Exploração visual interativa de dados coletados pelo Sistema Integrado de Monitoramento Ambiental - SIMA

Alisson Fernando Coelho do Carmo[1], **Milton Hirokazu Shimabukuro**[1], **Enner Herenio de Alcântara**[1]

[1]Programa de Pós-Graduação em Ciências Cartográficas (PPGCC)
Faculdade de Ciências e Tecnologias (FCT)
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)
Presidente Prudente – SP – Brasil

alisondocarmo@gmail.com, {miltonhs,enner}@fct.unesp.br

Abstract. *Researches on environmental phenomena behavior have been benefited from sensor technological advances. Periodic collecting produces long time series, with large data sets composed by a great number of variables. Interactive visual exploration and analysis techniques can be applied to support such data sets processing in order to facilitate the identification of erros and trends, as well as, the detection of patterns. In this context, this paper presents an approach using visual representations for data, associated with interaction techniques, to support the process of identification and understanding the data structures and possible failures in data collected and stored by Integrated System for Environmental Monitoring.*

Resumo. *Estudos sobre o comportamento de fenômenos ambientais tem se beneficiado da evolução dos sensores. Coletas periódicas geram longas séries temporais, com grande quantidade de dados e variáveis. Técnicas de exploração e análise visual interativa de dados podem ser utilizadas para auxiliar a análise do grande volume de dados e facilitar a identificação de erros, tendências e observação de padrões. Neste contexto, este trabalho apresenta uma abordagem que utiliza técnicas de representação visual e interativa de dados para auxiliar a identificação e compreensão das estruturas de dados e possíveis falhas presentes no conjunto de dados coletados e armazenados pelo Sistema Integrado de Monitoramento Ambiental.*

1. Introdução

A necessidade de registrar, acompanhar e entender os fenômenos e comportamentos do meio ambiente sempre esteve presente no cotidiano do homem e é um de seus objetos de estudo. Tal tarefa tem se beneficiado do desenvolvimento tecnológico, principalmente relacionado à evolução da tecnologia empregada em dispositivos sensores. A utilização de sensores na coleta de dados ambientais permite a aquisição automática e periódica de dados. A periodicidade oferece uma nova possibilidade de análise, possibilitando a integração do atributo tempo com os valores coletados, originando longas séries temporais.

A principal característica da análise de séries temporais é o histórico dos valores registrados. Desta forma, é possível investigar o conjunto de dados em busca de padrões

e dependências que podem ser evidenciados durante a observação do comportamento dos registros ao longo do tempo.

No entanto, a manipulação de conjuntos de dados de séries temporais requer alguns cuidados. A constante aquisição dos dados faz com que o volume de dados coletados permaneça em crescimento frequente. Para que esta grande quantidade de dados possa ser analisada e interpretada, é necessária a utilização de recursos computacionais capazes de processar e sumarizar o conjunto.

Existem diversas metodologias que podem ser aplicadas para a análise de séries temporais. Um dos recursos que podem potencializar a análise é a utilização de técnicas de Visualização de Informação. Estas técnicas buscam representar visualmente os dados para facilitar a interpretação. Esta operação pode ser integrada com outros recursos que se beneficiem da percepção do analista durante a exploração dos dados, e garantam a interação durante a exploração dos dados, cenário conhecido como Visual Analytics.

Neste contexto, este trabalho tem o objetivo de apresentar os resultados parciais obtidos com a investigação sobre a utilização de técnicas de exploração e análise visual de dados oriundos de sensores ambientais. O presente texto discute alguns fatores relacionados com as características do conjunto de dados coletados pelo Sistema Integrado de Monitoramento Ambiental (SIMA).

As demais Seções descrevem os principais aspectos sobre o desenvolvimento deste trabalho, que compõe uma pesquisa em andamento. Na Seção 2 são apresentados alguns conceitos relacionados. A exploração e manipulação dos dados é descrita na Seção 3. Os resultados dos testes realizados com a exploração do conjunto de dados, sobretudo com a representação visual dos dados são abordados na Seção 4 para então, subsidiar as considerações finais apresentadas na Seção 5.

2. Conceitos relacionados

Nesta Seção são apresentados os principais conceitos relacionados com o desenvolvimento deste trabalho.

2.1. Séries Temporais

As séries temporais constituem uma importante configuração de dados, as quais os representam de forma ordenada em relação ao tempo. O estudo de séries temporais geralmente é focado em dois principais fatores que são diretamente relacionados, referentes à compreensão da forma que os valores da série são gerados e ao estudo do comportamento da série, permitindo a estimativa de valores ausentes em instantes de tempo da série, bem como a predição de valores futuros. Existem diversas técnicas tradicionais para análise de séries temporais, principalmente baseadas em cálculos estatísticos. Para potencializar os resultados obtidos na análise de séries temporais, outros recursos computacionais podem ser utilizados, como abordado por Esling e Agon [Esling and Agon 2012], que apresentam um levantamento sobre diferentes algoritmos e ferramentas que permitem aplicar técnicas de mineração de dados para a descoberta de conhecimento em séries temporais por meio do comportamento geométrico da variação dos dados.

2.2. Sistema Integrado de Monitoramento Ambiental

O SIMA é formado por um conjunto de tecnologias aplicadas à coleta de dados e monitoramento da hidrosfera [INPE 2013]. Ele é composto de uma rede de plataformas que

possui sensores aquáticos e sensores capazes de coletar atributos relacionados ao ar. As plataformas SIMA realizam a leitura dos sinais dos sensores com a periodicidade de uma hora. Após a leitura, os dados coletados são transmitidos via satélite, estando este visível para a plataforma ou não. Por esta razão, pode ocorrer de alguns dados serem perdidos no momento da transmissão. Servidores intermediários em estações terrestres são responsáveis por receber os dados transmitidos e realizar a verificação da existência de erros na transmissão dos sinais. Após esta validação dos dados, estes são transmitidos ao servidor no centro de armazenamento, os quais passam pelo processo de decodificação, processamento e armazenamento, e ficam disponíveis em um portal da internet. Alcântara et al [Alcântara et al. 2013] apresentam uma análise utilizando algumas métricas estatísticas sobre as plataformas e discutem sobre os principais problemas que podem estar presentes, relacionados à degradação dos sensores e comunicação com o satélite.

2.3. Visual Analytics

A complexidade envolvida no processo de exploração e análise de dados pode ser incrementada em razão de alguns aspectos intrínsecos, como o volume de dados a ser considerado, a quantidade de parâmetros a serem interpretados, e a integridade e qualidade associada a estes dados. Neste sentido, técnicas de análises e recursos computacionais podem oferecer as ferramentas necessárias para viabilizar esta tarefa e torná-la mais natural. O termo Visual Analytics (VA) foi apresentado por Thomas e Cook [Thomas and Cook 2005] no cenário em que a representação visual não era suficiente para viabilizar a análise direta de grandes quantidades de dados. Então, técnicas de interação e manipulação visual foram agregadas ao processo de análise para garantir a permanência do analista no centro do processo, de modo que sua capacidade de percepção e cognição visual pudesse ser utilizada para refinar o processo de exploração e construção do raciocínio analítico. No contexto de análise de dados obtidos por sensores, o fator tempo pode oferecer outras oportunidades de análise, como apresentado por Maciejewski et al [Maciejewski et al. 2010], que discutem sobre alguns benefícios que podem ser obtidos com a integração entre as representações tradicionais de séries temporais com outras técnicas de visualizações espaço temporais.

3. Exploração dos dados

Para a realização deste trabalho, foram considerados os dados coletados e armazenados pelo projeto SIMA. Para tanto, foram utilizadas as planilhas eletrônicas de dados exportados pelo portal na internet, levando em consideração o período de funcionamento de todas as plataformas registradas até início do ano de 2013. Os dados foram inseridos em um Sistema Gerenciador de Banco de Dados (SGBD) para facilitar a manipulação e filtragem dos dados a serem processados. O SGBD utilizado foi o PostgreSQL, escolha motivada por ser um sistema *open source* que possui integração com a extensão espacial PostGIS, também *open source*. O primeiro fator observado durante a exploração dos dados registrados, foi o tempo de atividade de cada plataforma, como pode ser visto na Figura 1, que apresenta o período em que as plataformas estiveram ativas - coletando e transmitindo dados.

Explorando os dados foi possível observar que mesmo durante o período de atividade da plataforma algumas amostras não estavam registradas. A ausência de dados de

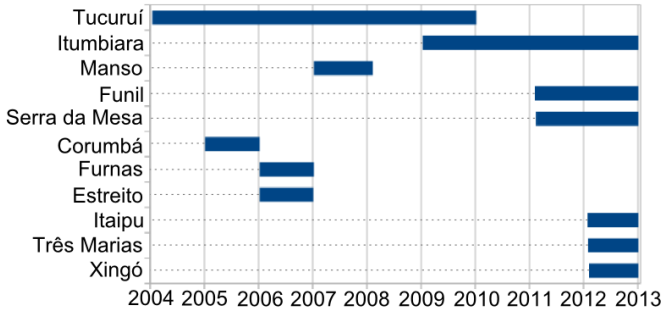

Figura 1. Períodos de atividades das plataformas SIMA

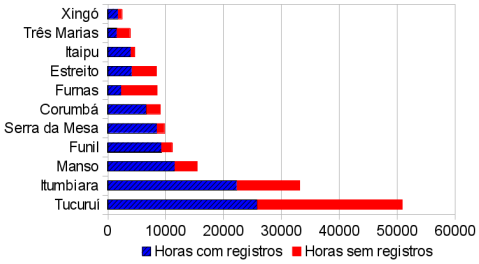

Figura 2. Horas ativas das plataformas

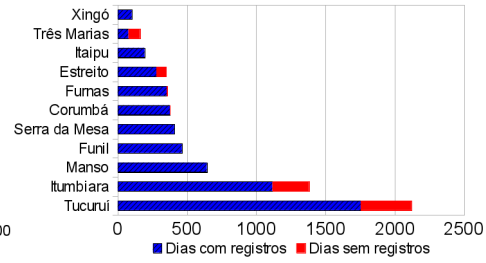

Figura 3. Dias ativos das plataformas

coletas ocorre tanto em determinadas horas do dia, mostrada na Figura 2, como também em dias completos, exibida na Figura 3.

Os gráficos apresentados nas Figuras 1, 2 e 3 permitem observar a ausência de dados de um modo geral. No entanto, esta visualização considera apenas a existência/ausência de registros, e não aborda a integridade dos dados existentes. A Seção 4 apresenta abordagens iniciais para a observação da integridade dos dados das plataformas SIMA.

4. Resultados preliminares

Inicialmente, para verificar o comportamento geral do estado das variáveis e sua integridade, foi utilizada uma visualização capaz de representar uma matriz de dados ordenados de acordo com o tempo. A Figura 4 exibe a representação de um período de dados da plataforma Três Marias, na qual cada linha representa uma variável diferente. A plataforma Três Marias foi escolhida em razão do maior índice proporcional de ausência de dados - objeto de estudo deste teste - tanto relativo a cada hora como ao dia completo.

A escala de cores utilizada na Figura 4 representa os valores de acordo com sua intensidade, variando em uma escala linear contínua. Nesta visualização foram considerados apenas os registros diários existentes no conjunto de dados, tornando a linha do tempo sequencial, sem lacunas. Vale ressaltar que a escala de tempo nesta representação não é linear, pois considera apenas os registros armazenados, organizando-os sequencialmente, independente da ausência de dados.

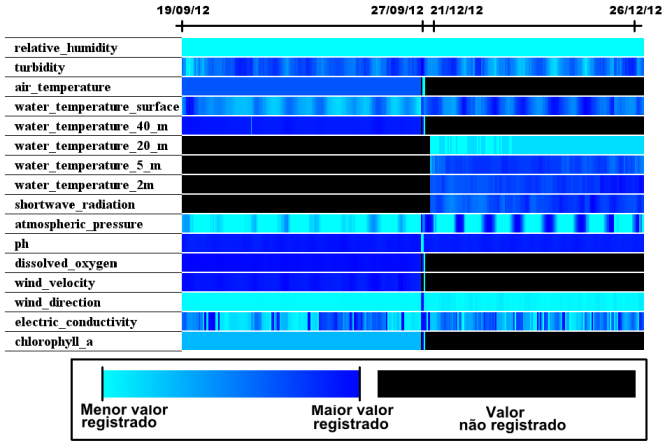

Figura 4. Visualização em matriz de um período da plataforma Três Marias

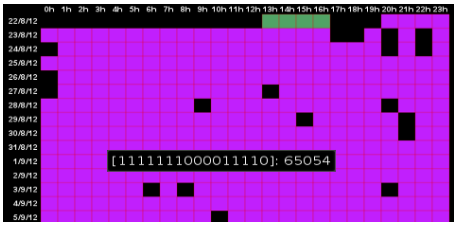

Figura 5. Visualização Calendar View (início do intervalo) dos dados da plataforma Três Marias

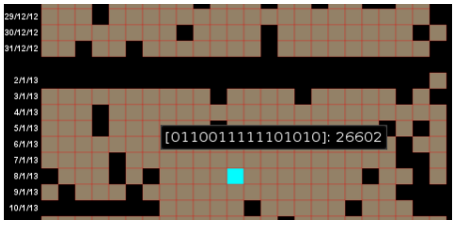

Figura 6. Visualização Calendar View (fim do intervalo) dos dados da plataforma Três Marias

Para evidenciar as falhas dos sensores, foi utilizada a técnica de representação conhecida como *Calendar View*. Nesta representação, a relação de tempo é segmentada em observações diárias em cada linha da matriz, e as colunas identificam as horas do dia. Os valores registrados de cada variável foram convertidos em uma sequência binária de 16 bits, em que cada bit representa a presença (1) ou ausência (0) do dado da respectiva variável. Uma vez definida a sequência binária para cada registro e realização do mapeamento discreto para a escala de cores do espaço RGB (*Red, Green, Blue*), é possível identificar o conjunto de sensores com falhas por meio da cor que representa o registro, facilitando a observação de padrões de falhas. As Figuras 5 e 6 apresentam um exemplo de dois períodos, inicial e final, respectivamente, da plataforma Três Marias.

Nesta representação a conversão da sequência binária ocorre de forma discreta, de modo que todas as combinações de falhas nas 16 variáveis (2^{16}) são mapeadas proporcionalmente para o espaço de cor RGB, garantindo que cada uma das combinações seja representada por uma cor diferente. Na Figura 5, que representa o início do período ativo da plataforma, predomina a cor referente à sequência binária "1111111000011110"que representa o erro nos dados de cinco sensores. Já nos últimos registros, representados na Figura 6, a cor predominante se altera, indicando a sequência "0110011111101010", ou seja, com erros em seis sensores, sendo todos diferentes do primeiro cenário.

5. Considerações finais

No contexto de análise de séries temporais geradas por sensores é comum a inconsistência de alguns dados, pois o processo de aquisição e armazenamento dos valores envolve diversos fatores. Os erros dos dados podem ser inseridos em razão de falha do sensor na leitura, erro na transmissão ou ainda na conversão do sinal elétrico do sensor para um valor discreto. Por esta razão, a manipulação e estudo do comportamento das séries temporais pode trazer benefícios associados à completitude do conjunto de dados, pois os valores perdidos podem ser estimados considerando o histórico temporal.

Existem diferentes abordagens para a análise de séries temporais. Este trabalho apresentou a utilização de técnicas de análise visual para auxiliar a exploração dos dados, contando com a capacidade humana de cognição visual. Os recursos de interatividade, por meio da manipulação da representação visual, permite que o analista refine os resultados gerados a partir de sua própria percepção. Comparando a visualização apresentada na Figura 4 com a representação dos mesmos dados utilizando a representação *Calendar View*, exibidas nas Figuras 5 e 6, é possível verificar a existência de falhas nos respectivos sensores. Desta forma, pode ser observado que diferentes visualizações podem ser utilizadas para representar o mesmo conjunto de dados, sendo que cada uma delas implica em leituras diferentes da representação, evidenciando características específicas.

Agradecimentos

Os autores agradecem o Programa de Pós-Graduação em Ciências Cartográficas (PPGCC) da Faculdade de Ciências e Tecnologia/UNESP (FCT/UNESP)- Campus de Presidente Prudente - por permitir o desenvolvimento desta investigação como tema de projeto de mestrado; a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio dedicado ao projeto; e ao Instituto Nacional de Pesquisas Espaciais (INPE) pela cessão dos dados SIMA.

Referências

[Alcântara et al. 2013] Alcântara, E., Curtarelli, M., Ogashawara, I., Stech, J., and Souza, A. (2013). A system for environmental monitoring of hydroelectric reservoirs in brazil. *Ambiente e Água - An Interdisciplinary Journal of Applied Science*, 8(1).

[Esling and Agon 2012] Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Comput. Surv.*, 45(1):12:1–12:34.

[INPE 2013] INPE, H. (2013). Sima: Sistema integrado de monitoramento ambiental. `http://www.dsr.inpe.br/hidrosfera/sima/`. Acessado em Agosto de 2013.

[Maciejewski et al. 2010] Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W. S., Grannis, S. J., and Ebert, D. S. (2010). A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220.

[Thomas and Cook 2005] Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.