Análise baseada em fractais para identificação de mudanças de tendências em múltiplas séries climáticas

Santiago Augusto Nunes[1], **Luciana A. S. Romani** [1,2], **Ana M. H. Avila** [3], **Caetano Traina Jr** [1], **Elaine P. M. de Sousa** [1], **Agma J. M. Traina** [1]

[1]ICMC - Universidade de São Paulo - São Carlos - Brasil
[2]Embrapa Informática Agropecuária - Campinas - Brasil
[3]Cepagri - Universidade Estadual de Campinas - Campinas - Brasil

bynael@grad.icmc.usp.br, alvim@icmc.usp.br, avila@cpa.unicamp.br,
{caetano, parros, agma}@icmc.usp.br

Abstract. *In the last few decades, huge amounts of climate data have been gathered and stored by several institutions. The analysis of these data has become an important task due to worldwide climate changes and the consequent social and economic effects. In this work, we propose an approach to analyze multiple climate time series in order to identify intrinsic temporal patterns and trend changes. By dealing with multiple time series as multidimensional data streams and combining fractal-based analysis with clustering, we can integrate different climate variables and discover general behavior changes over time.*

Resumo. *Nas últimas décadas, grandes volumes de dados climáticos têm sido coletados e armazenados por diversas instituições. A análise destes dados é uma tarefa cada vez mais importante, considerando os cenários de mudanças climáticas globais e os consequentes efeitos sociais e econômicos. Neste trabalho, é apresentada uma abordagem para análise integrada de múltiplas séries climáticas, com o objetivo de identificar padrões temporais intrínsecos aos dados e mudanças de tendências. O processo de análise proposto trata múltiplas séries temporais como data streams multidimensionais, e combina técnicas baseadas em fractais para o monitoramento de data streams com agrupamento de dados.*

1. Introdução

Nas últimas décadas, os resultados dos trabalhos em Climatologia têm mostrado que o clima do planeta está mudando, com aumento das temperaturas e mudanças na distribuição de chuvas. De acordo com a Organização Mundial de Meteorologia, clima pode ser definido como a descrição média das condições atmosféricas inferidas a partir de observações contínuas durante um período de pelo menos 30 anos [Zhai et al. 2005]. Além disso, também devem ser considerados os desvios da média que caracterizam a variabilidade natural e consequentemente os fenômenos extremos.

As primeiras pesquisas sobre clima tinham por objetivo estudar a influência das condições climáticas sobre as pessoas, seus costumes, hábitos, distribuição geográfica e atividades como agricultura. Com o aumento da população mundial, essas pesquisas intensificaram-se [Ayoade 1996] e atualmente os meteorologistas têm analisado grandes

quantidades de dados provenientes de diferentes sensores e de saídas de modelos de previsão, a fim de entender condições extremas e anomalias climáticas.

Os resultados de várias análises mostram que eventos extremos vêm aumentando em frequência, duração e intensidade nos últimos anos [Alexander et al. 2006]. Consequentemente, esse aumento pode levar à intensificação de desastres naturais. Chuvas fortes em um único dia ou vários dias com chuvas muito acima da média podem causar inundações que trazem problemas tanto para ambientes urbanos quanto rurais. Nesse cenário, é importante entender as tendências dos fenômenos extremos, para que seja possível uma preparação para tais situações adversas, criando condições de mitigar os problemas e tomar decisões em tempo hábil.

Dados climáticos de estações de superfície, sensores remotos, radares meteorológicos e de outros sensores, têm aumentado muito, gerando semanalmente grandes volumes de dados. Além disso, os modelos de mudanças climáticas têm sido processados para diferentes cenários, contribuindo para o aumento da quantidade de dados climáticos na ordem de terabytes por simulação. A análise de todos esses dados torna-se, portanto, cada vez mais desafiadora para os pesquisadores. Os meteorologistas têm usado métodos estatísticos bem conhecidos, como análise de componentes principais, mas ainda há necessidade do desenvolvimento de novas técnicas para recuperar informação relevante e extrair padrões interessantes dos conjuntos de dados, analisando as variáveis de maneira integrada e mostrando a dependência entre elas em séries de dados cada vez mais longas. Neste contexto, duas tarefas relevantes são:

1. Analisar eficientemente múltiplas séries temporais climáticas para encontrar padrões e alterações de tendências nos dados.
2. Identificar extremos climáticos que possam ser indícios de mudanças no clima de uma determinada região.

Este artigo propõe um processo de análise de séries temporais climáticas que tem por objetivo oferecer suporte a essas tarefas. Na abordagem proposta, múltiplas séries temporais são tratadas como uma *data stream* multidimensional, isto é, cada série é considerada um atributo da *data stream*. Com isso, é possível analisar de maneira integrada várias séries climáticas longas, cada uma representando uma variável climática.

O processo de análise proposto utiliza técnicas de monitoramento de *data streams* baseado em conceitos da teoria de fractais, o que permite a identificação de padrões e possíveis mudanças de tendência considerando também o relacionamento intrínseco existente entre séries definidas por variáveis climáticas distintas. Além disso, essas técnicas são combinadas com métodos de agrupamento de dados, com o objetivo de encontrar padrões, tanto similares quanto distintos, que são revelados em janelas temporais de granularidades diferentes. Os resultados iniciais obtidos nos experimentos realizados com séries climáticas reais indicam que a abordagem proposta pode ser uma ferramenta útil para os especialistas do domínio, ou seja, para os meteorologistas.

Na Seção 2 são sintetizados conceitos da teoria de fractais e sua aplicação em análise de *data streams*, formando a base deste trabalho. Na Seção 3 é apresentada a abordagem proposta. Os experimentos e resultados são descritos na Seção 4. Finalmente, na Seção 5 são apresentados a conclusão e trabalhos futuros.

2. Fractais e a Análise de *Data Streams*

Um fractal é definido pela propriedade de auto-similaridade, ou seja, é um objeto que apresenta as mesmas características para diferentes variações em escala e tamanho. Portanto, partes do fractal são similares, exata ou estatisticamente, ao fractal como um todo [Schroeder 1991]. Dentre os conceitos da teoria de fractais, a dimensão fractal é um dos mais relevantes para as áreas de banco de dados e mineração de dados, pois provê uma estimativa da dimensionalidade intrínseca de um conjunto de dados, isto é, da quantidade de informação que o conjunto efetivamente representa, independente da dimensão E do espaço em que está definido. A dimensionalidade intrínseca representa o comportamento não-uniforme do conjunto de dados, indicando a existência ou não de correlações entre seus atributos [Faloutsos and Kamel 1994, Traina et al. 2005]. Por exemplo, um conjunto de pontos distribuídos ao longo de uma linha num espaço tridimensional (E=3) tem seus 3 atributos linearmente correlacionados e dimensionalidade intrínseca igual a 1.

A dimensão fractal de conjuntos estatisticamente auto-similares, como é o caso de grande parte dos conjuntos de dados reais, pode ser determinada pela Dimensão de Correlação Fractal D_2. Para conjuntos multidimensionais definidos em um espaço de E dimensões, o valor de D_2 pode ser calculado usando a Equação 1 [Schroeder 1991] e o método *Box-Counting*, onde: r é o tamanho do lado das células em um hiper-reticulado que divide o espaço de endereçamento do conjunto de dados, e $C_{r,i}$ é a contagem de pontos dentro da célula i.

$$D_2 \equiv \frac{\partial log(\sum_i C_{r,i}^2)}{\partial log(r)} \quad r \in [r_1, r_2] \tag{1}$$

Conceitos da teoria de fractais têm sido utilizados em diversas tarefas de análise e mineração de dados, tais como estimativa de seletividade [Baioco et al. 2007], detecção de agrupamentos [Barbará and Chen 2000], previsão em séries temporais [Chakrabarti and Faloutsos 2002], identificação de correlações entre atributos [Sousa et al. 2007b] e análise de distribuição dos dados [Traina et al. 2005].

A informação de comportamento intrínseco fornecida pela dimensão fractal D_2 também pode ser aplicada na identificação de mudanças de comportamento em *data streams*. A idéia geral é monitorar a *data stream* por meio do cálculo continuado de D_2 ao longo do tempo. Assim, alterações significativas em sucessivas medidas de D_2 podem indicar alterações na distribuição dos dados.

Uma técnica para mensurar D_2 em *data streams* foi proposta em [Sousa et al. 2007a], com o algoritmo *SID-meter*. Nessa abordagem, uma *data stream* é definida como uma sequência ordenada de eventos $< e_1, e_2, \ldots, e_n >$, sendo cada evento representado por um vetor de E medidas. Os eventos que ocorrem dentro de um determinado intervalo de tempo são tratados como um conjunto de dimensão E, viabilizando o cálculo continuado de D_2 para sucessivas sequências de eventos.

O *SID-meter* é baseado em janelas temporais deslizantes divididas em períodos de contagem sequenciais (n_c), tal que cada período inclui um número pré-determinado de eventos (n_i) da *stream*. Portanto, $n_i * n_c$ determina o tamanho da janela e n_i representa seu deslocamento. O valor de D_2 é calculado considerando todos os eventos na janela e atualizado a cada n_i eventos. A Figura 1 ilustra janelas temporais consecutivas divididas em 5 períodos de contagem sobre uma *data stream* composta dos atributos a_1, a_2 e a_3.

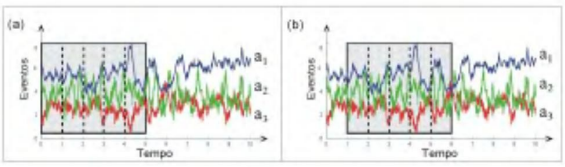

Figura 1. Janelas temporais consecutivas sobre uma *data stream* tridimensional.

3. A Abordagem Proposta

Este trabalho propõe um processo de análise de múltiplas séries temporais que combina monitoramento de *data streams* e agrupamento de dados. Nesta abordagem, múltiplas séries temporais são tratadas como uma *data stream* multidimensional, isto é, cada série é considerada um atributo da *data stream*. Por exemplo: séries climáticas de temperatura e precipitação são integradas definindo uma *data stream* bidimensional.

O processo proposto, ilustrado na Figura 2, possui duas etapas principais: 1) processamento *off-line* da *data stream* utilizando uma extensão do algoritmo *SID-meter*; 2) agrupamento das medidas de dimensão fractal obtidas na etapa anterior.

O *SID-meter*, originalmente, foi projetado para considerar apenas um tamanho pré-definido de janela deslizante. Para ser aplicado em dados climáticos de modo mais adequado, foi criada uma extensão do algoritmo para oferecer suporte a janelas deslizantes de tamanhos distintos, aplicadas simultaneamente durante uma única leitura do conjunto de dados. Esta extensão permite a análise temporal do comportamento da *stream* em diferentes granularidades de tempo, buscando detectar padrões que ocorrem, por exemplo, bimestralmente, semestralmente ou anualmente.

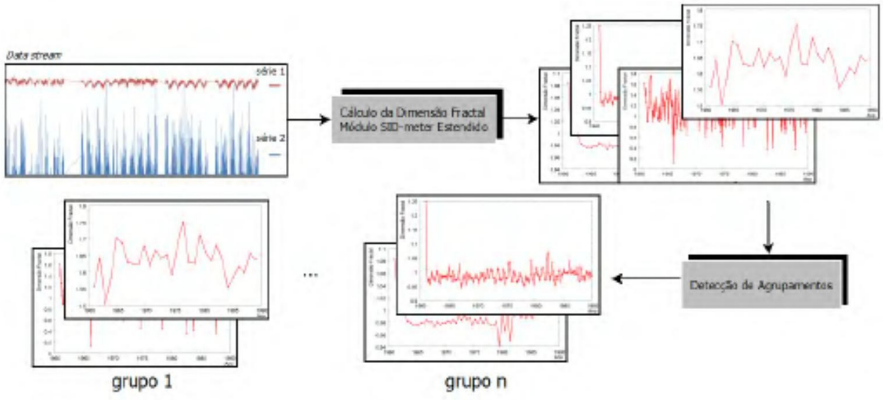

Figura 2. Processo de análise de múltiplas séries temporais.

As janelas de diversos tamanhos são geradas a partir de parâmetros de inicialização definidos de acordo com as necessidades do especialista: o menor e o maior tamanho de janela, isto é, valores mínimo e máximo de n_c (períodos de contagem) e o valor de n_i (eventos por período). Além disso, são definidos os valores de incremento automático para n_c e n_i. O resultado deste processamento são vários gráficos contendo os valores de D_2 ao longo do tempo para diferentes granularidades temporais.

Os gráficos gerados pelo *SID-meter* podem ser considerados séries temporais

de medidas de D_2. Assim, visando um estudo mais refinado, o processo de análise desenvolvido agrupa as séries de D_2 com o objetivo de identificar padrões similares que se repetem em diferentes granularidades de tempo, e padrões que são revelados somente quando a série é analisada com granularidades mais específicas. Neste trabalho, foi utilizado o método de particionamento *K-Medoids* [Kaufman and Rousseeuw 1990], com a função de comparação *Dynamic Time Warping* (DTW) para mensurar a similaridade entre as séries. A DTW realiza um relaxamento durante a comparação dos padrões, encontrando similaridades mesmo que existam deslocamentos ou deformações nas séries. O *K-Medoids* foi escolhido para a implementação inicial por sua simplicidade e por permitir a escolha de elementos do próprio conjunto como centros dos agrupamentos.

4. Experimentos e Resultados

O processo de análise proposto foi aplicado a séries temporais climáticas. Para a realização de experimentos foram utilizados dois conjuntos de dados:

1. Dados reais: séries climáticas reais, fornecidas pelo Agritempo[1], compostas de medidas diárias de precipitação e temperatura média aferidas por 25 estações meteorológicas de superfície do Estado de São Paulo, no período de 1961 a 1990.

2. Dados estimados: séries climáticas compostas de medidas estimadas de temperatura média e precipitação, obtidas no *site* do *WMC Global Climate Resource*[2]. As medidas são estimadas por interpolação espacial a partir de médias mensais de medidas reais coletadas por estações de superfície, gerando séries climáticas para todos os pontos de uma malha global com latitude e longitude variado de 0,5 em 0,5 grau. Para efeito de comparação nos experimentos, foram selecionadas as séries estimadas no período de 1961-1990, para pontos com latitude e longitude mais próximos às localizações das estações reais de SP que fazem parte do banco de dados do Agritempo.

Para cada um dos conjunto de dados, foi definida uma *data stream* bidimensional composta dos atributos `precipitação` e `temperatura média`. Para a inicialização dos parâmetros do *SID-meter* foi definido o número de períodos de contagem (n_c) variando de 2 a 5, com deslocamento (n_i) mínimo de 1 mês e máximo de 1 ano. Isto significa uma janela mínima de 2 meses e uma janela máxima de 5 anos.

Os gráficos de variação da dimensão fractal D_2 gerados pelo *SID-meter* para os dois conjuntos de dados apresentaram comportamentos significativamente distintos. Os dados reais resultaram em gráficos com menos variações no valor de D_2, que se manteve em torno de 1, como ilustrado na Figura 3a. Isso indica que as variáveis `precipitação` e `temperatura média` possuem correlação. Esse comportamento condiz com a expectativa dos meteorologistas, uma vez que essas duas variáveis são correlacionadas, variando de uma correlação mais forte em determinado período para uma correlação mais fraca em outros. Já nos gráficos gerados com os dados estimados por interpolação, as variações na D_2 foram maiores e mais bruscas, com os valores próximos a 2 indicando pouca ou nenhuma correlação entre os atributos, como ilustrado na Figura 3b. Isso é uma indicação clara de que métodos usualmente aplicados na área de meteorologia para a geração de medidas estimadas ainda requerem aprimoramento, mesmo do ponto de vista puramente numérico.

[1] Sistema de Monitoramento Agrometeorológico - http://www.agritempo.gov.br/
[2] http://climate.geog.udel.edu/~climate

A diferença nas correlações entre medidas reais e medidas estimadas foi indicada também ao executar o algoritmo de agrupamento *K-Medoids*: os gráficos gerados com dados reais ficaram em grupos completamente separados dos grupos contendo gráficos correspondentes a dados estimados. Foram criados 6 grupos para dados reais, cada um contendo gráficos para o mesmo tamanho de janela porém com deslocamentos distintos. Por outro lado, para os dados estimados, foram gerados 4 grupos contendo gráficos para diferentes tamanhos de janela e deslocamento. Como os tamanhos de janela podem variar bastante, o agrupamento de gráficos com diferentes granularidades temporais facilita o processo de análise do especialista, permitindo que ele(a) possa selecionar mais rapidamente quais janelas temporais evidenciam fenômenos a serem estudados.

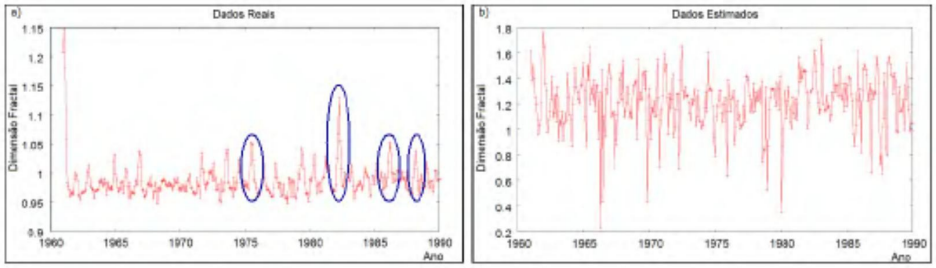

Figura 3. Variação de D_2 para janelas de 3 meses: (a) dados reais - Agritempo; (b) dados estimados por interpolação - WMC.

O gráfico apresentado na Figura 3a mostra o cálculo de D_2 para uma janela de 3 meses com deslocamento de 1 mês. Pode-se observar que apesar de D_2 não superar muito o valor 1, o gráfico apresenta alguns picos, destacados na figura. O maior deles corresponde ao ano de 1983, em que houve a ocorrência de um El Niño forte (82/83). O El Niño Oscilação Sul (ENOS) é um fenômeno de grande escala que ocorre no Oceano Pacífico na costa da América do Sul e se caracteriza pelo aumento das temperaturas da água do mar naquela região. Ele provoca chuvas intensas na região Sul do Brasil e secas no Nordeste [Berlato and Fontana 2003]. O Estado de São Paulo, em particular, sofre influência da Zona de Convergência do Atlântico Sul (ZCAS), segundo estudos sobre a influência do fenômeno ENOS no regime de chuvas da América do Sul durante as monções [Grimm and Tedeschi 2008]. Em anos de El Niño, há um aumento no número de ocorrência de eventos de chuvas fortes entre os meses de outubro e fevereiro, com uma quebra em janeiro. Entretanto, em se tratando do total mensal de precipitação, o sinal é menos pronunciado, ocorrendo em algumas regiões e em outras não.

Na Figura 3a, alguns outros picos positivos também coincidem com anos de El Niño, como o pico em 1986, que corresponde a um El Niño (86-88) moderado, e o pico em 1976, que corresponde a um El Niño (76-77) fraco. Este resultado mostra que quando houve alguma variação na relação entre a precipitação e a temperatura, isto foi refletido no gráfico de dimensão fractal, uma vez que as alterações só aparecem quando ocorre uma mudança no comportamento das variáveis que estão sendo analisadas.

No entanto, no gráfico da Figura 3b, referente a dados estimados, não foi possível identificar correspondência entre o comportamento fractal apresentado e fenômenos climáticos com anomalias. Uma explicação possível é que, para a geração das séries climáticas, tenham sido utilizados poucos pontos com medidas reais para estimar os

demais valores, incorrendo em muitas estimativas imprecisas. Isso indica que a utilização de métodos de interpolação para melhorar a cobertura de regiões com poucas estações de superfície (o que é um procedimento comum empregado pelos meteorologistas) na realidade perturba o processo de análise, ao distorcer as correlações existentes entre diferentes variáveis climáticas.

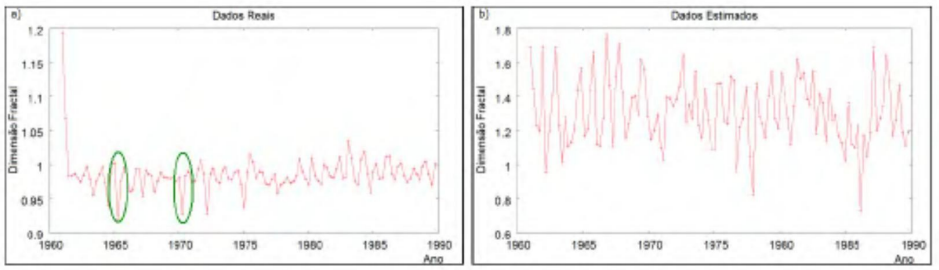

Figura 4. Variação de D_2 para janelas de 6 meses: (a) dados reais - Agritempo; (b) dados estimados por interpolação - WMC.

Dados do mesmo período mostrado na Figura 3a foram utilizados para gerar o gráfico apresentado na Figura 4a, porém com uma janela de 6 meses com deslocamento de 3 meses. Esse gráfico apresenta o mesmo padrão de comportamento do anterior (Figura 3a) e, consequentemente, ambos foram incluídos no mesmo grupo pelo *K-Medoids*. Como o tamanho da janela é maior, o número de pontos no gráfico diminui, tornando mais tênues os picos positivos, como em 1983, que corresponde a um El Niño. Por outro lado, picos negativos, como em 1966 e 1971, coincidem com outra anomalia importante, a La Niña, quando ocorre o resfriamento das águas e aumento na pressão atmosférica na região leste do Pacífico [Berlato and Fontana 2003]. Em anos de La Niña ocorre o oposto ao El Niño, diminuindo o número e a intensidade de eventos extremos e, como os gráficos mostram, passa a existir uma maior correlação entre as variáveis medidas.

Finalmente, o gráfico gerado com os dados estimados para a mesma configuração de janela (Figura 4b) possui muita variação e não foi possível identificar padrões como nos dados reais. Novamente, isso indica que a estimativa realizada para a geração das séries climáticas não reflete com precisão o comportamento dos dados reais, e esta é uma informação relevante para os meteorologistas.

5. Conclusão e Trabalhos Futuros

Neste artigo foi apresentada uma abordagem para análise de múltiplas séries temporais, combinando monitoramento de *data streams* e agrupamento de dados.

De modo geral, os resultados iniciais mostraram que a análise baseada na variação da dimensão fractal é capaz de indicar alterações de comportamento dos dados que coincidem com fenômenos climáticos. Além disso, o agrupamento dos gráficos resultantes do *SID-meter* mostrou que padrões semelhantes aparecem num mesmo grupo, permitindo ao especialista realizar estudos mais refinados considerando granularidades temporais de interesse em cada grupo.

Ressalta-se que comparar séries de dados meteorológicos reais com dados estimados, e identificar quando os dados estimados apresentam comportamento de fato compatível com os dados reais, é de fundamental importância para os meteorologistas,

dada a necessidade crescente de estudos mais detalhados sobre o comportamento do clima num cenário em que nem sempre a rede de estações disponíveis é suficiente, tornando necessária a pesquisa baseada em dados estimados.

A extensão da análise para séries climáticas de outras regiões do Brasil, que também são fortemente influenciadas por fenômenos como El Niño e La Niña, serão os próximos passos do trabalho. Além disso, pretende-se explorar outros algoritmos de detecção de agrupamentos, incluindo algoritmos específicos para agrupamento de séries temporais. Análises comparativas com outras técnicas também serão realizadas.

6. Agradecimentos

Os autores agradecem a Embrapa, Fapesp, CNPq, Capes e Microsoft Research pelo apoio financeiro e ao Agritempo pelos dados climáticos reais utilizados neste trabalho.

Referências

Alexander, L. et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, 111:1–22.

Ayoade, J. O. (1996). *Introdução à climatologia para os trópicos*. Ed.Brestrand Brasil, Rio de Janeiro.

Baioco, G. B., Traina, A. J. M., and Traina, C. (2007). Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *SSDBM 2007*, pages 6–16, Banff, Canada. ACM Press.

Barbará, D. and Chen, P. (2000). Using the fractal dimension to cluster datasets. In *ACM SIGKDD*, pages 260–264, Boston, MA.

Berlato, M. A. and Fontana, D. C. (2003). *El Niño e La Niña: impactos no clima, na vegetação e na agricultura do Rio Grande do Sul; aplicações de previsões climáticas na agricultura.* Editora da UFRGS, Porto Alegre.

Chakrabarti, D. and Faloutsos, C. (2002). F4: large-scale automated forecasting using fractals. In *CIKM*, volume 1, pages 2–9, McLean, VA - EUA. ACM Press.

Faloutsos, C. and Kamel, I. (1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *ACM PODS*, pages 4–13, Minneapolis, MN.

Grimm, A. and Tedeschi, R. G. (2008). Enso and extreme rainfall events in south america. *Journal of Climate*, 22:1589–1609.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons.

Schroeder, M. (1991). *Fractals, Chaos, Power Laws*. W. H. Freeman and Company.

Sousa, E. P. M., Traina, C., Traina, A. J. M., and Faloutsos, C. (2007a). Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing Journal*, 25:33–59.

Sousa, E. P. M., Traina, C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2007b). A fast and effective method to find correlations among attributes in databases. *DMKD*, 14(3):367 – 407.

Traina, C., Sousa, E. P. M., and Traina, A. J. M. (2005). Using fractals in data mining. In *New Generation of Data Mining Applications*, pages 599–630 (Chapter 24). Wiley/IEEE Press.

Zhai, P. et al. (2005). Guidelines on climate watches. Technical report, World Meteorological Organization. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/GuidelinesonClimateWatches.pdf.