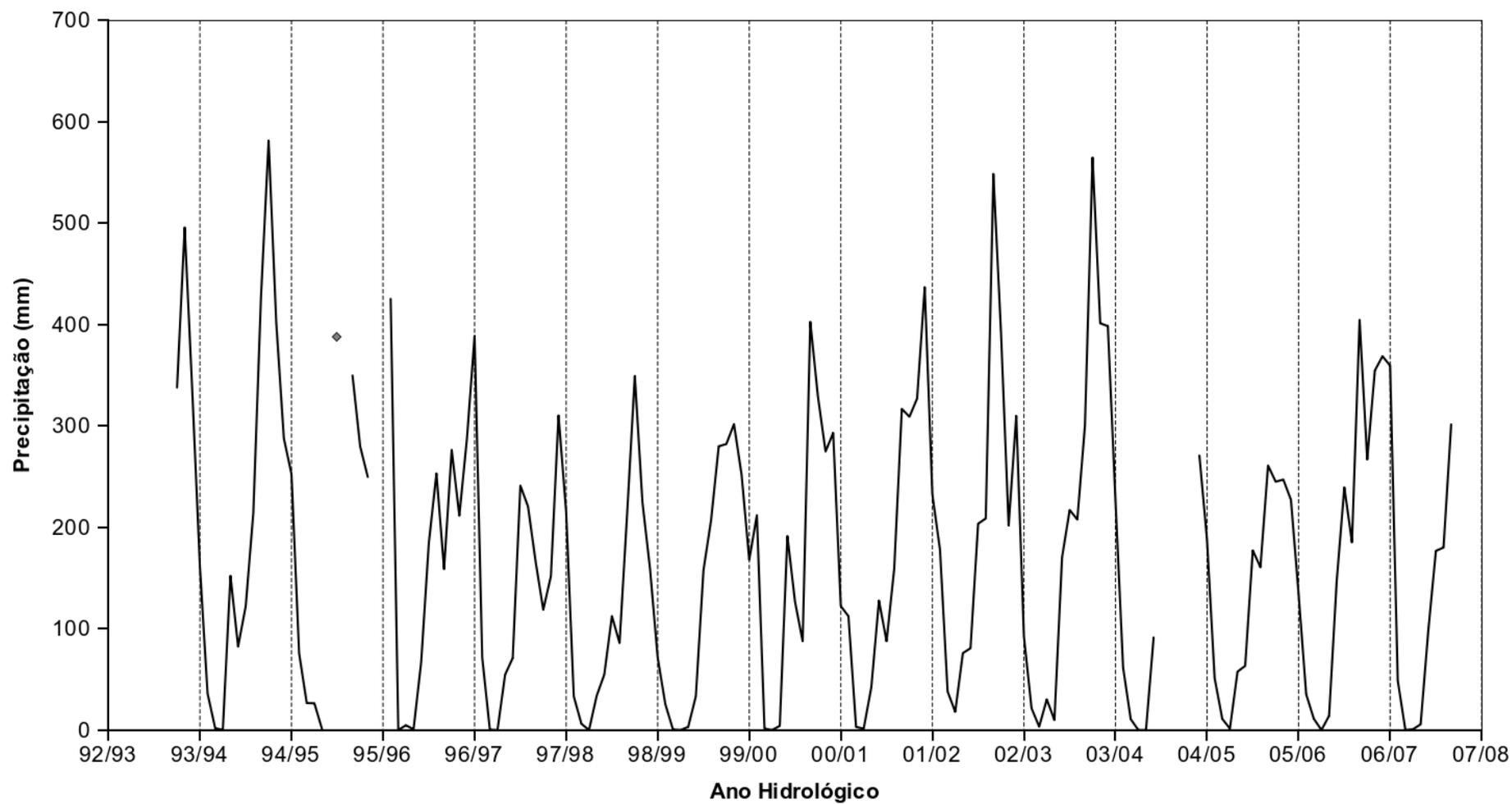


Figura 3.25 – Série temporal de precipitação, registrada na estação de 857000, durante o período de Janeiro de 1993 a Dezembro de 2006.

3.2.2.1 Separação do Sinal e Ruído

A interpretação dos gráficos é semelhante àquela apresentada no preenchimento da série temporal de vazão. Pode-se observar na Figura 3.26 que o autovetor 1 representa a tendência principal da série e está bastante separado dos demais autovetores. O componente harmônico anual resultou em um par de autovetores com valores singulares próximos (autovetores 2 e 3), formando o platô no gráfico. Porém, neste caso, a partir do autovetor 3 já foi difícil identificar autovetores que correspondessem ao sinal da série, uma vez que seus valores singulares estão bastante próximos e possivelmente são autovetores misturados aos ruídos.

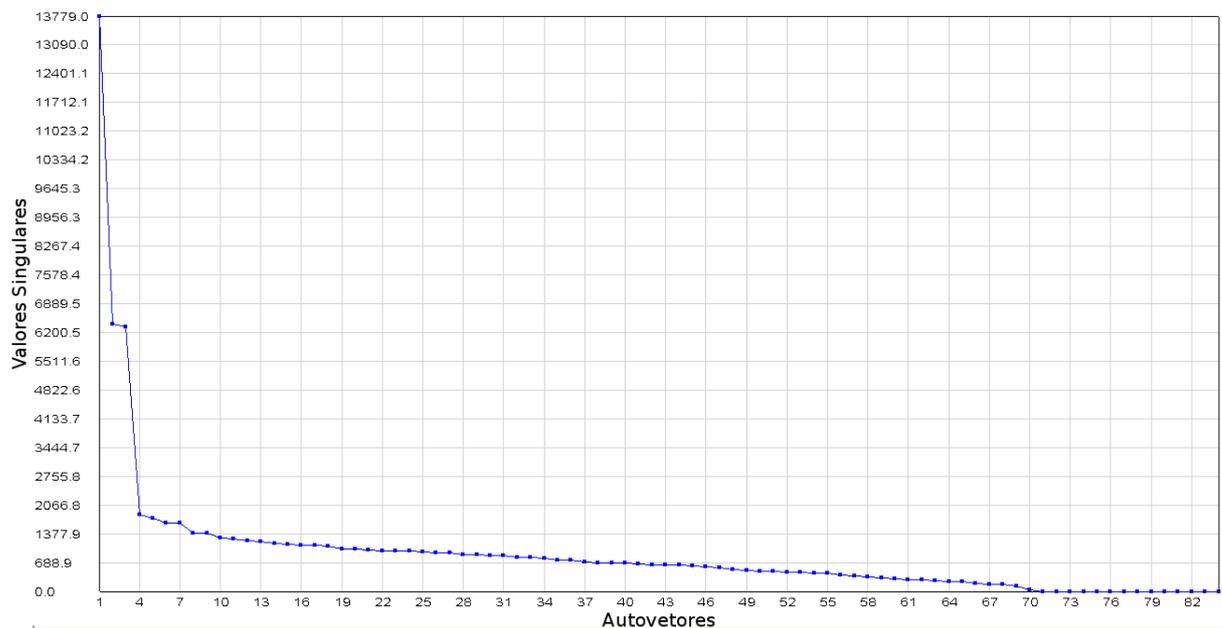


Figura 3.26 – Valores singulares dos autovetores defasados.

Deve-se notar que a separação do sinal foi menos eficiente neste caso, com a nítida separação apenas do autovetor de tendência e dos autovetores que correspondem a sazonalidade anual. Pela representação 2D dos autovetores (Figura 3.27), pode-se observar a principal tendência da série, representada pelo autovetor 1, e que os autovetores 2 e 3 representam, realmente, a sazonalidade anual.

Os autovetores 4, 5, 6 e 7 foram considerados como resultado da mistura entre um componente sazonal e um componente de tendência. Já os autovetores 8 e 9 apresentaram uma periodicidade semestral mas, por estarem muito misturados aos ruídos, foram desprezados na etapa de reconstrução.

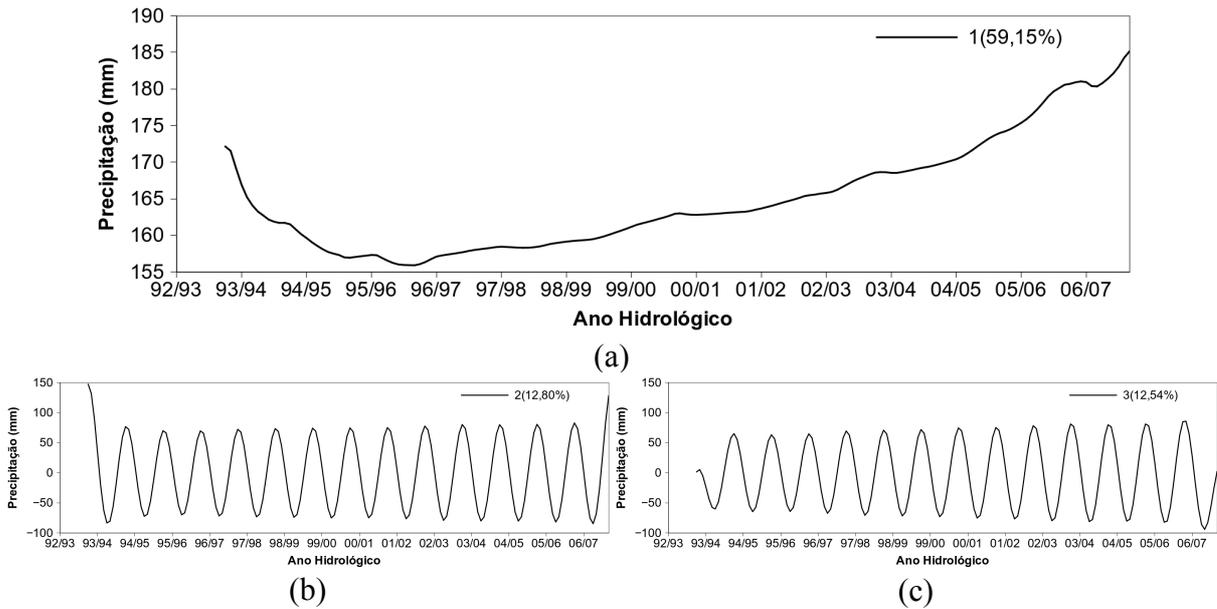


Figura 3.27 – Três primeiros autovetores reconstruídos, representativos do sinal, resultantes da decomposição da série temporal com $L=84$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12.

3.2.2.2 Reconstrução do Sinal da Série

Como apenas três autovetores foram extraídos de maneira eficiente para esta série, o agrupamento pode ser feito apenas de duas maneiras. Pode-se agrupar todos os autovetores para que a reconstrução da série seja feita, ou formar 2 grupos: um apenas com o autovetor 1 (tendência) e o outro com os autovetores 2 e 3 (componente harmônico). A Figura 3.28 (a), mostra a reconstrução da série utilizando os três autovetores, ou seja, mostra a reconstrução do sinal extraído da série. Novamente, a reconstrução ocorreu mesmo onde há a ocorrência de dados faltantes, preenchendo as lacunas no conjunto de dados. Os resíduos da etapa de reconstrução são mostrados na Figura 3.28 (b).

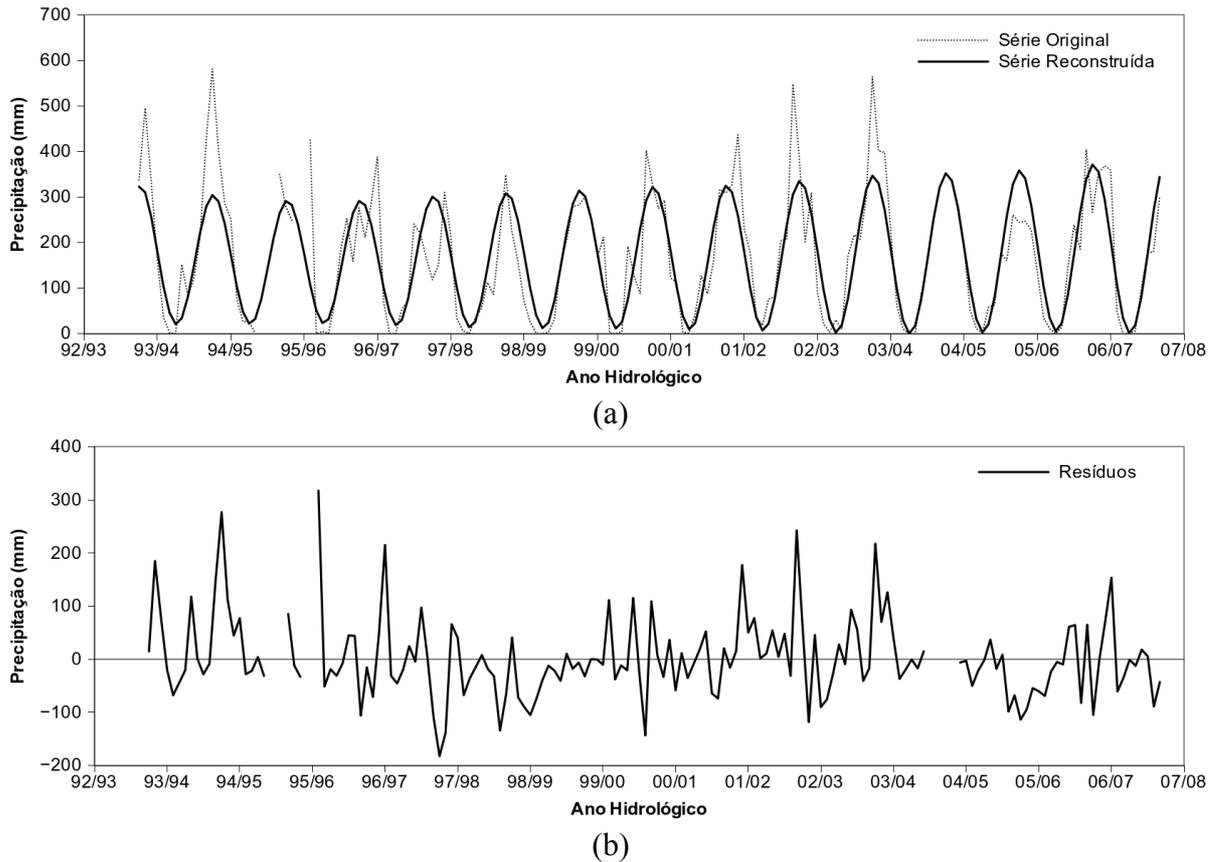


Figura 3.28 – Sinal reconstruído da série e resíduos da reconstrução. (a) série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2 e 3) e; (b) erros absolutos (resíduos) cometidos na etapa da reconstrução.

Os resíduos da etapa de reconstrução variaram de -182,31 a +317,63 mm cometidos na reconstrução dos valores referentes a janeiro de 1997 e maio de 1995, respectivamente.

3.2.2.3 Comparação entre os Métodos SSA e Regressão Linear Simples (RLS)

As estações mais próximas a estação Santa Rosa são mostradas na Tabela 3.1.

Tabela 3.1 – Estações pluviométricas mais próximas a estação Santa Rosa

Código	Nome	Rio	Estado	Município	Operadora	Distância ¹
758000	Barra do São Manuel	Tapajós	AM	Borba	Eletronorte	188 km
957000	Minação São Benedito	Teles Pires	PA	Itaituba	Eletronorte	52,2 km
957001	Novo Planeta	Teles Pires	MT	Aripuanã	CPRM	76,9 km
958000	Núcleo Ariel	Juruena	MT	Aripuanã	Desativada	142,7 km
1057001	Trivelato	Teles Pires	MT	Alta Floresta	CPRM	122,4 km
1058002	Núcleo Ariel	Rio Juruena	MT	Aripuanã	CPRM	142,0 km

¹ Distância entre cada uma das estações e a estação Santa Rosa (857000).

As estações Mineração São Benedito e Núcleo Ariel (958000) foram descartadas, pois a série de dados disponível para estas estações não cobrem os anos de interesse. A série de dados do posto Mineração São Benedito abrange o período de janeiro de 1978 a maio de 1983 e a série do posto Núcleo Ariel (958000) abrange o período de março de 1982 a setembro de 1990. As estações Barra do São Miguel e Núcleo Ariel (1058002) também foram descartadas. Estas duas estações, embora possuam registros de janeiro de 1993 a dezembro de 2005, apresentaram falhas nos mesmos meses das falhas de dados na estação Santa Rosa. As falhas de dados da estação Barra do São Miguel coincidiram com as falhas que vão de outubro de 2003 a fevereiro de 2004 enquanto uma falha da estação Núcleo do Ariel (1058002) coincidiu com a falha de setembro de 1994, na estação Santa Rosa.

Do exposto acima, e como destacado por Bertoni & Tucci (2004), nota-se os principais problemas na utilização do Método da Ponderação Regional e do Método de Regressão Linear Simples. Nem sempre existem estações próximas àquela que se deseja preencher. Além disso, nem sempre as estações próximas existentes possuem dados o suficiente para que se possa fazer o preenchimento. Além disso, as estações próximas podem apresentar falhas nos mesmos meses daqueles que se deseja preencher na estação de interesse. Das estações próximas, após uma análise prévia, apenas as estações Novo Planeta (957001) e Trivelato (1057001) puderam ser utilizadas no preenchimento das falhas de dados pelo método de RLS. Estas duas estações estão significativamente correlacionadas (Tabela 3.2) com a estação Santa Rosa, e podem ser utilizadas para o preenchimento pelo método da regressão.

Tabela 3.2 – Coeficientes de correlação entre as séries temporais de precipitação acumulada mensal.

Estações Correlacionadas	N*	Coeficiente de correlação (ρ)	Parâmetros estatísticos		
			t	df	p-value
Novo Planeta × Santa Rosa	151	0,8292	18,1132	149	<2,2x10 ⁻¹⁶
Trivelato × Santa Rosa	152	0,8615	20,7800	150	<2,2x10 ⁻¹⁶
Santa Rosa (SSA) × Santa Rosa	159	0,8249	18,2828	157	<2,2x10 ⁻¹⁶

* Corresponde ao número de dados utilizados no teste de correlação.

Os testes de hipóteses para os coeficientes a e b da equação de regressão linear simples entre as estações Novo Planeta e Santa Rosa, Figura 3.29 (a), foram significativos ao nível de significância de 5% (p -value = 0,0040 e p -value = $1,7 \times 10^{-39}$, respectivamente). Os coeficientes da regressão entre as estações Trivelato e Santa Rosa, Figura 3.29 (b), também foram significativos a 5% (p -value = 0,0065 e p -value = $5,3 \times 10^{-46}$, respectivamente). Os resultados dos testes de hipótese nos coeficientes das equações indicam que as equações que

melhor ajustam a relação entre os valores de precipitação registrados na estação Novo Planeta e Santa Rosa e entre o registrado nas estações Trivelato e Santa Rosa, são as equações mostradas nos Figuras 3.29 (a) e (b), respectivamente. Entretanto, os coeficientes de determinação das retas (R^2), foram relativamente baixos.

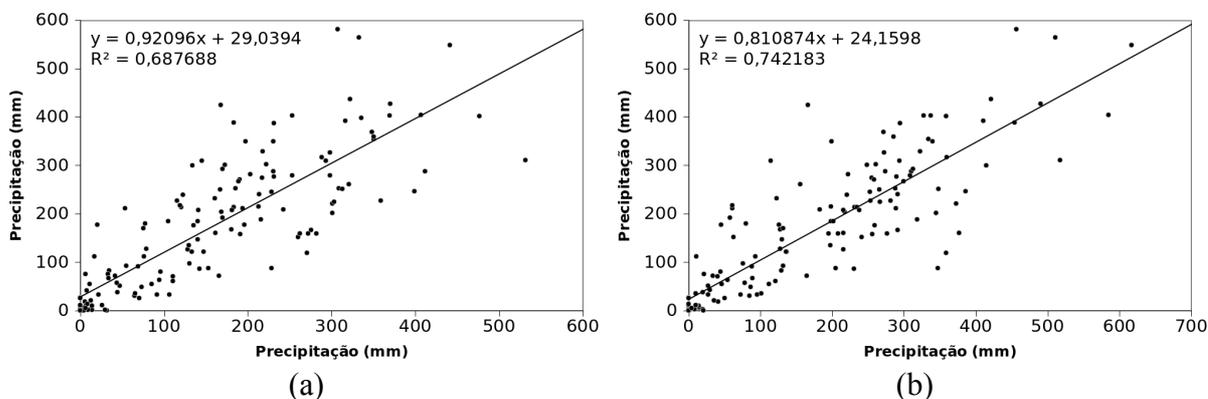


Figura 3.29 – Regressões lineares (a) entre a precipitação registrada na estação Novo Planeta - eixo das abcissas - e a registrada na estação Santa Rosa - eixo das ordenadas; (b) entre a precipitação registrada na estação Trivelato - eixo das abcissas - e a registrada na estação Santa Rosa - eixo das ordenadas.

O coeficiente de determinação mede a relação entre a variação explicada pela equação de regressão linear e a variação total da variável dependente. Assim, considerando o valor de R^2 , pode-se afirmar que cerca de 69% dos pares de dados da variação total da precipitação registrada na estação Santa Rosa pode ser explicada pela variação da precipitação registrada na estação Novo Planeta e cerca de 74% dos pares de dados a variação total da precipitação registrada na estação Santa Rosa pode ser explicada pela variação da precipitação registrada na estação Trivelato.

A partir das equações lineares, estimou-se cada um dos valores da série do posto Santa Rosa para que se pudesse calcular os Erros Médios Absolutos e Erros Médios Relativos (Tabela 3.3), para comparação com o preenchimento pelo método SSA.

Tabela 3.3 – Erros Médios Absolutos e os Erros Médios Relativos do preenchimento das falhas de dados.

Parâmetros	Método de Preenchimento		
	SSA	RLS	RLS
		Novo Planeta × Santa Rosa	Trivelato × Santa Rosa
Erro Médio Absoluto (EMA)	56,1	58,6	54,8
Erro Médio Relativo (EMR)	3,50%	3,10%	3,20%

Considerando os EMAs e os EMRs, nota-se que os valores são muito próximos. O

preenchimento através da equação linear entre a estação Novo Planeta e a Estação Santa Rosa apresentou o menor EMR, enquanto a regressão linear entre a estação Trivelado e a estação Santa Rosa apresentou o menor EMA. Embora a técnica de RLS apresente EMA na mesma ordem de grandeza dos erros do método SSA, o coeficiente de determinação indica que isto ocorre em 69% dos casos entre as estações Santa Rosa e Novo Planeta e em 74% dos casos entre as estações Santa Rosa e Trivelato, ou seja, nos outros 16% e 26% dos casos os erros serão maiores e não poderão ser explicados pela equação de regressão.

Uma vez que os preenchimentos foram bastante semelhantes, destaca-se que no método SSA não é necessária nenhuma informação além da própria série de dados que se quer preencher, enquanto a técnica de regressão depende da existência de outras séries, que, além disso, devem obedecer vários critérios para que possam ser utilizadas no preenchimento. Não obstante, o preenchimento através do método SSA, é capaz de extrair informações a respeito da série, tal como a tendência e os efeitos sazonais, além de poder ser utilizado como um método de previsão.

3.2.3 Séries Temporais de Variáveis de Qualidade da Água

No contexto da gestão de recursos hídricos o monitoramento da qualidade da água é um instrumento usado pelos governos em escala mundial para inferir o estado da qualidade das águas superficiais. Neste sentido existe um grande interesse nas séries temporais de variáveis de qualidade para que se possa extrair informação sobre as mudanças da qualidade da água ao longo do tempo. Portanto, o preenchimento das falhas de dados é de suma importância para se extrair informação sobre a qualidade da água já que todos os modelos de análise de séries temporais partem do pressuposto de que as séries devam ser completas. O não cumprimento desta condição implica na síndrome de riqueza de dados, pobreza de informação.

As séries temporais provenientes do programa de monitoramento da qualidade da água do Estado de Mato Grosso, apresentam muitos problemas no que tange a qualidade dos dados, seja em função da grande quantidade de falhas nas séries temporais, seja em função do grande número de dados censurados presentes nas variáveis DQO, DBO, fósforo, nitrogênio (NASCIMENTO, 2010) ou outros problemas no processo de medição das variáveis. Como todas as variáveis de qualidade, de todas as estações da bacia do Rio Cuiabá apresentam falhas de dados, é de extrema importância a aplicação de métodos de preenchimento visando a

recuperação das informações perdidas e que são importantes no processo de gestão de recursos hídricos e na tomada de decisão.

Todas as 12 estações de amostragem do Rio Cuiabá possuem falhas de dados. A estação Acorizal foi escolhida por possuir um alto percentual de falhas, de 39% a 47%, dependendo da variável. As principais características das variáveis de qualidade da água em cada uma das estações de amostragem são mostradas na Tabela 3.4. Ao contrário do que acontece com as estações pluviométricas e fluviométricas, a grande maioria das estações de monitoramento da qualidade da água apresentam muitas falhas. Em alguns casos, o percentual de falhas é de quase 50%.

Tabela 3.4 – Características das séries temporais das variáveis de qualidade da água de algumas estações de amostragem da bacia do Rio Cuiabá.

Estações		Temperatura da água	Turbidez	Sólidos	DQO	NTK	Fosfato Total
Acorizal	<i>N</i>	159	159	159	159	159	159
	Observações	93	97	92	84	92	93
	Falhas	66	62	67	75	67	66
	%	41,5%	39,0%	42,1%	47,2%	42,1%	41,5%
Barão de Melgaço	<i>N</i>	160	160	160	160	160	160
	Observações	96	99	88	98	95	96
	Falhas	64	61	72	62	65	64
	%	40,0%	38,1%	45,0%	38,8%	40,6%	40,0%
Córrego do Barbado	<i>N</i>	140	140	140	140	140	140
	Observações	93	95	87	93	90	90
	Falhas	47	45	53	47	50	50
	%	33,6%	32,1%	37,9%	33,6%	35,7%	35,7%
Nobres	<i>N</i>	112	112	100	112	112	112
	Observações	73	74	66	74	73	73
	Falhas	39	38	34	38	39	39
	%	34,8%	33,9%	34,0%	33,9%	34,8%	34,8%
Passagem da Conceição	<i>N</i>	162	162	162	162	162	162
	Observações	104	108	100	108	103	104
	Falhas	58	54	62	54	59	58
	%	35,8%	33,3%	38,3%	33,3%	36,4%	35,8%
Porto Cercado	<i>N</i>	162	162	162	162	162	162
	Observações	97	100	89	99	96	97
	Falhas	65	62	73	63	66	65
	%	40,1%	38,3%	45,1%	38,9%	40,7%	40,1%
São Gonçalo	<i>N</i>	119	119	119	119	119	119
	Observações	81	85	76	84	81	79
	Falhas	38	34	43	35	38	40
	%	31,9%	28,6%	36,1%	29,4%	31,9%	33,6%

Dados provenientes do monitoramento realizado pela Secretaria de Estado do Meio Ambiente de Mato Grosso.

Quimpo & Jing-Yea (1970) demonstraram que a frequência mensal de amostragem é ideal em termos de custo benefício para os programas de monitoramento da qualidade da água. Levando em consideração este resultado e o período de monitoramento das variáveis de qualidade da água na estação Acorizal, as séries temporais deveriam conter 159 registros ($N=159$). Entretanto, existem de 84 a 97 valores realmente observados, dependendo da variável. Das séries temporais das variáveis temperatura, turbidez, sólidos, DQO, nitrogênio e fósforo, apenas as variáveis turbidez e sólido totais apresentam um comportamento periódico em função do ciclo de estiagem e cheia característico da região. Em função das outras variáveis não apresentarem comportamento nitidamente periódico e da menor quantidade de falhas na série da Turbidez, se optou por demonstrar a aplicação do método SSA no preenchimento desta variável.

Embora as séries apresentem grandes quantidades de falhas durante todo o período de monitoramento, destaca-se a grande quantidade de falhas que ocorrem a partir do ano hidrológico 03/04, período em que não existem observações consecutivas. Nos anos hidrológicos 97/98 a 02/03, cuja frequência amostral foi quase mensal, se verificou apenas nove falhas.

3.2.3.1 Turbidez - Estação Acorizal (66255000)

A série temporal da variável turbidez reproduz o ciclo sazonal característico da região (Figura 3.30), e tem comportamento similar ao das variáveis vazão média e precipitação, exceto pelo fato dos picos na série ocorrerem no início de cada ano hidrológico e os menores valores ocorrerem ao final de cada ano.

A Figura 3.30 mostra que os maiores valores de turbidez presentes na série temporal são 81 UNT, medido em maio de 1993, 101 UNT, medido em agosto de 1996, 63 UNT medido em maio de 2002 e 72 UNT medido em abril de 2003. Os demais valores estão compreendidos na faixa de 1 a 53,6 UNT. Embora seja nítida uma sazonalidade de 12 meses, existe um comportamento irregular tanto nos picos, quanto nos mínimos da série.

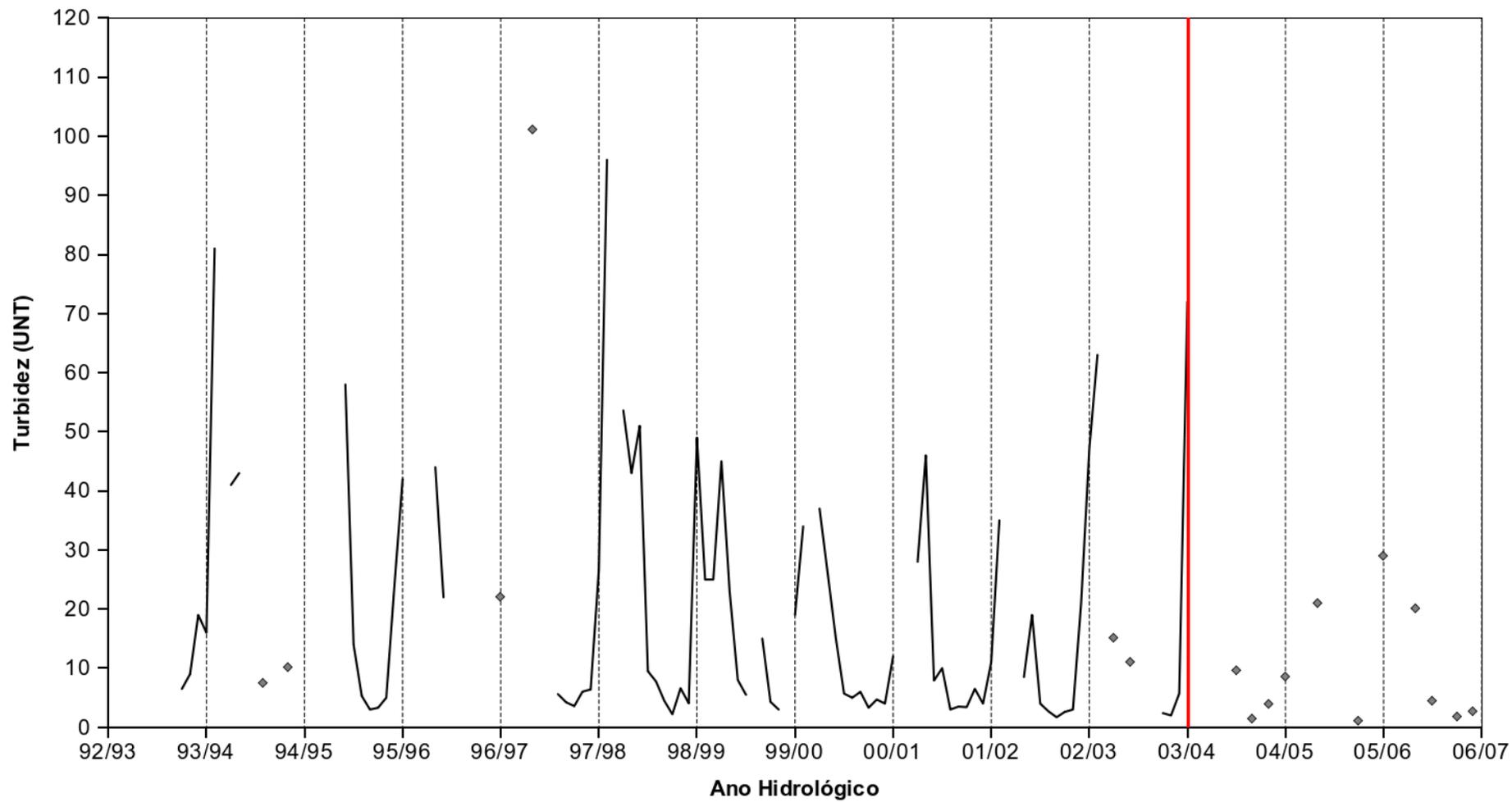


Figura 3.30 – Série temporal da variável Turbidez durante o período de Janeiro de 1993 a Março de 2006.

3.2.3.1.1 Decomposição da Série Original

Considerando o tamanho da série ($N=159$), foi escolhido um tamanho de janela igual a 72 ($L=72$), e limiar igual a 33, o que gerou 88 vetores defasados completos (W). A partir da interpretação do gráfico dos valores singulares dos autovetores, bem como das representações 2D dos autovetores, foram selecionados os autovetores 1, 2, 3, 4 e 5 como sendo representativos do sinal. O autovetor 1 representativo para a tendência da série, os autovetores 2 e 3 para o componente harmônico de período igual a 12 meses, e os autovetores 4 e 5 para o componente harmônico de período igual a 4 meses. A série reconstruída com os cinco autovetores mencionados é mostrada na Figura 3.31.

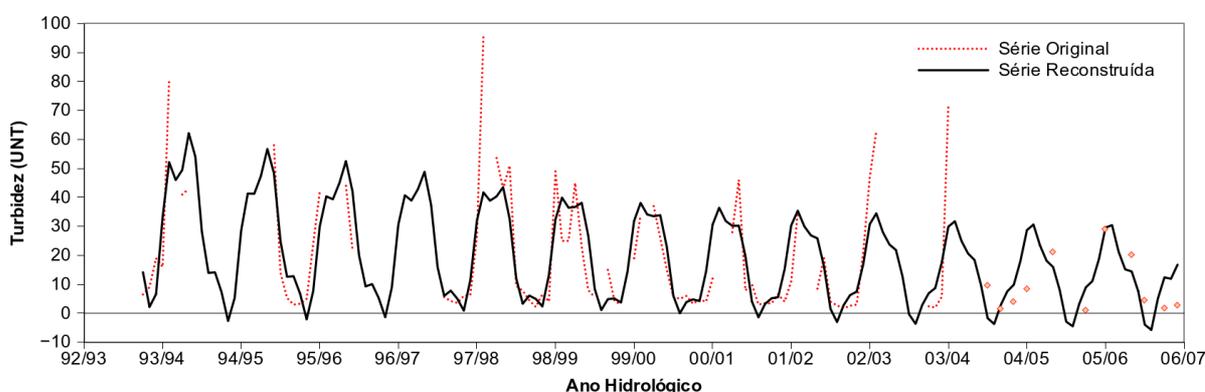


Figura 3.31 – Série temporal reconstruída a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5).

É importante notar, na Figura 3.31, que muitos valores estimados foram negativos. Estes valores são de difícil interpretação. A existência de valores negativos também ocorreu nas séries de precipitação, porém, neste caso, é razoável considerar que o valor provavelmente seria igual a zero uma vez que a não ocorrência de chuva nos meses de seca é bastante provável considerando as características da região e, de fato, alguns valores medidos ao longo do período de monitoramento foram iguais a zero.

Para minimizar o problema dos valores negativos, o tamanho da janela foi reduzido gradativamente. Porém, não se obteve resultados satisfatórios com nenhum tamanho de janela razoavelmente grande (>24) e divisível por 12. Considerando estes problemas, a série foi truncada no final do ano hidrológico 02/03, e os anos hidrológicos 03/04, 04/05 e 05/06 foram desconsiderados. Esta decisão foi tomada considerando que os dados existentes nestes anos hidrológicos não foram suficientes para caracterizar o período, uma vez que muitas falhas estão presentes e não há nenhuma medida em meses consecutivos. Assim, a série preenchida foi uma subsérie com $N=124$. Na Figura 3.30, o ponto de truncamento da série é representado

pela linha vermelha, no ano hidrológico 03/04. Considerando a série truncada e fixando o limiar em 26, obteve-se 65 vetores defasados completos.

3.2.3.1.2 Separação do Sinal e Ruído

Pode-se observar na Figura 3.32, os valores singulares dos autovetores extraídos na etapa da reconstrução. Assim como para a série completa, a série truncada resultou na separação de, nitidamente, cinco autovetores representativos para o sinal. O autovetor 1, representativo da tendência, os autovetores 2 e 3 que corresponde a um componente harmônico e os autovetores 4 e 5, também correspondentes a um componente harmônico.

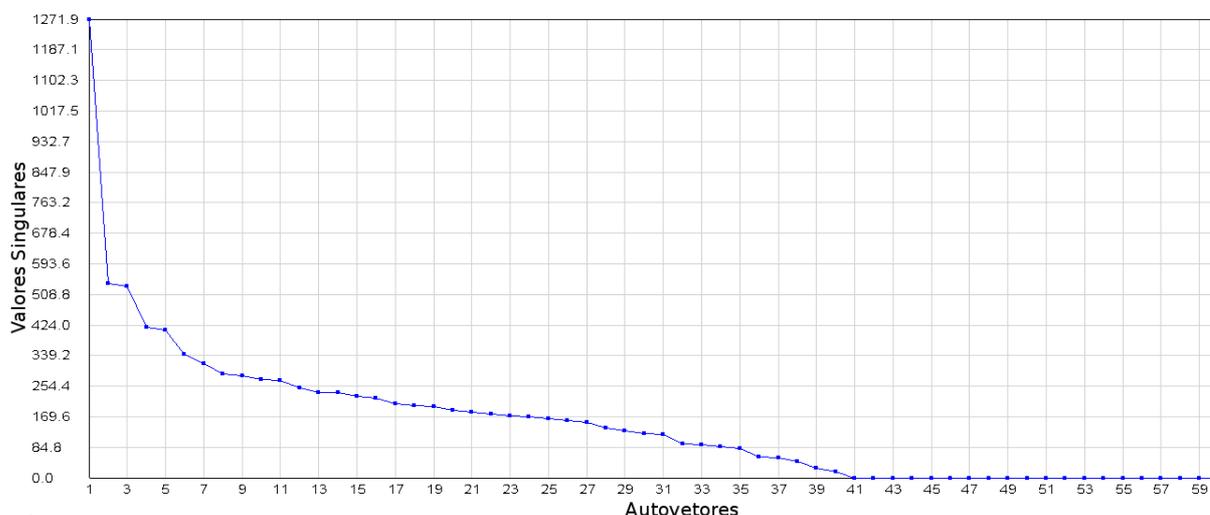


Figura 3.32 – Valores singulares dos autovetores defasados.

Observando a representação 2D dos autovetores, pode-se dizer que o autovetor 1 descreve a tendência principal da série, Figura 3.33 (a). A tendência da série, de maneira geral, apresenta comportamento linear e de diminuição dos valores. Os autovetores 2 e 3 representam o componente harmônico de período igual a 12, como pode-se observar na Figura 3.33 (b) e (c). Os autovetores 4 e 5 representam o componente de período igual a 4 meses, Figuras 3.33 (d) e (e) e, neste caso, há uma clara variação da amplitude dos autovetores que representam este componente harmônico, sendo a amplitude consideravelmente maior no início da série, diminuindo com o passar do tempo.

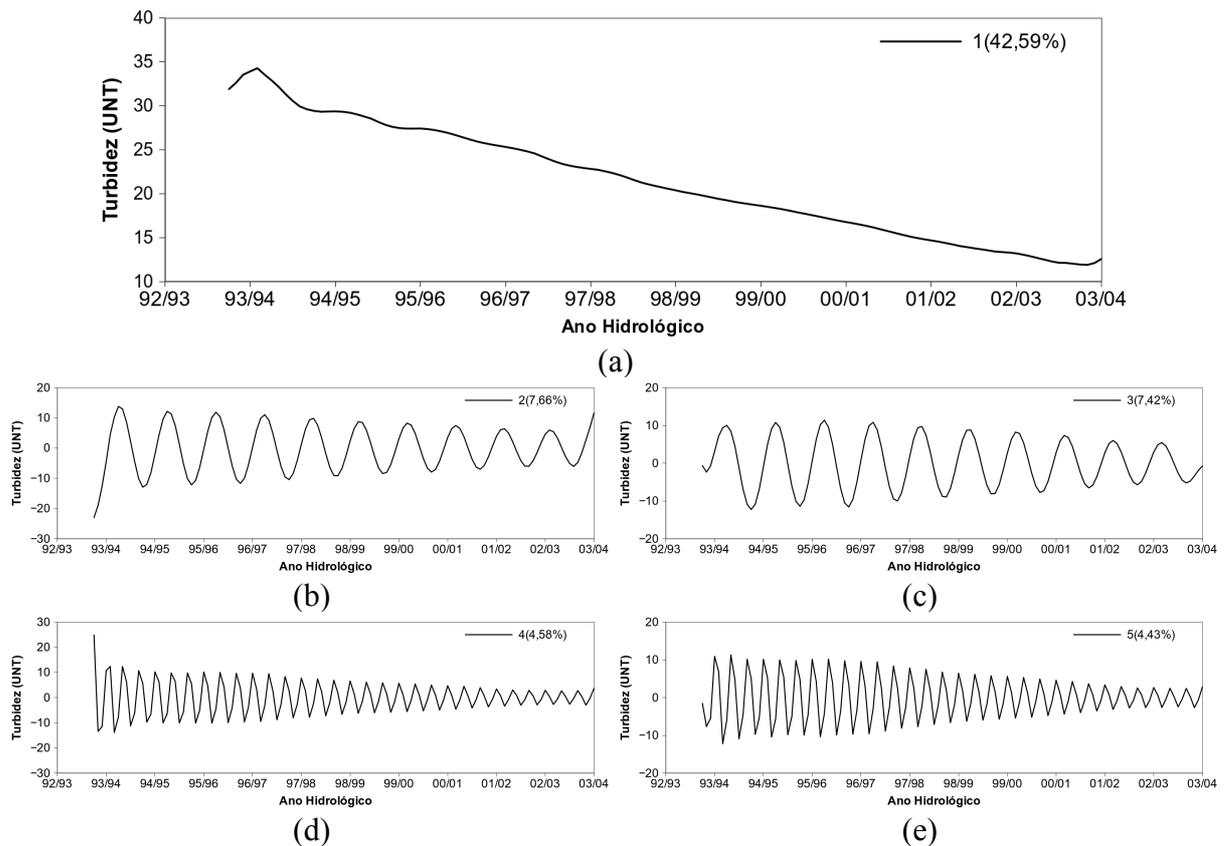
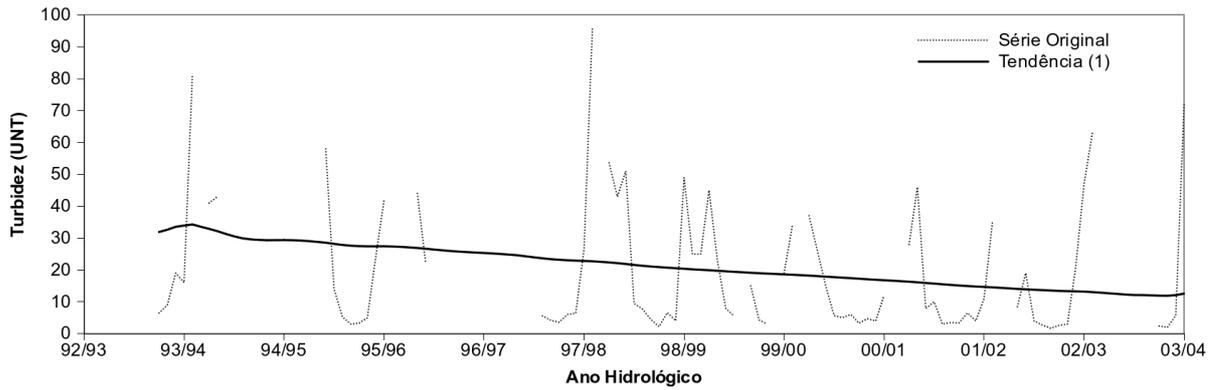


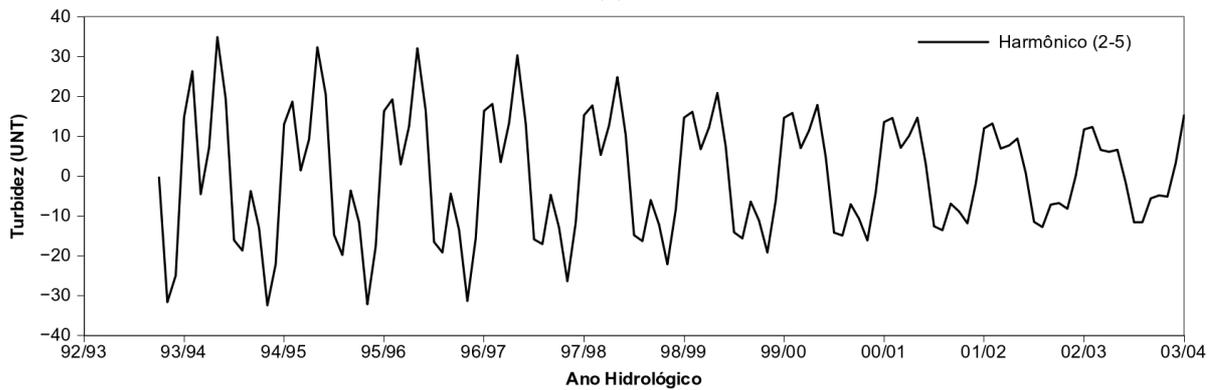
Figura 3.33 – Cinco primeiros autovetores reconstruídos representativos do sinal considerando $L=60$. (a) componente da tendência principal da série; (b) e (c) componente harmônico de período igual a 12; (d) e (e) componente harmônico de período igual a 4.

3.2.3.1.3 Reconstrução do Sinal da Série

Inicialmente, foi reconstruído apenas o autovetor representativo para a tendência da série, Figura 3.34 (a), tornando nítida a tendência de diminuição dos valores da série temporal. O agrupamento dos autovetores que representam os componentes harmônicos de frequência igual a 12 (autovetores 2 e 3) e igual a 4 (autovetores 4 e 5), e a posterior reconstrução, teve como resultado a série mostrada na Figura 3.34 (b). Este resultado mostra a capacidade do método de SSA modelar comportamentos complexos a partir do agrupamento dos autovetores. Tanto a tendência quanto o componente harmônico da série foram reconstruídos mesmo nos períodos de falha de dados, demonstrando a capacidade do método em estimar os dados faltantes.



(a)



(b)

Figura 3.34 – Séries reconstruídas através do método da projeção. (a) reconstrução da tendência da série utilizando o autovetor 1, sobreposta a série original; (b) reconstrução do componente harmônico da série com o agrupamento dos autovetores 2, 3, 4 e 5.

A partir do agrupamento de todos os autovetores representativos do sinal foi reconstruído o sinal da série temporal original e foi feito o preenchimento simultâneo dos dados faltantes, como pode ser observado na Figura 3.35.

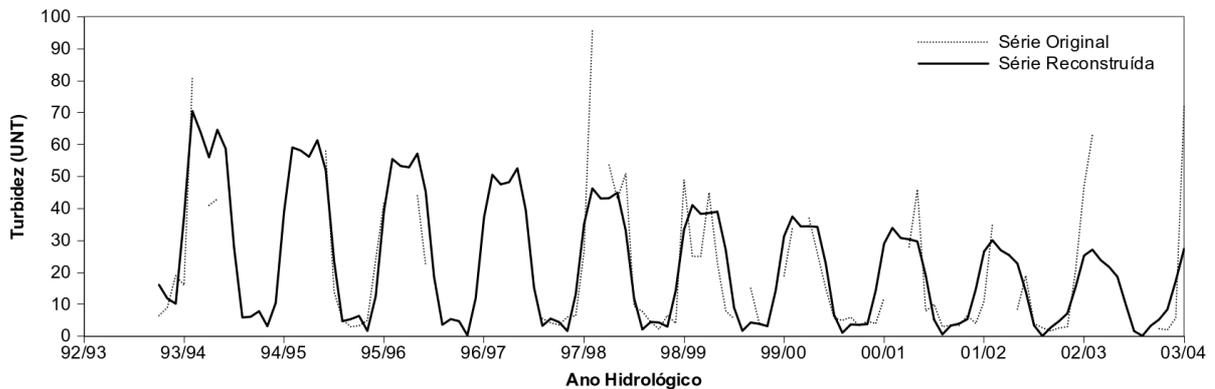


Figura 3.35 – Sinal reconstruído da série temporal a partir do agrupamento dos autovetores representativos do sinal (autovetores 1, 2, 3, 4 e 5), sobreposta a série temporal original.

Os resíduos da etapa de reconstrução variaram de -23,26 a 49,69 UNT cometidos, respectivamente, na reconstrução dos valores referentes a abril de 1998 e dezembro de 1999.

Além destes limites, erros acentuados, de 48,39 e 44,64 UNT, foram observados na reconstrução dos valores correspondentes a março de 1999 e dezembro de 2005. O resíduos da etapa de reconstrução são mostrados na Figura 3.36.

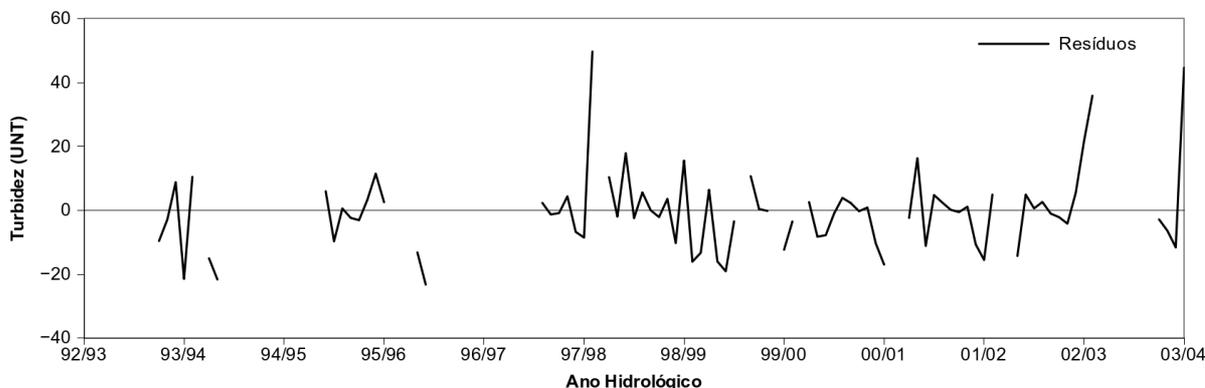


Figura 3.36 – Erros absolutos (resíduos) cometidos na etapa da reconstrução do sinal da série temporal de turbidez.

O EMA e o EMR da reconstrução foram iguais a 9,0 e 0,7% respectivamente. A série original e a série reconstruída não apresentaram uma alta correlação ($\rho = 0,7883$), porém foi significativa ao nível de significância de 5% ($t = 11,743$, $df = 84$, $p\text{-value} < 2,2 \times 10^{-16}$). Assim, pode-se dizer que a reconstrução da série foi satisfatória, e representativa para a série original, além do preenchimento das falhas ter sido coerente com as características de sazonalidade apresentadas pela série original truncada. Entretanto, destaca-se a impossibilidade de se preencher a série original em função dos valores negativos obtidos na reconstrução, sendo necessário truncar a série para que o número de falhas fosse diminuído, e a reconstrução fosse coerente.

Embora o método tenha se mostrado adequado para o preenchimento das variáveis vazão, precipitação e turbidez, no caso desta última, a qualidade dos dados gerados pelo programa de monitoramento dificultou a aplicação do método, uma vez que a quantidade de falhas é muito grande. Ainda assim, o método SSA foi capaz de extrair os autovetores harmônicos e o autovetor de tendência, sem que a série fosse truncada e estes foram essencialmente os mesmos autovetores extraídos na série truncada.

3.2.3.1.4 Reconstrução da Série Utilizando Autovetores de Sinal e Ruído

A Figura 3.37 mostra seis autovetores extraídos e que são proveniente de ruído. Estes autovetores não podem ser interpretados como provenientes de algum agente ou evento hidrológico e por isso devem ser descartados na etapa de reconstrução.

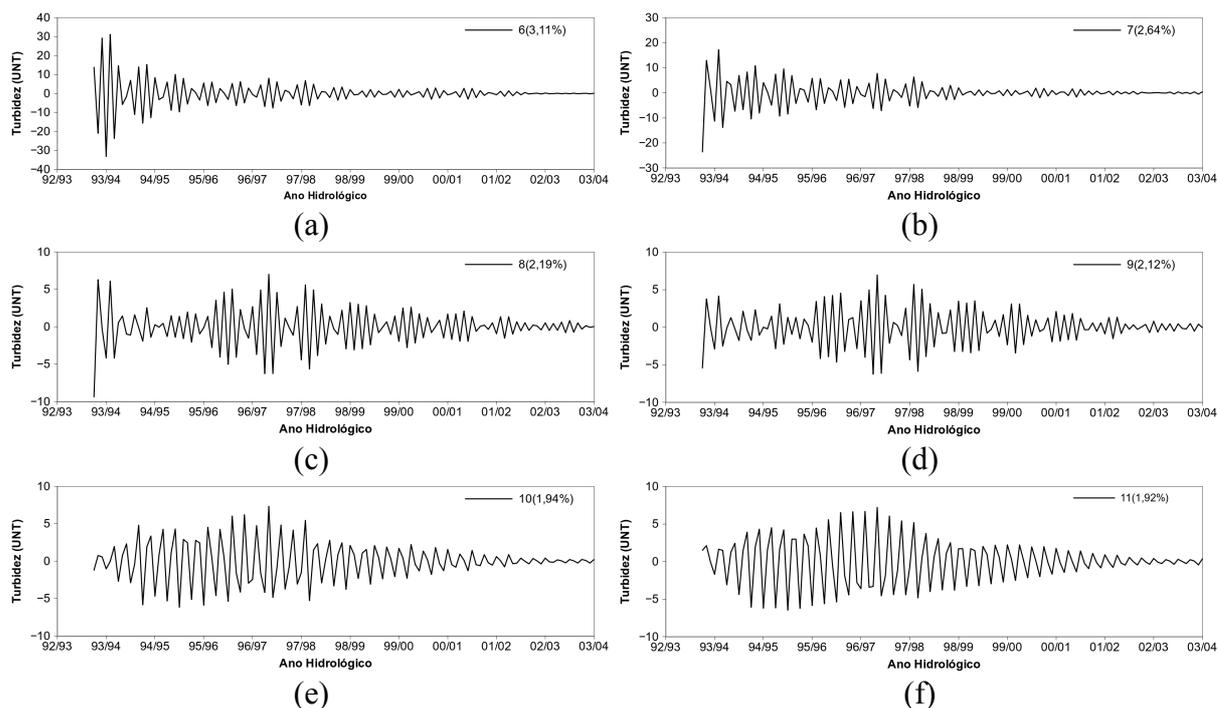


Figura 3.37 – Seis primeiros autovetores reconstruídos provenientes de ruídos na série temporal.

Uma vez que o método SSA decompõe a série em autovetores que, somados, representam a série original, a inclusão dos ruídos na reconstrução resulta numa série melhor ajustada aos dados originais, como pode ser observado na Figura 3.38. Esta série foi reconstruída incluindo 25 autovetores provenientes de ruído (6-30). Entretanto, a inclusão destes autovetores faz com que os valores estimados para os dados faltantes não tenham significado, ou seja, não são considerados informação.

Além disso, muitos valores estimados com a inclusão dos ruídos foram negativos e não seguem o comportamento observado na série original, muito embora os erros na estimativa dos valores observados sejam menores ($EMA=2,40$; $EMR=0,3\%$) que os erros encontrados a partir da reconstrução da série utilizando apenas autovetores representativos do sinal. Incluindo todos os autovetores de ruído, a série reconstruída tem a mesma forma da série original, porém não é possível fazer o preenchimento dos dados faltantes. Portanto, a separação dos autovetores representativos do sinal e os autovetores provenientes de ruído é de fundamental importância para uma reconstrução significativa da série temporal e para um preenchimento adequado e que apresente o mesmo comportamento da série original.

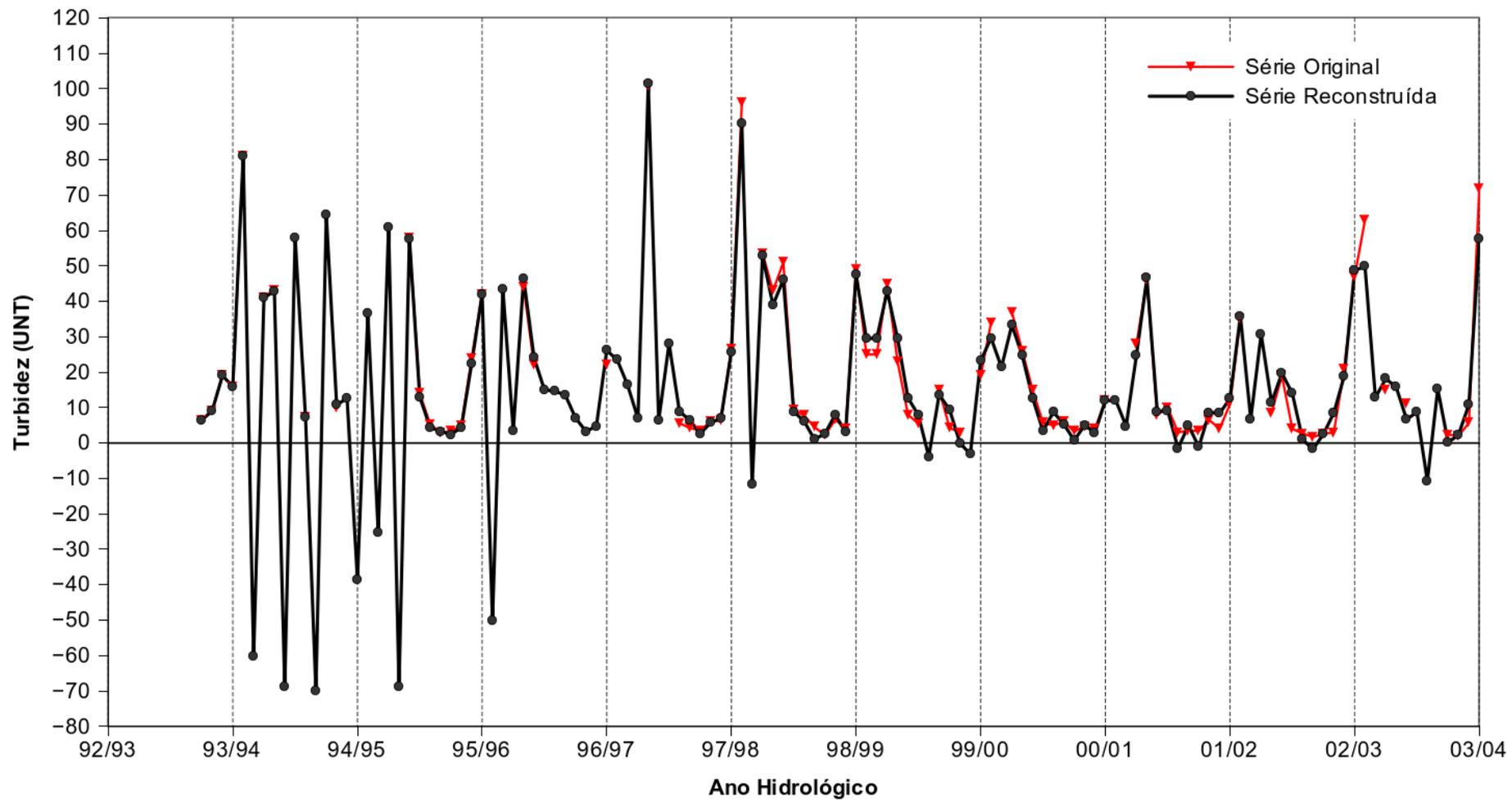


Figura 3.38 – Série reconstruída a partir do agrupamento dos trinta primeiros autovetores, sobreposta a série temporal original.

3.3 Previsão de dados em variáveis ambientais

A seguir são apresentados os resultados do uso do método SSA para a previsão das variáveis vazão média mensal da estação Cuiabá e de precipitação acumulada mensal da estação Vila Alegre.

3.3.1 Vazão Média Mensal - Estação Cuiabá (66260001)

A série original de vazão da estação Cuiabá ($N=552$) foi truncada no mês de novembro de 2003. O número de observações da série truncada foi reduzido para 513 ($N=513$). A partir da fixação do tamanho da janela ($L=252$), a série truncada foi decomposta. Seguindo as mesmas etapas descritas na análise da série, os autovetores extraídos foram agrupados e se procedeu a reconstrução da série. Tomando como base a série reconstruída, foi feita a previsão para os 39 meses retirados da série original (Figura 3.39).

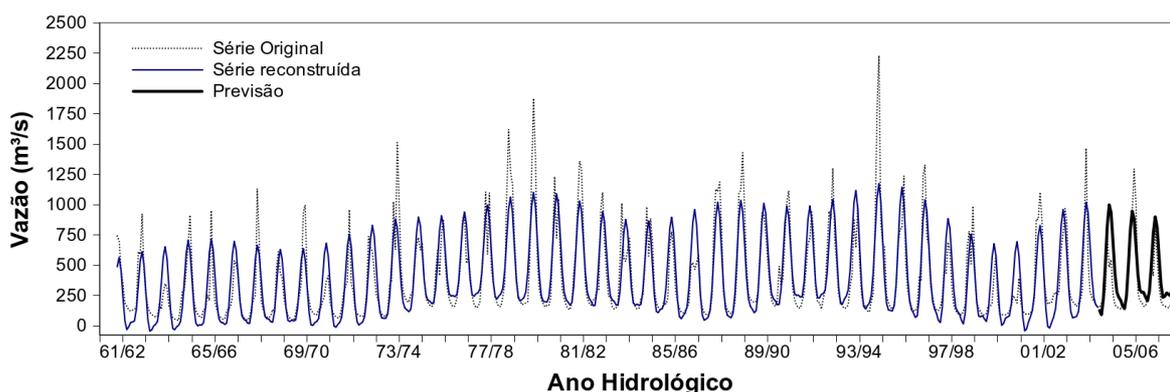


Figura 3.39 – Reconstrução da série temporal truncada e previsão dos valores retirados da série, sobreposta a série temporal original.

Os Erros Médio Absoluto (EMA) e Médio Relativo (EMR) cometidos na previsão foram $137,2 \text{ m}^3/\text{s}$ e $0,45\%$, respectivamente, e mantiveram-se dentro do esperado, tendo em vista os erros cometidos na análise desta série. Além disso, os valores previstos foram bastante semelhantes aos valores da série reconstruída para os mesmos meses.

Como os erros da previsão mantiveram-se dentro do esperado, foi feita a previsão para dois anos hidrológicos utilizando a série completa ($N=552$). A previsão pode ser observada na Figura 3.40.

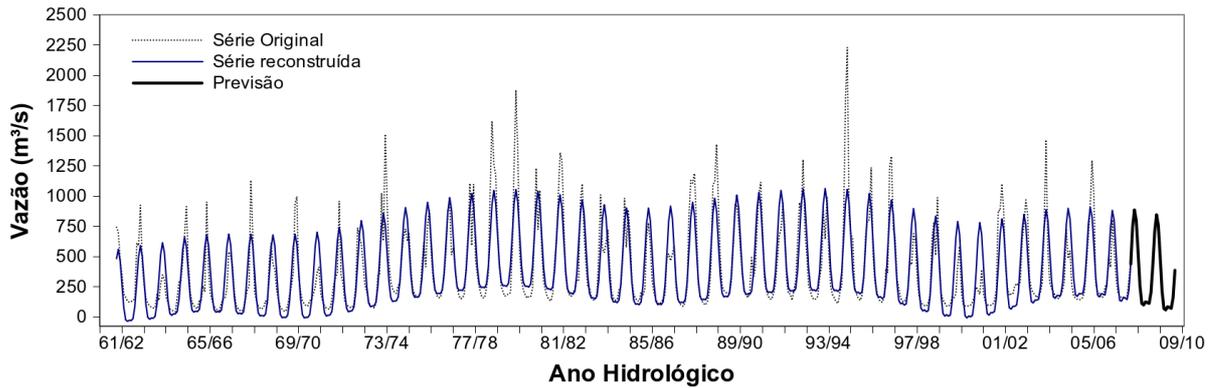


Figura 3.40 – Reconstrução da série temporal original e previsão para dois anos hidrológicos, sobreposta a série temporal original.

O método foi capaz de prever os valores futuros mantendo a estrutura extraída da série e o comportamento da série original. Os resultados das previsões para a variável vazão média demonstram, mais uma vez, a robustez do método.

3.3.2 Precipitação Acumulada Mensal - Estação Vila Alegre (1359001)

A série original de precipitação da estação Vila Alegre ($N=168$) foi truncada no mês de janeiro de 2005. O número de observações da série truncada foi reduzido para 144 ($N=144$). A partir da fixação do tamanho da janela ($L=72$), a série truncada foi decomposta e, a seguir, reconstruída procedendo-se a escolha dos autovetores representativos para o sinal da série. A partir disso foi feita a previsão para os 24 meses retirados da série original, como pode ser observado na Figura 3.41.

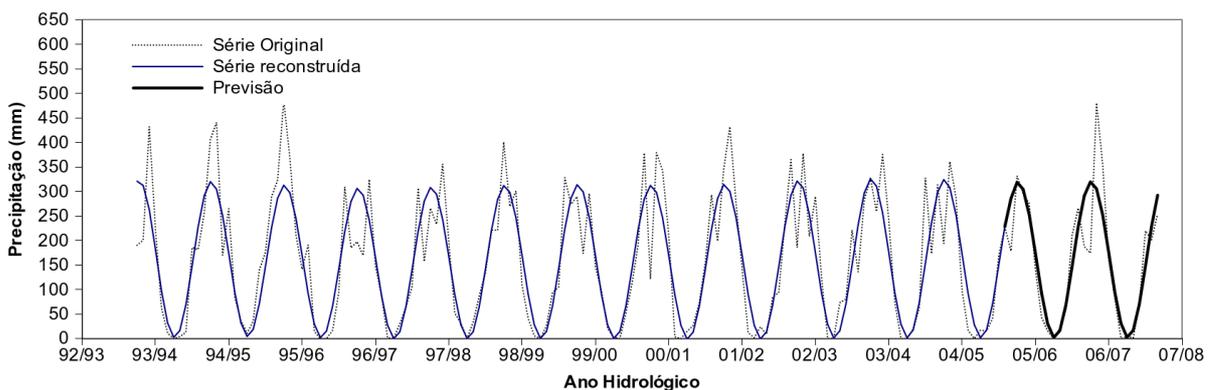


Figura 3.41 – Reconstrução da série temporal truncada e previsão dos valores retirados da série, sobreposta a série temporal original.

Neste caso, o EMA e o EMR cometidos na previsão foram $44,34 \text{ m}^3/\text{s}$ e $0,79\%$, respectivamente. Estes erros estão bem próximos dos erros encontrados na análise da série e, em função disso, foi feita a previsão para dois anos hidrológicos utilizando a série completa

($N=168$). A previsão pode ser observada na Figura 3.42.

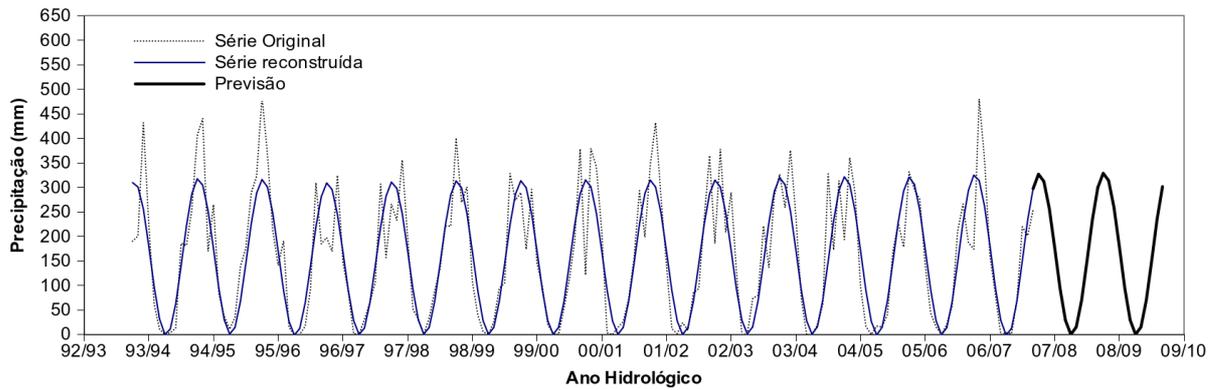


Figura 3.42 – Reconstrução da série temporal original e previsão para dois anos hidrológicos, sobreposta a série temporal original.

O método foi capaz, novamente, de prever os valores futuros mantendo a estrutura extraída da série e o comportamento da série original. As principais conclusões acerca da utilização do método na análise e preenchimento de falhas de dados em séries temporais de variáveis ambientais são apresentadas a seguir.

Capítulo 4

Conclusões

Com base nos resultados apresentados no capítulo anterior, conclui-se que o método de Análise de Espectro Singular (SSA) é uma ferramenta matemática robusta que pode ser utilizada para analisar séries temporais completas, preencher falhas de dados de séries incompletas e prever valores futuros de variáveis ambientais. Um enfoque mais detalhado de cada uma das três aplicações do método revelou os seguintes aspectos:

a) *Análise de Séries Temporais Completas*: o método permitiu determinar a característica de tendência da série, extrair componentes sazonais e separar as componentes que representam o sinal (informação) dos ruídos. No caso das variáveis analisadas neste estudo o método foi eficiente na extração dos componentes de tendência e componentes sazonais de todas as séries. Todas as séries apresentaram, pelo menos, um componente sazonal referente ao período de 12 meses. Análises dos autovetores, periodogramas e covariâncias médias confirmaram a sazonalidade em função do ciclo estiagem/cheia, característico da região. Em função dos diferentes regimes de escoamento ao longo do Rio Cuiabá, foram extraídos na Estação Cuiabá e na Estação Porto Cercado diferentes autovetores harmônicos.

b) *Preenchimento de Falhas de Dados*: o método também permitiu preencher falhas de dados de séries incompletas das variáveis vazão média e precipitação, com Erros Médios Relativos iguais a 0,2% e 3,5% respectivamente. O baixo valor encontrado para o Erro Médio Relativo durante a reconstrução da série de vazão média deve-se a estabilidade da série original, que não apresentou valores discrepantes ao longo de todo o período de monitoramento. No caso da variável turbidez, o grande número de falhas de dados tornou a aplicação do método mais difícil e a interpretação dos resultados foi dificultada pelo aparecimento sistemático de valores negativos na série reconstruída. Logo, a série foi truncada e os anos hidrológicos com muitas

falhas de dados consecutivas foram excluídos da série. Com a diminuição do tamanho da série e da proporção de dados faltantes, o preenchimento foi realizado de forma eficiente. Além disso, o método foi capaz de modelar a tendência e os componentes harmônicos da série simultaneamente ao preenchimento. Entretanto, não foi possível recuperar as informações perdidas nos anos hidrológicos de 03/04, 04/05 e 05/06 devido ao grande número de falhas de dados consecutivas na série de turbidez.

c) *Previsão de Valores Futuros*: O método foi capaz de prever valores futuros para um período de 24 meses mantendo a estrutura e o comportamento das séries originais. Os resultados das previsões para as variáveis vazão média e precipitação acumulada mensal produziram Erros Médios Relativos iguais a 0,45% e 0,79%, respectivamente. Para determinar o valor futuro da sequência de observações foram realizados três passos: separar os autovetores correspondentes ao sinal dos componentes de ruído, determinar o padrão característico dos autovetores que representam o sinal e utilizar este padrão para prever valores futuros.

d) *Vantagens em Relação a Outros Métodos*: a grande maioria dos estudos realizados no Brasil sobre preenchimento de falhas de dados de variáveis ambientais utilizam os métodos de regressão linear e/ou média móvel. No caso específico da variável precipitação, os métodos de regressão linear simples, ponderação regional e krigagem (método geoestatístico de regressão para análise e interpolação que pressupõe que os dados encontram-se espacialmente correlacionados) também são utilizados. Para utilizar de forma eficiente o método da ponderação regional, a série temporal deve possuir, no mínimo, dez anos e deve-se ter disponível dados de pelo menos três estações vizinhas. Além disso, as estações devem estar numa região climatológica semelhante àquela da estação que se deseja preencher. Este fato é um fator limitante na grande maioria dos estudos. No caso do preenchimento pelo método da regressão, pode-se optar pela regressão linear simples, na qual é necessário dados de apenas uma estação vizinha, ou pela regressão linear múltipla, na qual é possível relacionar dados de várias estações próximas para que o preenchimento seja feito. O maior problema na aplicação destes métodos é encontrar estações vizinhas e representativas para que se possa fazer o preenchimento. Geralmente, falhas em longos períodos e a grande variabilidade dos dados dificultam o ajuste por métodos usualmente utilizados de interpolação. A principal vantagem do método SSA para o preenchimento de falhas de dados em relação aos métodos de regressão linear simples, regressão linear múltipla, método da ponderação regional e

krigagem, deve-se a sua natureza auto-adaptativa que dispensa a dependência de séries temporais de estações vizinhas. O preenchimento é realizado durante o processo de reconstrução usando exclusivamente os próprios dados da série.

e) *Gestão de Recursos Hídricos e Programas de Monitoramento*: o método SSA pode ser aplicado como ferramenta de auxílio à gestão de programas de monitoramento, especialmente no monitoramento da qualidade da água, onde os dados são provenientes de campanhas de campo e posterior análise em laboratório e cujo custo é elevado. Em função do custo elevado, a maioria dos programas de monitoramento no Brasil realiza apenas quatro amostragens durante um ano hidrológico. Porém, levando em conta que as variáveis ambientais são governadas por processos estocásticos ou aleatórios, se essas quatro amostragens não forem distribuídas no tempo de forma adequada, a amostragem pode não ser significativa do ponto de vista estatístico. Como a frequência de amostragem afeta diretamente não apenas os custos do sistema de monitoramento, mas também a informação que pode ser obtida a partir dos dados disponíveis, existe a necessidade de se utilizar métodos estatísticos como critério quantitativo para a escolha dos intervalos de amostragem. O método SSA, como demonstrado neste estudo, pode ser uma ferramenta útil na recuperação de informações perdidas, minimizando o custo do monitoramento e maximizando a qualidade da informação produzida a partir dos dados coletados.

f) *Sugestões Para Estudos Futuros*: no que diz respeito ao formalismo matemático, existe a necessidade de estudos mais detalhados para determinar a causa do aparecimento de valores negativos no processo de reconstrução da série. Resultados preliminares indicam que isso ocorre em três situações: i) quando existem longas falhas consecutivas, ii) quando as falhas ocorrem com certa regularidade e iii) quando existem valores iguais a zero ou muito próximos a zero na série original. No que diz respeito a aplicação do método e, no caso específico das variáveis de qualidade de água (comumente utilizadas em programas de monitoramento), é importante determinar qual deve ser a mínima frequência de amostragem para que os dados observados possam ser utilizados para a análise, preenchimento e previsão de valores futuros.

Referências Bibliográficas

ANA – Agência Nacional de Águas – **HIDROWEB**. Séries Históricas – estações pluviométricas e fluviométricas. Disponível em: <<http://hidroweb.ana.gov.br/>>. Acesso em: julho de 2009.

BERTONI, J. C. & TUCCI, C. E. M. (2004) Precipitação. In: TUCCI, C. E. M. (Org.) **Hidrologia: Ciência e Aplicação**. Terceira edição, primeira reimpressão. - Porto Alegre, Editora da UFRGS/ABRH. 943p.

BLOOMFIELD, P. (2000) **Fourier Analysis of Time Series – An introduction**. Second Edition, John Wiley & Sons, INC. 261p.

BROCKWELL, P. J. & DAVIS, R. A. (2002) **Introduction to Time Series and Forecasting**. Second Edition, Springer. 434p.

BROERSEN, P. M. T., WAELE, S. and BOS, R. (2004) Autorregressive spectral analysis when observations are missing. **Automatica**, 40 (9), p. 1495-1504.

BROERSEN, P. M. T. (2006) Automatic spectral analysis with missing data. **Digital Signal Processing**. 16, p. 754-766.

BROOMHEAD, D. S. & KING, G. P. (1986a) Extracting qualitative dynamics from experimental data. **Physica D**. 20, p. 217-236.

BROOMHEAD, D. S. & KING, G. P. (1986b) On the qualitative analysis of experimental dynamical systems. In: SARKAR, S. (Ed.) **Nonlinear Phenomena and Chaos**. Adam Hilger, Bristol. p. 113-144.

BUCHSTABER, V. M. (1994) Time series analysis and grassmannians. In: GINDIKIN, S. (Ed.) **Applied Problems of Random Transform**. Volume 162 of AMS Transactions – series 2. AMS, Province. p. 1-17.

CHATFIELD, C. (2004) **The analysis of Time Series: an introduction**. Sixth edition,

Chapman & Hall/CRC.

COWPERTWAIT, P. S. P & METCALFE, A. V. (2009) **Introductory Time Series with R**. Springer. 254p.

CRYER, J. D. & CHAN, K-S. (2008) **Time Series Analysis: With Applications in R**. Second Edition, Springer. 491p.

DANILOV, D. (1997) Principal components in time series forecast. **Journal of Computacional and Graphical Statistics**, 6, p. 112-121.

EHLERS, R. S. (2009) **Análise de Séries Temporais**. 5ª edição. 114p.

ELSNER, J. B. & TSONIS, A. A. (1996) **Singular Spectrum Analysis: A new tool in Time Series Analysis**. Plenum Press, New York and London. 164p.

FAN, J. & YAO, Q. (2003) **Nonlinear Time Series: Nonparametric and Parametric Methods**. Springer, New York, INC. 551p.

GHIL, M., ALLEN, M. R., DETTINGER, M. D., IDE, K., KONDRASHOV, D., MANN, M. E., ROBERTSON, A. W., SAUNDERS, A., TYAN Y., VARADI, F. and YIOU, P. (2002) Advanced spectral methods for climatic time series, **Review of Geophysics**, 40 (1), p. 1–41.

GOLYANDINA, N. & OSIPOV, E. (2007). The “Caterpillar”-SSA method for analysis of time series with missing values. **Journal of Statistical Planning and Inference**. 137, p. 2642–2653.

GOLYANDINA, N., NEKRUTKIN, V. and ZHIGLJAVSKY, A. (2001) **Analysis of Time Series Structure: SSA and Related Techniques**. Chapman & Hall/CRC, London. 305p.

HAIR, J. F., TATHAM, R. L., ANDERSON, R. E. e BLACK, W. (2005) **Análise Multivariada de Dados**. 5º edição, Bookman, 600p.

HASSANI, H. (2007) Singular Spectrum Analysis: Methodology and Comparison. **Journal of Data Science**. 5, p. 239-257.

HOCKE, K. & KÄMPFER, N. (2009), Gap filling and noise reduction of unevenly sampled

data by means of the Lomb-Scargle periodogram. **Atmos. Chem. Phys.** v9, p. 4197-4206.

KIRCHGÄSSNER, G. & WOLTERS, J. (2007) **Introduction to Modern Time Series Analysis**. Springer. 274p.

KONDRASHOV, D., FELIKS, Y. and GHIL, M. (2005) Oscillatory modes of extended Nile River records (A.D. 622-1922). **Geophysical Research Letters**. 32 (10), L10702.

KONDRASHOV, D. & GHIL, M. (2006), Spatio-temporal filling of missing points in geophysical data sets. **Nonlinear Processes in Geophysics**. v.13, p. 151-159.

LINDSEY, J. K. (2004) **Statistical Analysis of Stochastic Process in Time**. New York. Cambridge University Press. 338p.

LIPSCHUTZ, S. & LIPSON, M. L. (2001) **Álgebra Linear**. 3ª edição, McGraw-Hill Companies, Inc, 400p.

MARQUES, C. A. F., FERREIRA, J. A., ROCHA, A., CASTANHEIRA, J. M., MELO-GONÇALVES, P., VAZ, N. and DIAS, J. M. (2006) Singular spectrum analysis and forecasting of hydrological time series. **Physics and Chemistry of the Earth**. 31, p. 1172-1179.

MORETTIN, P. A. & TOLOI, C. M. C. (2006) **Análise de Séries Temporais**. 2ª edição, rev. E ampl. - São Paulo: Edgard Blücher. 538p.

MOSKVINA, V. G. & ZHIGLJAVSKY, A. A. (2000) Applications of the singular spectrum analysis for change-point detection in time series. **Journal of Time Series Analysis**. (submetido).

NASCIMENTO, O. da C. (2010) **Análise da Qualidade dos dados gerados pelos Programas de Monitoramento da Qualidade das Águas das Bacias Hidrográficas do Rio Cuiabá e Rio das Garças**. Dissertação (Mestrado em Recursos Hídricos) – Universidade Federal de Mato Grosso, Cuiabá. 131p.

POLLOCK, D. S. G. (1999) **A Handbook of Time-Series Analysis, Signal Processing and Dynamics**. Academic Press, 782p.

QUIMPO, R. G. & JING-YEA, Y. A. N. G. (1970) Sampling consideration in stream discharge and temperature measurements. **Water Resources Research**, 6 (6), p. 1711-1774.

SAHU, B. K. (2003) **Time Series Modelling in Earth Sciences**. First Edition. Taylor & Francis, 294p.

SCHOELLHAMER, D. H. (2001) Singular Spectrum Analysis for time series with missing data. **Geophysical Research Letters**. 28 (16), p. 3187-3190.

UYSAL, M. (2007), Reconstruction of Time Series Data with Missing Values. **Journal of Applied Sciences**. v.7, p. 922-925.

Livros Grátis

(<http://www.livrosgratis.com.br>)

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)
[Baixar livros de Literatura de Cordel](#)
[Baixar livros de Literatura Infantil](#)
[Baixar livros de Matemática](#)
[Baixar livros de Medicina](#)
[Baixar livros de Medicina Veterinária](#)
[Baixar livros de Meio Ambiente](#)
[Baixar livros de Meteorologia](#)
[Baixar Monografias e TCC](#)
[Baixar livros Multidisciplinar](#)
[Baixar livros de Música](#)
[Baixar livros de Psicologia](#)
[Baixar livros de Química](#)
[Baixar livros de Saúde Coletiva](#)
[Baixar livros de Serviço Social](#)
[Baixar livros de Sociologia](#)
[Baixar livros de Teologia](#)
[Baixar livros de Trabalho](#)
[Baixar livros de Turismo](#)