**MÉTODO DE DETECÇÃO DE OUTLIERS BASEADO NA AMPLITUDE DO CICLO DIÁRIO MEDIANO MÓVEL**

Graduando em Meteorologia UFSM1, Jônatan Dupont Tatsch2

1 Graduando em Meteorologia, UFSM

2 Departamento de Física, UFSM

**Resumo**

**Abstract**

1. **Introdução**

Em medições de fluxos com os sistemas de covariância de vórtices turbulentos (ECOS) é inerente a ocorrência de pontos aberrantes (*outliers*) tanto nas séries temporais de alta-frequência (10 Hz) quanto de baixa-frequência (média de 30 min). Os valores instantâneos de alta frequência são geralmente removidos antes do cálculo dos fluxos médios de 30 minutos. Entretanto, valores aberrantes podem ainda ocorrer nas séries médias de 30 min por diferentes razões como, mudanças na área de abrangência da medida (*footprint*), mudanças nas condições de turbulência, problemas instrumentais (gotas de água no anemômetro sônico ou no analisador de gás infravermelho) e do sistema de alimentação de energia. Diferenças nos métodos usados para identificação destes dados ou na forma como são implementados nos softwares de processamento de dados de ECOS (e.g, EddyPro e EdiRe) podem levar a incertezas significativas na comparação da magnitude de fluxos (água, energia e CO2) entre locais e tipos de ecossistemas (Fratini e Mauder 2014 – ref de sua dissertação), além a qualidade preenchimento de falhas. Neste trabalho propõe-se um método para detecção de dados aberrantes nas séries de fluxos médios de 30 min baseado na comparação entre na derivada local do valor de fluxo e a amplitude do ciclo diário mediano móvel numa janela temporal de N dias antes e após a meia-hora considerada. O método difere daqueles comumente disponíveis nos softwares de processamento de dados de fluxo pelo fato de ser fisicamente baseado e não apenas estatístico. O método usa estatísticas robustas na sua implementação, tendo a mediana e o desvio padrão mediano absoluto como estimativas dtimativas de tendência central e dispersão dos dados. A frequência de dados detectados pelo método foi comparada as classes de qualidade dos dados de fluxos (Foken 2003).

1. **Materiais e Métodos**

Os dados deste trabalho são de uma Torre de Fluxo (29°43’26,76” S, 53°34’92” W, altitude de 88 m acima do nível médio do mar) da Rede SULFLUX, e localiza-se em uma área de vegetação rasteira, característica do bioma Pampa, da Universidade Federal de Santa Maria (UFSM) no município de Santa Maria, RS. Mais detalhes sobre o sítio experimental encontram-se em Valente (2016).

A medição dos escalares em alta frequência (10 Hz) foi realizada por um Sistema Eddy Covariance, composto por um anemômetro sônico tridimensional (Wind Master Pro), e um analisador de gás infravermelho caminho fechado (LI-7200) em funcionamento até 15/12/2014, e após substituído por outro de caminho aberto (LI-7500), instalados a uma altura de 3 (três) metros na torre que possui 30 (trinta) metros de comprimento.

O cálculo dos fluxos médios de 30 minutos foi realizado pelo software EddyPro® utilizando o fluxograma de processamento de dados que incluiu ….cite quais as opções de configuração foram usadas conforme sua dissertação (não precisa das refs), apenas descreva o corresponde "I\_R2\_MM\_30s\_MCP" da sua dissertação (qual rotação de coordenadas, remoção de tendência, cte de tempo e etc).

2.1 Método de detecção de dados espúrios

Na análise da série temporal dos fluxos de energia na base média de 30 minutos, ficou constatado visualmente a existência de valores fisicamente inconsistentes, isto é, fora da variação esperada para a série de dados. Nesse caso, para a detecção automática destes valores aberrantes, foi elaborado um método simples baseado na comparação entre a diferença do valor de cada meia hora em relação a meia hora anterior () e a amplitude do ciclo diário mediano móvel (CDMM) com uma janela de N dias antes e após a meia hora considerada, da seguinte forma:

(1)

Em que, na equação (1) é o fluxo na meia hora , e indica o elemento da série temporal de tamanho . Considerando que o CDMM é representado pelo vetor , onde a amplitude de é obtida como na equação (2):

(2)

Dessa forma, se o valor é marcado como sendo spike. A partir de testes realizados alterando-se o N, chegou-se a conclusão que N = 60 dias foi suficiente para detectar a maioria dos spikes. Na realidade, o parâmetro que teve o maior impacto na marcação dos dados foi o fator de amortecimento, pois este fator a medida que fica abaixo de 1, aumenta a quantidade de dados marcados como spikes. (Tu achas que aqui tem que explicar sobre o fator de amortecimento? É esse fator o responsável pelo sombreado do gráfico? - SIM DEVE FALAR E QUAL FOI USADO- coloque como estava na dossertação mas adicione isso - verifique formatação! Aqui no meu dropbox abri com o word e as formas estavam em posições estranhas - verifica no pdf gerado que elas estejam ok.).

2.2 Classes de controle de qualidade dos fluxos de energia

As classes de controle de qualidade, de acordo com Foken et al. (2004), variam de 1 a 9 e são determinadas por critérios de estacionariedade e de turbulência desenvolvida, em que a classe 1 indica o fluxo de melhor qualidade, mas a classe 9 o de pior qualidade. A classificação de qualidade é independente para cada fluxo, p.ex.: em uma meia-hora o rótulo de qualidade do Fluxo *H* pode ser 1 (qc\_H = 1) e a de *LE* pode ser 5 (qc\_LE = 5), logo a estimativa de *H* nessa meia-hora é de qualidade superior à de *LE*. As classes podem ser reagrupadas em 3 classes, conforme Foken (2008): (1) com fluxo de melhor qualidade e inclui as classes 1-3; (2) com fluxos de qualidade boa e inclui as classes 4-6; (3) com fluxos de qualidade baixa e inclui as classes 7-8.

1. **Resultados e Discussões**

No período analisado, o que totaliza 31536 pontos médios de meia hora, verificou-se que houve uma quantidade maior de dados marcados como spikes em H (556 pontos) do que em LE (apenas 145 pontos). Isso pode ser percebido observando a figura 1.

Em relação a H, isso representa aproximadamente 2% do total de dados, sendo que destes, apenas 105 pontos foram detectados quando H > 0, e a maioria (451 pontos) ocorreu na condição de H negativo (H < 0). Por outro lado, os spikes de LE representam somente 0,5% do todo e a maior quantidade foi marcada no período diurno.

|  |
| --- |
|  |
|  |

Figura 1 – Variação sazonal do (a) Fluxo de calor sensível (H) e (b) Fluxo de calor latente (LE). Os pontos em vermelho são aqueles marcados como outliers de acordo com o método de detecção descrito neste trabalho.

Quando se faz a comparação entre os dados marcados como aberrantes (qc\_rcr\_H ou qc\_rcr\_LE idênticos a 1) e a qualidades dos dados (qc\_H ou qc\_LE, variando de 1 a 9), a tabela 1 nos informa que quase 70% dos pontos marcados como espúrios, tanto de H quanto de LE foram rotulados pelo sistema de classificação de Foken (2003) como de alta qualidade (classes 1 a 3).

Nesse sentido, é importante dizer que que embora eles tenham sido considerados como de alta qualidade, em uma inspeção visual seriam facilmente apontados como pontos aberrantes.

Tabela 1 – Quantidade de dados marcados como pontos aberrantes suspeitos (qc\_rcr\_H ou qc\_rcr\_LE), de acordo com o rótulo de qualidade dos fluxos (qc\_H ou qc\_LE).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| qc\_rcr\_H | qc\_H | Total | qc\_rcr\_LE | qc\_LE | Total |
| 1 | 1 | 303 | 1 | 1 | 85 |
| 2 | 65 | 2 | 15 |
| 3 | 0 | 3 | 0 |
| 4 | 82 | 4 | 15 |
| 5 | 25 | 5 | 6 |
| 6 | 20 | 6 | 8 |
| 7 | 41 | 7 | 11 |
| 8 | 13 | 8 | 1 |
| 9 | 7 | 9 | 4 |

1. **Conclusões**

Destacar as % de dados detectados e enfatizar que mesmo dados de boa qualidade pelo sistema do Foken podem ocorrer a maior ocorrência de dados aberrantes. E que este sistema de qualidade portanto não elimina a ocorrência de dados espúrios, o que pode propagar a incerteza para os fluxos preenchidos por métodos de preenchimento de falhas quando eles não são eliminados adequadamente por técnicas efetivas e fisicamente baseadas de detecção de dados aberrantes.

**Referências Bibliográficas**

Inserir refs