ANÁLISE ESTATÍSTICA MULTIVARIADA EM GESTÃO DE RECURSOS HÍDRICOS: ESTUDO DE CASO DA BACIA DO ALTO IGUAÇU

Camila de Carvalho Almeida [1]; Cristovão V. Scapulatempo Fernandes[2] & Marianne S. França Sieciechowicz[3]*

Resumo - A análise estatística multivariada permite sintetizar grandes conjuntos de dados, podendo servir para a identificação de parâmetros chaves no monitoramento de qualidade da água e assim contribuir para a gestão de recursos hídricos. No Brasil, os parâmetros de qualidade são estabelecidos pela Resolução CONAMA 357/05 que está baseada em concentrações, porém, sabe-se que somente este tipo de análise não transmite de forma eficaz a situação do corpo hídrico, já que não leva em consideração a variação da vazão. No presente trabalho, os dados de concentração e carga obtidos no monitoramento de qualidade da água da Bacia do Alto Iguaçu foram avaliados através de técnicas estatísticas multivariadas com o objetivo de identificar os parâmetros que se destacam em cada tipo de dado. Foram utilizadas as técnicas de Análise de Componentes Principais (ACP), Análise Fatorial (AF) e Análise de Agrupamentos (AA). Os resultados para concentração e carga foram diferentes, sendo que a análise de carga apresentou resultados mais satisfatórios do que concentração para algumas técnicas. No geral, os parâmetros que mais se destacaram foram aqueles ligados à degradação da matéria orgânica, aos sólidos e ao nitrogênio, portanto, os parâmetros que refletiram de maneira direta as fontes de poluição ao longo da bacia.

Palavras-Chave – qualidade da água, gestão de recursos hídricos, análise estatística multivariada.

MULTIVARIATE STATISTICAL ANALYSIS FOR WATER RESOURCES PLANNING AND MANAGEMENTE: CASE STUDY OF ALTO IGUAÇU BASIN

Abstract – Multivariate statistical analysis allows synthesizing large data sets, which may serve to identify key parameters in monitoring water quality and contribute to the management of water resources. In Brazil, the quality parameters are established by CONAMA Resolution 357/05 which is based on concentrations, however, it is known that only this type of analysis does not convey effectively the situation of the water body, since it does not take into account the flow variation. The objective of this work was to apply multivariate statistical techniques in concentration data and load obtained in monitoring water quality of the Alto Iguaçu Basin to identify the parameters that stand out in each data type. We used the techniques of Principal Component Analysis (PCA), Factor Analysis (FA) and Cluster Analysis (AA). The results for the concentration and load are different, and the load analysis showed that the most satisfactory results for some concentration techniques. In general, the parameters that stood out were those linked to the degradation of organic matter and nitrogen to solids, so the parameters that reflect a direct pollution sources along the basin.

Keywords – water quality, water resources management, multivariate statistical analysis.

[1] Universidade Federal do Paraná, carvalho_camila@ymail.com
[2] Universidade Federal do Paraná, cris.dhs@ufpr.br
[3] Instituto de Tecnologia para o Desenvolvimento - LACTEC, marianne.franca@lactec.org.br

1. INTRODUÇÃO

A taxa de crescimento das populações urbanas só vem aumentando com o passar dos anos e geralmente de forma rápida e irregular, por isso, um dos grandes desafios atuais é conseguir proporcionar às populações urbanas condições adequadas de saneamento básico e ambiental.

Neste contexto, surge a necessidade do monitoramento da qualidade da água que permite avaliar o que está ocorrendo de fato na bacia com a perspectiva de identificação dos parâmetros que estão sendo mais afetados, os trechos e regiões mais sensíveis e quais são as fontes potenciais dessa poluição. Como o comportamento de um corpo hídrico depende de vários fatores, para que o monitoramento seja eficiente é preciso um longo tempo de medição de diversos parâmetros, o que gera um alto custo monetário e um grande conjunto de dados de difícil interpretação. Por isso é importante estudar os resultados do monitoramento sobre diferentes aspectos e utilizar ferramentas que permitam uma melhor compreensão do conjunto de informações.

De acordo com Pinto e Maheshwari (2011), desde a última década vários países vêm tentando desenvolver métodos mais adequados para avaliar a saúde da água doce para espécies bióticas e seres humanos. Dentre esses métodos, as técnicas de estatística multivariada vêm se sobressaindo, pelo fato de possibilitarem a análise de grandes e complexos conjuntos de dados e de conseguirem simplificar esta estrutura, identificando as variáveis mais representativas. A ideia central da utilização da técnica de análise estatística multivariada em gestão dos recursos hídricos é a de permitir identificar quais são as melhores ações a serem tomadas pela gestão pública, com o objetivo de melhorar a qualidade da água na região, e quais são os parâmetros e pontos de monitoramento essenciais no monitoramento dos corpos hídricos.

A bacia do Alto Iguaçu, localizada na Região Metropolitana de Curitiba, é um caso típico de corpo hídrico que sofre com os impactos gerados pela ocupação humana significativa de cerca de três milhões de habitantes. Adicionalmente a isso, existem trechos bastante industrializados e um sistema de coleta e tratamento de esgoto ineficiente, que combinados com áreas de ocupação irregular expressivas e dinâmicas resultam em um processo de degradação ambiental bastante significativo (COMEC, 2012).

Dentro deste contexto, faz-se necessária uma visão e interpretação mais abrangente dos resultados do monitoramento para que uma gestão dos recursos hídricos possa ser realizada de uma maneira mais eficiente, melhorando a qualidade da água de rios importantes para a população humana, como é o caso da bacia em destaque.

Sendo assim, o objetivo deste trabalho é melhorar o entendimento das alterações de qualidade da água através da aplicação de técnicas de estatística multivariada para identificar os parâmetros mais representativos no conjunto de dados de concentração e de carga da Bacia do Alto Iguaçu na Região Metropolitana de Curitiba.

2. MATERIAIS E MÉTODOS

2.1 Área de estudo

A Bacia do Alto Iguaçu, localizada na Região Metropolitana de Curitiba, é altamente povoada e industrializada e não conta com um completo e eficiente sistema de saneamento básico, o que acaba se refletindo na forma de poluição hídrica.

De acordo com a Coordenação da Região Metropolitana de Curitiba (COMEC), a região envolve atualmente 29 municípios, com uma população que passa dos três milhões de habitantes. Um trecho dessa região, atravessado pelos rios Iraí e Iguaçu é caracterizado pela planície plana, com uma grande extensão de várzeas naturais, que em alguns trechos são intensamente exploradas para extração de areia. Estas regiões, juntamente com as áreas de mananciais, vêm passando por um crescente processo de ocupação irregular, que acaba interferindo ainda mais na qualidade da água da bacia (FRANÇA, 2009).

Para o monitoramento, foram escolhidos estrategicamente seis pontos de forma a abranger toda a região e serem de fácil acesso, com eles, foi possível estudar 107 km do Rio Iguaçu. A Bacia do Alto Iguaçu e a localização dos seis pontos de monitoramentos estão mostrados na Figura 1.

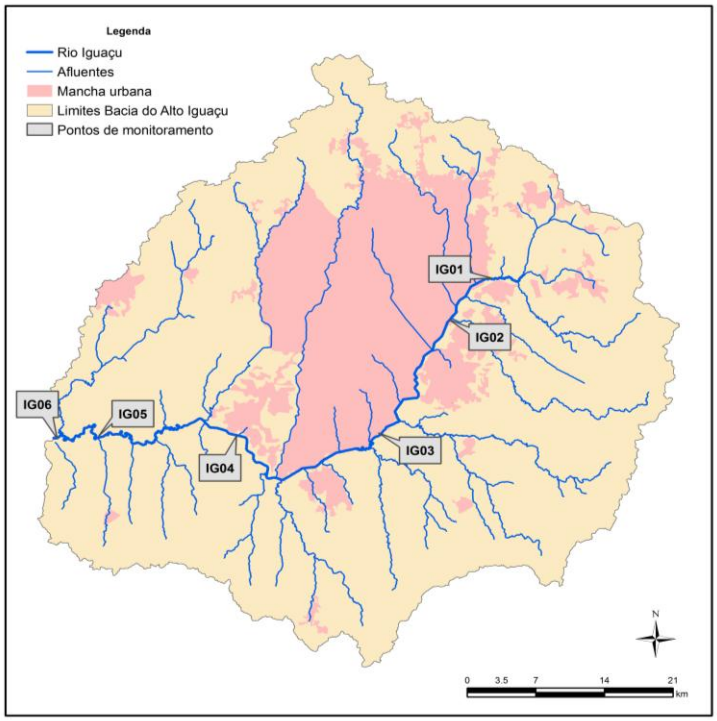

Figura 1 – Mapa da Bacia do Alto Iguaçu e os pontos de monitoramento
Fonte: Adaptado de PORTO et al. (2007)

2.2 Monitoramento da qualidade da água

A avaliação da qualidade da água apresentada no presente estudo está vinculada ao Projeto "Bacias Críticas: Bases Técnicas para a definição de Metas Progressivas para seu Enquadramento e a Integração com os demais Instrumentos de Gestão" (PORTO *et al.*, 2007), desenvolvido em parceria entre a UFPR e a USP, durante o período de 2005 a 2007. Foram utilizados dados de 35 campanhas de monitoramento, que continuou a ser realizado após o término do projeto.

Os parâmetros físico-químicos determinados *in situ* através de sensores foram: oxigênio dissolvido (OD, em mg/L), temperatura da água (ºC), condutividade elétrica (µS/cm), pH e turbidez (NTU). Amostras de água foram coletadas para posterior análise dos seguintes parâmetros de

qualidade da água: demanda bioquímica de oxigênio (DBO), demanda química de oxigênio (DQO), nitrogênio orgânico (N_org), nitrogênio amoniacal (N_NH_3) nitrito, nitrato, fósforo total, sólidos sedimentáveis, sólidos suspensos totais, sólidos dissolvidos totais e carbono orgânico total (COT).

A profundidade Secchi foi determinada pelo disco Secchi e a vazão foi obtida pela curva-chave elaborada pelo ÁGUASPARANÁ tendo como base os níveis da régua telemétrica observados nos dias de coleta.

2.3 Método de Análise Estatística

Com os dados obtidos no monitoramento foi realizada a análise de componentes principais (ACP), a análise fatorial (AF) e a análise de agrupamentos (AA). Para isto, foi utilizado o *R: A language and environment for statistical*, que é um programa livre e gratuito e que serve para análises estatísticas e gráficas.

A ACP é basicamente utilizada para reduzir a dimensão de bancos de dados com muitas variáveis inter-relacionadas através da criação de um conjunto, chamado de componentes principais (CPs), que contém as variáveis mais representativas do problema estudado. Assim como a ACP, a análise fatorial tem como objetivo diminuir bancos de dados com muitas variáveis relacionadas, o que é feito através da criação de fatores, chamados fatores comuns, que são não correlacionados entre si e que supostamente medem aspectos comuns do conjunto de dados. Os fatores são obtidos a partir da estrutura de dependência entre as variáveis e com eles é possível saber o quanto cada fator está associado a cada variável e quanto o conjunto de fatores explica a variabilidade total dos dados (BARROSO e ARTES, 2003). A análise de agrupamentos tem como objetivo descobrir agrupamentos naturais dentro do conjunto de variáveis de forma que os semelhantes sejam agrupados em um mesmo grupo (FERREIRA, 2008; JOHNSON e WICHERN, 1998).

Essas técnicas foram aplicadas aos dados de concentração e de carga referentes à Bacia do Alto Iguaçu a partir de uma matriz contendo os dados de todos os pontos de monitoramento nas linhas e os parâmetros de qualidade de água nas colunas. Optou-se por excluir as linhas com dados faltantes, para evitar interferências no resultado final.

Para avaliação dos dados de concentração, foram consideradas 19 variáveis de qualidade da água (DBO, DQO, COT, turbidez, profundidade Secchi, SDT, SST, SS, condutividade, pH, temperatura, OD, nitrogênio amoniacal, nitrogênio orgânico, nitrato, nitrito, nitrogênio total, fósforo total e vazão) e como observações 87 coletas realizadas nos pontos IG01, IG02, IG03, IG04, IG05 e IG6, resultando, portanto numa matriz de 87 linhas por 19 colunas. Para avaliação dos dados de carga, foram considerados os mesmos 19 parâmetros, porém 84 coletas.

3. RESULTADOS E DISCUSSÃO

Tanto para concentração quanto para carga as variâncias observadas foram de ordens de grandezas bem diferentes, por isso, as técnicas multivariadas foram aplicadas a partir da matriz de correlação das variáveis originais. Obtiveram-se mais correlações significativas ($\geq$7,0) para os dados de carga do que para concentração, 23 e 2, respectivamente, e em comum obteve-se maiores correlações entre nitrogênio amoniacal e nitrogênio total.

3.1 Análise de Componentes Principais

Na ACP, utilizou-se o Critério de Kaiser para definir o número de componentes principais a serem retidas, que são aquelas com autovalores maiores que 1. Dessa forma, para o conjunto de dados relativos à concentração, ficaram retidas 6 CPs que explicaram 75% da variância do conjunto. Para o conjunto de dados de carga, foram retidas 5 CPs que explicaram 82% da variância do conjunto inicial. França (2009), em um estudo similar na mesma bacia, ao analisar concentrações encontrou 5 CPs com 78% da variância explicada. Portanto, os dados de carga se destacaram no sentido de explicar uma maior variância do conjunto de dados.

Os pesos das variáveis originais nas CPs são representados pelos autovetores da matriz de correlação, sendo que as variáveis com maior peso são as mais relevantes na determinação das componentes principais. No presente estudo, devido à maioria dos pesos serem baixos, optou-se por analisar os valores de correlação entre os escores das CPs e as variáveis originais para destacar as variáveis mais importantes. Dessa forma, observou-se como parâmetros importantes para a concentração a DBO, a DQO, o nitrogênio amoniacal e o total, todos encontrados na primeira CP. Já para a carga, 10 variáveis (DBO, DQO, COT, SDT, SST, N_NH_3, N_org, nitrito, nitrogênio total e vazão) ficaram retidas na primeira CP e pH e SS na segunda. Portanto, para carga, dos 19 parâmetros iniciais, 12 explicaram 60% da variância do banco de dados.

Cabe destacar que os parâmetros encontrados estão ligados principalmente à degradação da matéria orgânica, aos sólidos e ao nitrogênio, portanto, a parâmetros que refletem de maneira direta as fontes de poluição ao longo da bacia.

3.2 Análise Fatorial

Antes de realizar a AF é preciso verificar se o conjunto de dados apresenta distribuição normal, no presente estudo isso foi feito através da observação do gráfico de distribuição Qui-Quadrado dos dados de concentração, que se assemelharam a uma reta. Na elaboração do gráfico observou-se que os dados de carga formam uma matriz singular e que, portanto, tem um discriminante nulo, logo, não apresenta matriz inversa e sendo assim não tem distribuição normal. Isto não impede que a AF seja realizada. Porém os fatores não podem ser estimados pelo método da máxima verossimilhança, que era uma das intenções deste estudo. Então para os dados de concentração os fatores foram estimados pela máxima verossimilhança e também pelas componentes principais, enquanto que os de carga só pelas componentes principais. Cabe destacar que a matriz de cargas não é inversível porque representa uma combinação linear das colunas, pois a maioria das colunas foi multiplicada pela mesma série histórica de vazões.

Foi realizado o Teste de Esfericidade de Bartlett que resultou num p-valor nulo e 171 graus de liberdade tanto para concentração quanto para carga. Um outro valor que precisa ser calculado previamente é a Medida de Adequacidade de Kaiser Meyer Olkin que resultou em 0,5 para concentração e para carga não pôde ser calculada pelas condições já expostas anteriormente.

Depois de todas as verificações necessárias a AF foi realizada e, a partir do critério de Kaiser, já utilizado na ACP, reteram-se 6 fatores para os dados de carga que explicaram 75% da variância da amostra para o método das CP's e 62% para o método da máxima verossimilhança. Na análise de carga, 5 fatores explicaram aproximadamente 82% da variância. Os pesos obtidos foram muito baixos, por isso, decidiu-se realizar a rotação varimax.

Após a rotação, para carga encontrou-se como parâmetros com maior peso no Fator 1 o nitrogênio amoniacal e total, no Fator 3 condutividade e pH, no Fator 4 a turbidez e COT e no 5 a temperatura.

Tabela 1 – Composição dos fatores para concentração

	Componente Principal	**Máxima Verossimilhança**
Fator 1	N_NH_3 e nitrogênio total	Turbidez, profundidade, SST, N_Org, Nitrato
Fator 2	Turbidez	Condutividade, pH, fósforo
Fator 3	Condutividade, pH	N_NH_3, N_Org e nitrogênio total
Fator 4	-	DBO e DQO
Fator 5	DBO, SS e nitrato	Nitrito
Fator 6	COT	N_Org

Pelo método da máxima verossimilhança, dos 19 parâmetros iniciais, 16 são capazes de explicar 82% da variância, que é um bom valor. Para componentes principais, a variância explicada é um pouco menor, porém, o número de variáveis é pequeno, 9, comparado ao número original. Com esta grande diferença encontrada para os dois métodos, decidiu-se calcular as comunalidades e observar quais parâmetros ficariam abaixo de |0,5| e que poderiam, portanto ser descartados e a análise ser refeita sem eles. Essa segunda parte foi realizada somente para os dados de concentração.

Dessa forma, a AF final resultou em 3 fatores capazes de explicar 58% da variância da amostra, com apenas 8 parâmetros: DBO, SST, condutividade, pH, nitrogênio amoniacal, nitrogênio orgânico, nitrogênio total e fósforo. Isso não implica que monitorando somente estas variáveis conseguirá se obter um bom diagnóstico da variação da qualidade da água na Bacia do Alto Iguaçu.

França (2009) obteve para o método das componentes principais 11 parâmetros capazes de explicar 87,08% da variância. Os parâmetros em comum nos dois estudos foram nitrogênio amoniacal,nitrogênio orgânico, condutividade, pH, SST e DBO. Alguns parâmetros excluídos inicialmente neste estudo foram obtidos como variáveis importantes no estudo anterior e vice-versa.

3.2 Análise de Agrupamentos

Com o cálculo das correlações buscou-se identificar qual apresentava a maior correlação e portanto era a mais adequada para a realização da AA. Para concentração, a ligação que mostrou-se mais adequada foi a média e para carga a obtida pelo método do centróide. Para análise, separou-se o dendograma obtido em três grupos.

Para concentração, o primeiro continha as coletas 27 e 46 que pertencem ao P2 e P3 respectivamente, essas coletas apresentam dois valores de monitoramento que podem ser considerados anormais. A coleta 27 tem uma DBO de apenas 4,74 mg/L para uma vazão não muito alta, esse ponto é caracterizado justamente por apresentar altos valores para parâmetros que refletem poluição por matéria orgânica, logo o valor encontrado é bastante atípico. Já a coleta 46 apresenta um valor de condutividade muito elevado quando comparado aos demais para o mesmo ponto (561 µS/cm). Assim, concluiu-se que estas coletas foram agrupadas por apresentarem esses valores atípicos para os pontos de monitoraramento que eles representam. O segundo grupo ficou com as coletas 29, 33, 50, 51, 59 e 75. As três primeiras mais a 59 foram coletadas no mesmo dia, o que sugere que a bacia estava numa situação atípica neste dia (baixa ou alta vazã). O terceiro grupo

continha as demais coletas. Deste modo, observa-se que o fator que discriminou os agrupamentos foi a atipicidade das amostras.

No agrupamento das cargas, os dois primeiros grupos só continham amostras dos três pontos de jusante da bacia (P4, P5 e P6). Além disso, a maioria das coletas agrupadas no primeiro grupo apresentavam uma vazão bem mais alta que a normalmente encontrada em seus respectivos pontos de monitoramento. Já no segundo grupo, a maioria das coletas foram realizadas em condições de vazões abaixo das normais, resultando em altas e incomuns concentrações para valores como DBO e COT. O terceiro grupo continha as demais coletas. Assim, o fator que pode ter discriminado os agrupamentos, neste caso, foi a vazão.

Na análise de carga, os agrupamentos permitiram identificar o que se procurava com a realização da AA, uma vez que se obtiveram coletas com vazões elevadas, que normalmente proporcionam padrões de qualidade melhor e de vazões baixas, relacionadas a estados de qualidade de água mais degradados.

França (2009), ao realizar esta análise para concentrações, obteve dois agrupamentos, um contendo todas as coletas realizadas no ponto P1, que representam amostras de água de melhor qualidade e outro com as demais coletas.

4. Conclusão

A análise de parâmetros de qualidade da água em bacias com forte influência antrópica é sempre um desafio, principalmente para garantir a sua interpretação adequada, em especial, no que concerne à aplicação dos instrumentos de gestão de recursos hídricos. Neste estudo, os parâmetros que mais se destacaram estão ligados às frações de nitrogênio, principalmente o amoniacal. Interessante que os parâmetros mais comumente tidos como de grande importância na avaliação da qualidade da água, como DBO e OD não se evidenciaram estatisticamente em muitos dos resultados obtidos.

Apesar de serem comuns estudos estatísticos para diagnósticos de qualidade da água a partir de concentrações, os resultados obtidos neste estudo mostrou que a análise a partir de cargas pode ser mais significativa em algumas técnicas. Na AA, o objetivo de agrupar coletas de melhor e pior qualidade foi superior para carga do que para concentração. Nessa análise ficou clara a influência da variação da vazão na melhora ou piora da qualidade da água na bacia, pois os grupos formados estavam diretamente ligados a uma alta ou baixa vazão. Na ACP, 60% da variância do conjunto de dados pôde ser explicada por 13 parâmetros, portanto, se houvesse a necessidade de diminuir o número de parâmetros a serem monitorados, seria mais fácil estudar a possibilidade de descartar ou diminuir a frequência de monitoramento de alguma dessas seis variáveis que ficaram de fora das duas primeiras componentes principais. O que seria mais arriscado fazer considerando o resultado para concentração, que apontou só 3 variáveis como mais importantes na variação do banco de dados.

A análise fatorial não pôde ser realizada de forma correta para os dados de carga, porém os resultados obtidos para concentração foram bastante satisfatórios, podendo chegar à explicação de 60% da variância com 11 variáveis. Este resultado também poderia ajudar na não consideração de algum parâmetro. Cabe destacar que, analisando conjuntamente os resultados de carga obtidos pela ACP e os de concentração pela AF poder-se-ia tomar decisões muito mais relevantes para a gestão de recursos hídricos do que se fosse levado em conta somente os resultados de uma ou outra análise.

Ao comparar os resultados de concentração com os de França (2009) ficou evidente o quanto o número de dados interfere de forma importante nas técnicas estatísticas e que, portanto, quanto mais dados disponíveis melhor poderá se avaliar os resultados obtidos e mais confiáveis estes se tornarão. Isso mostra a importância de um bom e contínuo monitoramento, que torna possível observar as mudanças ocorridas ao longo dos anos, permitindo identificar os fatores que estão causando a degradação ou recuperação do corpo hídrico monitorado.

Por fim, cabe destacar a necessidade de se complementar o monitoramento para que estudos com uma série histórica mais significativa permita uma análise de qualidade da água mais consistente, ainda mais no que se refere à vazão e consequente análise de carga, que pode trazer uma nova visão sobre qualidade para a gestão de recursos hídricos.

REFERÊNCIAS

BARROSO, L. P.; ARTES, R. *Análise Multivariada: minicurso do 10º Simpósio de Estatística Aplicada à Experimentação Agronômica.* Lavras: Universidade Federal de Lavras, 2003.

COMEC – Coordenação da Região Metropolitana de Curitiba. Disponível em <http://www.comec.pr.gov.br> Último acesso em nov. 2012.

FERREIRA, D. F. *Estatística Multivariada.* Lavras: Universidade Federal de Lavras, 2008.

FRANÇA, M. S. *Análise multivariada dos dados de monitoramento de qualidade de água da Bacia do Alto Iguaçu: uma ferramenta para a gestão dos recursos hídricos.* 150 f. Dissertação (Mestrado em Engenharia) - Departamento de Hidráulica e Saneamento, Universidade Federal do Paraná, Curitiba, 2009.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis.* 4. ed. New Jersey: Prentice Hall, 1998.

PINTO, U.; MAHESWARI, B. L. River health assessment in peri-urban landscapes: An application of multivariate analysis to identify the key variables. *Water Research,* 45,p. 3915-3924, 2011.

PORTO, M. F. A.; FERNANDES, C. V. S.; KNAPIK, H. G.; FRANÇA, M. S.; BRITES, A. P. Z.; MARIN, M. C. F. C.; MACHADO, F. W.; CHELLA, M. R.; SÁ, J. F.; MASINI, L. *Bacias Críticas: Bases Técnicas para a definição de Metas Progressivas para seu Enquadramento e a Integração com os demais Instrumentos de Gestão.* Curitiba: UFPR – Departamento de Hidráulica e Saneamento, 2007. (FINEP/ CT-HIDRO). Projeto concluído.

R CORE TEAM (2012). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051 – 07 – 0, URL http://www.R-project.org/.