Ciência e Natura, Santa Maria v.38 Ed. Especial-
IX Workshop Brasileiro de Micrometeorologia, 2016, p. 498 – 503
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

Controle de qualidade em dados de alta frequência no projeto ATTO

Quality control of high frequency data in the ATTO project

Einara Zahn* e Nelson Dias†

Resumo

O controle de qualidade em dados de turbulência é uma etapa essencial antes de realizar qualquer tipo de análise. No entanto, não existem testes padronizados a serem seguidos, ficando a escolha dos critérios adotados a cargo de cada pesquisador. Neste contexto, o principal objetivo deste estudo é realizar um controle de qualidade em dados de alta frequência (10 Hz) medidos em dois níveis sobre uma floresta Amazônica (39,4 e 81,6 m, sendo a altura média da floresta aproximadamente 40 m). Inicialmente foram removidos os picos; na sequência, a variação do desvio padrão de cada série foi avaliada a fim de verificar casos de mau funcionamento do sensor ou baixos níveis de turbulência. Por fim, para detectar tendências nas séries e casos de não estacionariedade, o Reverse Arrangement Test foi empregado seguido de uma análise da variação temporal (diferença entre valores máximos e mínimos) de cada variável. Após todas as etapas do controle de qualidade, restaram 15,8% dos dados medidos em 81,6 m e 40,7% dos dados medidos em 39,4 m. Esses resultados mostram o grande número de séries inconsistentes, destacando a importância do controle de qualidade a fim de evitar que tais séries contaminem futuras análises.

Palavras-chave: *Turbulência, controle de qualidade, Reverse Arrangement Test.*

Abstract

Control quality in turbulence data is an essential step before carrying on any analysis. However, there are not standard tests to follow, and the evaluation criteria are left to the researcher. In this context, the main goal of this study is perform a quality control of high frequency data (10 Hz) measured at two levels above an Amazonian forest (39.4 and 81.6 m, where the average forest height is approximately 40 m). First of all, spikes were removed; in the following, changes in standard deviation in each variable were evaluated with the intention to detect sensor malfunction or low levels of turbulence. Finally, to detect trends and unsteady cases the Reverse Arrangement Test was applied, followed by an analysis of the temporal variation of each variable (difference between maximum and minimum values). After all control quality procedures, 15.8% of the data measured at 81.6 m and 40.7% of the data measured at 39.4 m remained. These results show the high quantity of inconsistent runs, highlighting the importance of the control quality in order to avoid that they contaminate the ensuing analysis.

Keywords: *: Turbulence, quality control, Reverse Arrangement Test.*

*Programa de Pós-Graduação em Engenharia Ambiental/UFPR
†Departamento de Engenharia Ambiental/UFPR

1 Introdução

Em micrometeorologia, um bom controle de qualidade dos dados de turbulência é uma etapa fundamental antes de se iniciar qualquer forma de análise. Quando este procedimento não é realizado de forma adequada, eventuais falhas nas séries temporais que não tenham sido detectadas e/ou corrigidas podem ser repassadas para as análises, contaminando os resultados e posteriormente prejudicando sua interpretação. Além disso, o controle de qualidade é essencial para assegurar que as hipóteses dos métodos empregados nas análises sejam satisfeitas. Exemplos são o Método das Covariâncias Turbulentas (MCT) e a Teoria de Similaridade de Monin Obukhov (TSMO); ambos pressupõem que o escoamento é estacionário na média, de forma que para obter um bom desempenho dos dois métodos faz-se necessário adotar critérios capazes de investigar casos em que esta hipótese não é seguida.

O controle de qualidade pode ser realizado em diferentes passos, sendo que o primeiro visa eliminar dados inverossímeis, tais como falhas ou picos nas séries (Foken, 2008). Os picos são caracterizados como flutuações de curta duração e de grande amplitude, que podem resultar do ruído aleatório na eletrônica dos sensores. Quando não são removidos das séries temporais, podem causar problemas em procedimentos simples, como no MCT. Neste caso, por exemplo, quando ocorrem picos correlacionados de temperatura e velocidade vertical, ambos medidos por um anemômetro sônico, pode haver contaminação do fluxo de calor ou da variância das séries das duas variáveis (Foken et al., 2005). Segundo Vickers e Mahrt (1997), um método de detecção dos picos pode ser formulado em termos de um número de desvios-padrão a partir da média ou alguma outra propriedade estatística. Detectado este pico, ele pode ser substituido a partir de uma interpolação linear. Este é o único procedimento que modifica o conjunto de dados, sendo que os procedimentos subsequentes do controle de qualidade são realizados sobre os dados já corrigidos.

Além do ruído adicionado aos dados pelos sensores, outros problemas com os instrumentos micrometeorológicos podem ser detectados a partir da comparação da variância (ou desvio padrão) com valores-limite préestabelecidos (Foken et al., 2005). Neste caso, estas estatísticas podem ser comparadas numa sequência de intervalos dentro da série temporal a fim de detectar problemas isolados. Um intervalo cuja variância das flutuações aproxime-se de zero, por exemplo, pode ser um indício de um período de falha do instrumento. Por outro lado, em alguns casos pode ser que a série tenha sido medida durante o período da noite, quando a estabilidade da atmosfera e os baixos níveis de turbulência fazem com que as variâncias tornem-se menores; neste caso, apesar da baixa variância não estar relacionada com erro de medição do sensor, estes baixos valores vão contra uma das premissas do MCT, que é o completo desenvolvimento da turbulência (Foken et al., 2005).

Os exemplos citados acima são apenas algumas das consequências causadas pela inconsistência dos dados medidos, e ressaltam a importância de escolher bons critéros para o controle de qualidade. No entanto, apesar de sua grande importância, não existe um procedimento padrão a ser seguido, sendo os métodos adotados de acordo com cada pesquisador. Um dos pontos destacados por Foken et al. (2005) é importância de realizar a análise visual durante as etapas do controle de qualidade, uma vez que isto permite confirmar se as séries foram corretamente detectadas.

O presente estudo visa empregar um controle de qualidade sobre um conjunto de dados de turbulência medidos sobre a floresta Amazônica no contexto do projeto ATTO. A primeira inspeção visual dos gráficos revelou a existência de muitas séries que fugiam das premissas do MCT e da TSMO, principalmente as séries medidas muito acima da floresta ($\sim$ 82 m). Dessa forma, o principal objetivo é detectar falhas ou condições adversas sobre os dados, e quando possível corrigir os dados, como a substituição dos picos, de modo que os dados finais sejam adequados ao uso destes dois métodos.

2 Materiais e Métodos

O conjunto de dados empregados nesta análise (cedidos pelo INPA – Instituto Nacional de Pesquisas na Amazônia) é parte das primeiras medições realizadas pelo projeto *Amazon Tall Tower Observatory — ATTO*, ou Observatório de Torre Alta da Amazônia, em Português. Estes dados piloto correspondem a medições realizadas nos níveis 39,4 e 81,6 m acima da floresta, cuja altura média é de 40 m, durante o mês de Abril de 2012.

As variáveis empregadas são as três componentes da velocidade do vento u, v e w, medidos por um anemômetro sônico 3D (R3, Gill Instruments Ltd.; CSAT3, Campbell Scientific Inc.), a temperatura virtual do ar θ_v, também medida pelo anemômetro sônico, e as densidades de dióxido de carbono, ρ_c, e vapor d'água, ρ_q, medidas por um IRGA (modelo LI7500A, LI-COR Inc). Os dados foram amostrados em uma frequência de 10 Hz e agrupados em blocos de 30 min, o que corresponde a 18000 medições por bloco. As medições ocorreram durante o mês de Abril, e o conjunto de dados resultante possui 1414 blocos medidos em 39,4 m e 1444 blocos medidos em 81,6 m. A seguir são discutidos os métodos adotados no controle de qualidade.

2.1 Remoção dos picos

O método de remoção de picos foi baseado no método proposto por (Vickers e Mahrt, 1997). Inicialmente as séries de cada variável foram separadas em janelas de dois minutos; posteriormente a média e o desvio padrão de cada janela foi calculada. Ao percorrer cada ponto da série, tal ponto foi considerado um pico quando seu valor absoluto era maior do que 3,5 vezes o desvio padrão a partir da média da janela à qual pertence. Uma vez detectado, o pico é substituído através de uma interpolação linear. No entanto, quando quatro ou mais pontos consecutivos são detectados, eles não são considerados picos, não sendo substituídos. Além disso, quando o número de picos é superior a 1% do total de dados, o bloco é descartado.

2.2 Análise do desvio padrão

O segundo método do controle de qualidade consistiu em verificar o desvio padrão a cada dois minutos (1200 pontos) ao longo das séries. Inicialmente, cada série teve a média móvel calculada (com janela de 900 pontos); posteriormente foram subtraídas as séries originais das médias móveis, resultando em uma nova série, agora composta por flutuações. Usando a temperatura como exemplo, sua série original, θ_v(t), menos a média móvel da série, θ_{v-mov}(t), resulta em θ'_v(t) = θ_v(t) - θ_{v-mov}(t). Vale lembrar que, uma vez que a janela adotada para a média móvel foi de 900 pontos, a série de flutuações possui 17101 pontos (18000 – 899). Na sequência, cada série temporal resultante foi separada em intervalos de 1200 pontos, resultando em 14 intervalos de 2 minutos (o último intervalo possui apenas 301 pontos e não é incluído na análise).

Com base na precisão de amostragem do sensor de cada variável, e após a realização de testes, valores-limite foram estipulados para o desvio padrão. Dessa forma, sempre que o desvio padrão de uma janela mostrou-se inferior ao respectivo valor-limite, para uma determinada variável, o bloco a qual tal série pertence foi descartado. A Tabela 1 exibe os valores-limite (λ) adotados para as seguintes variáveis: velocidade vertical, λ_w, velocidade longitudinal e transversal, λ_u e λ_v, respectivamente, temperatura, λ_{θ_v}, dióxido de carbono, λ_{ρ_c}, e vapor d'água, λ_{ρ_q}. As unidades são as mesmas fornecidas pelos sensores.

2.3 *Reverse Arrangement Test*

Seja uma sequência de N observações de uma variável aleatória x, onde as observações são denotadas por x_i, $i = 1,2,3,...,N$. Agora, contando o número de vezes em que $x_i > x_j$ para $i < j$, cada desigualdade é chamada um arranjo inverso, e o número total de arranjos é denotado por A. Da série temporal, $x_1, x_2,...,x_N$, define-se

$$h_{ij} = \begin{cases} 1 & \text{se } x_i > x_j, \\ 0 & \text{caso contrário.} \end{cases}$$

Então,

$$A = \sum_{i=1}^{N-1} A_i,$$

$$A_i = \sum_{j=i+1}^{N} h_{ij}.$$

Após calculados estes índices, e dado um nível de significância α, a região de aceitação para esta hipótese será $[A_{N;1-\alpha/2} < A \leq A_{N;\alpha/2}]$, cujos valores são tabelados (Bendat e Piersol, 1986, p. 97). Na presente análise, os parâmetros adotados foram N = 50, em que a série foi agora separada em 50 intrevalos, tendo sido a média de cada intervalo empregada na análise, e α = 0,05 como nível de significância. Com estes parâmetros o limite torna-se $[495 < A \leq 729]$.

O último método adotado foi de cunho subjetivo, e foi incluído a fim de detectar séries de escalares com variações muito grandes, ou séries não estacionárias, dentro do intervalo de 30 minutos. Para tanto, obteve-se a diferença entre o maior e o menor valor da média móvel de cada série temporal, tendo sido tal série descartada quando esta diferença era maior que um valor pré-estabelecido para o escalar em questão. Estes valores foram escolhidos com base na inspeção visual dos gráficos, em que definiu-se: $\Delta\rho_c$ = 3 mmol m^{-3}, $\Delta\theta_v$ = 1,7 ^{o}C e $\Delta\rho_q$ = 200 mmol m^{-3}.

3 Resultados e Discussão

A primeira etapa do controle de qualidade incluiu a exclusão dos blocos com menos de 30 minutos (12 blocos em 39,4 e 8 em 81,6 m), a remoção/substituição dos picos e a análise do desvio padrão. Dos 3 métodos, o mais impactante foi a análise do desvio padrão, uma vez que reduziu drasticamente a quantidade de blocos.

A Figura 1 exibe uma série temporal de temperatura excluída com base na análise do desvio padrão. Dos 14 intervalos que foram comparados com λ_{θ_v}, 6 ficaram abaixo deste valor-limite, sendo que o menor valor foi σ_{θ_v} = 0,0174 ^{o}C no segundo intervalo da série. Conforme mencionado anteriormente, esta parte do controle de qualidade reduziu drasticamente a quantidade de blocos, tendo restado nesta etapa apenas 21,5% dos blocos medidos em 81,6 m e 50,2% dos blocos medidos em 39,4 m.

Após analisar visualmente as séries remanescentes, notou-se ainda a existência de séries fortemente não estacionárias, o que poderia comprometer os resultados

Tabela 1: Valores-limite adotados para o desvio-padrão das variáveis analisadas.

λ_w	λ_u	λ_v	λ_{θ_v}	λ_{ρ_c}	λ_{ρ_q}
0,005 m s^{-1}	0,008 m s^{-1}	0,008 m s^{-1}	0,04 ° C	0,01 mmol m^{-3}	1 mmol m^{-3}

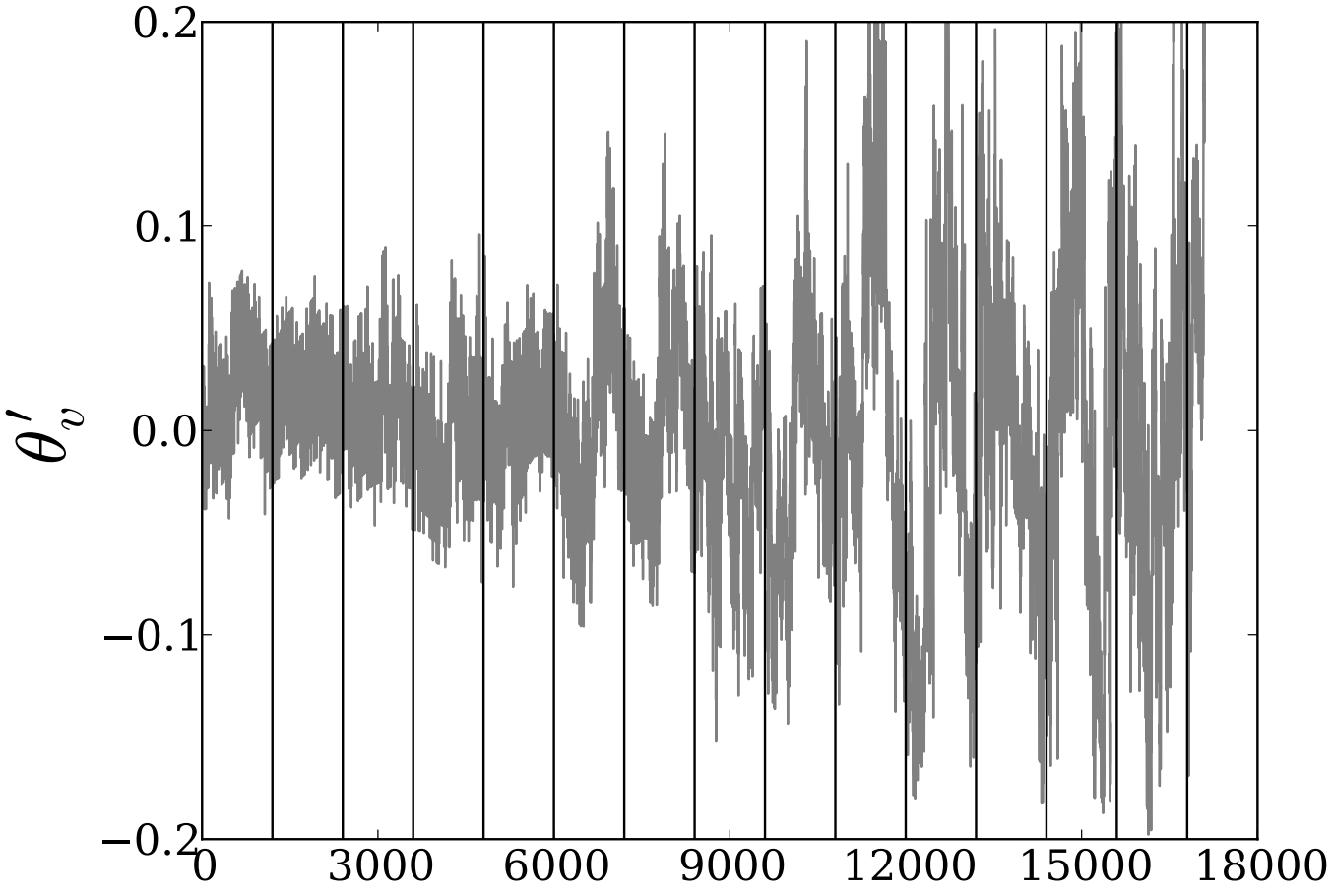

Figura 1: Série temporal de temperatura (medida em 39,4 m, dia 01/04/2012 às 18h00) detectada pelo método de análise do desvio padrão. O segundo intervalo resultou em σ_{θ_v} = 0,0174 ºC $< \lambda_{\theta_v}$ (= 0,04 ºC).

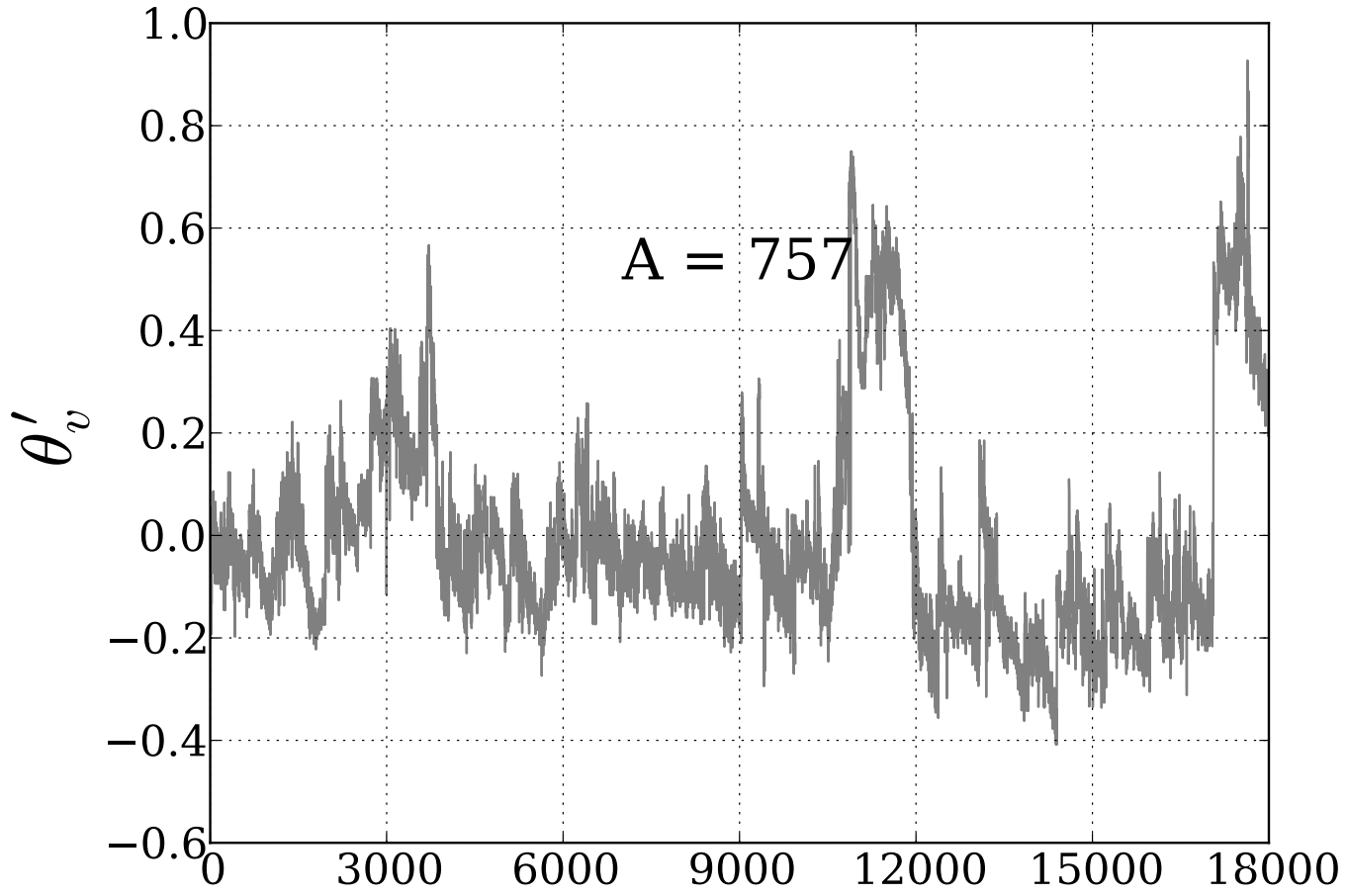

Figura 2: Série de temperatura (medida em 39,4 m, dia 03/04/2012 às 08h30) detectada pelo método *Reverse Arrangement Test*. Neste exemplo A = 757, superior ao limite superior de validade do teste (= 729).

finais de futuras análises realizadas com tais dados. Por este motivo, na sequência o método *Reverse Arrangement Test* foi empregado a fim de detectar séries com algum tipo de tendência. Tendo sido escolhidos N = 50 e α = 0,05, foram excluídos blocos em que alguma das séries apresentou o parâmetro A $<$ 495 ou $\geqslant$ 729.

A Figura 2 exibe um caso para temperatura em que A = 757, fazendo com que o bloco fosse removido pelo método. Neste caso as flutuações foram obtidas a partir da remoção da tendência linear, tendo sido ajustada uma função do tipo $\theta_{v-linear}$(t) = a + bt à série, e na sequência obteve-se a flutuação fazendo-se θ'_v(t) = θ_v(t) − $\theta_{v-linear}$. Analisando esta série, pode-se verificar uma grande variabilidade na temperatura, em que mudanças abruptas no sinal do escalar ocorrem, fazendo com que muitos arranjos inversos fossem detectados.

Por fim, as séries que permaneceram passaram ainda por outro controle, sendo desta vez o teste subjetivo. Nesta etapa detectou-se muitas séries de vapor d'água, ρ_q, tendo sido encontradas variações (diferença entre o maior e o menor valor da média móvel) de mais 1000 mmol m^{-3} (sendo o limite de 200 mmol m^{-3}). Séries de dióxido de carbono, ρ_c, também exibiram grande variabilidade, chegando a Δ_{ρ_c} = 34 mmol m^{-3}. Estas grandes diferenças aparentam ser algum tipo de influência externa, que não as condições meteorológicas, ou ainda erro dos sensores, uma vez que ambos os escalares são medidos pelo mesmo equipamento; isso se evidência quando verificadas as séries de temperatura, que nos poucos casos detectados pelo teste, não ultrapassou diferenças de 3,5 ºC. No entanto, estes fatores não foram analisados neste estudo, dessa forma uma explicação plausível para estas variações não pôde ser obtida.

Após esta fase final do controle, restaram 15,8% dos blocos medidos em 81,6 m e 40,7% dos blocos medidos em 39,4 m, o que destaca o grande número de séries temporais inconsistentes medidas acima do dossel vegetal em relação às medidas no topo.

4 Conclusões

O principal objetivo desta análise foi realizar um controle de qualidade sobre um conjunto de dados de turbulência, de forma que os dados resultantes fossem próprios para futuras análises, como o Método das Covariâncias Turbulentas e a Teoria de Similaridade de Monin Obukhov. Após a remoção dos picos, foram adotados 3 testes: análise do desvio padrão, *Reverse Arrangement Test* e uma análise das variações da série.

Logo no início, o primeiro teste reduziu cerca de 79% dos blocos em 81,6 m e 50% em 39,4 m; isso confirmou a hipótese inicial, em que a partir da inspeção visual dos gráficos, notou-se a existência de um maior número de séries inconsistentes no nível mais alto. Outro ponto relevante desta análise foi a possível influência dos dados medidos pelo IRGA, que mede vapor d'água e dióxido de carbono. Muitas blocos foram removidos devido à existência de grandes variações nestas duas séries, chegando a $\Delta_{\rho_q} > 1000$ mmol m^{-3} e $\Delta_{\rho_c} > 30$ mmol m^{-3}. No entanto, como o objetivo deste estudo foi apenas conduzir o controle de qualidade, possíveis influências externas sobre as séries não foram investigadas.

Por fim, restaram aproximadamente 15,8% dos blocos originais em 81,6 m e 40,7% em 39,4 m, destacando a importância do controle de qualidade devido ao grande número de blocos que contaminariam os resultados caso fossem mantidos.

Agradecimentos

Ao apoio fornecido, agradecemos à Sociedade Max Planck e ao Instituto Nacional de Pesquisas da Amazonia. Agradecemos também ao Ministério Federal de Educação e Pesquisa da Alemanha (BMBF contrato 01LB1001A) e ao Ministério da Ciência, Tecnologia e Inovação (MCTI/FINEP contrato 01.11.01248.00), assim como à Universidade Estadual do Amazonas (UEA), FAPEAM, LBA/INPA e SDS/CEUC/RDS-Uatumã e à CAPES pela bolsa de mestrado de Einara Zahn. Por fim, agradecemos A. Manzi, A. Araújo e Marta Sá pela cessão dos dados utilizados neste estudo.

Referências

Bendat, J. S., Piersol, A. G. (1986). *Random Data*, 2º edn. John Wiley & Sons.

Foken, T. (2008). *Micrometeorology*, 1º edn. Springer-Verlag Berlin Heidelberg.

Foken, T., Gockede, M., Mauder, M., Mahrt, L., Amiro, B., Munger, W. (2005). Post-field data quality control. Em: Lee, X., Massman, W., Law, B. (Eds) *Handbook of Micrometeorology: A Guide for Surface Flux Measurement and Analysis*, Springer Netherlands.

Vickers, D., Mahrt, L. (1997). Quality control and flux sampling problems for tower and aircraft data. *Journal of Atmospheric and Oceanic Technology*, *14*, 512–526.