BCG - Boletim de Ciências Geodésicas - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702016000300031

Artigo

AVALIAÇÃO DA QUALIDADE DE DADOS AMBIENTAIS POR MEIO DE TÉCNICAS DE ANALÍTICA VISUAL

Using Visual Analytics techniques to evaluate the Data Quality in environmental datasets

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. Email: alisondocarmo@gmail.com, miltonhs@fct.unesp.br, enner@fct.unesp.br.

Resumo:

O desenvolvimento tecnológico tem impulsionado a utilização de sensores para a realização de coletas automatizadas e periódicas de dados, como aqueles empregados no Sistema Integrado de Monitoramento Ambiental (SIMA), cujo conjunto de dados é utilizado neste trabalho. Apesar da automatização do processo de aquisição de dados, estes podem apresentar falhas decorrentes de problemas na coleta, na transmissão ou no armazenamento dos dados. A existência de grande quantidade de dados temporais multivariados e a possibilidade de falhas são indicativos da necessidade de utilização de recursos computacionais para apoiar o processo de análise. Neste trabalho são utilizadas técnicas de análise visual para a extração de características do conjunto de dados, as quais, posteriormente, podem impactar a qualidade da análise dos fenômenos associados. Os resultados obtidos demonstram os benefícios da utilização de representações visuais e interativas para a exploração do conjunto de dados, as quais facilitam a percepção de informações acerca de: disponibilidade dos dados; funcionamento dos sensores; e evidências de padrões de falhas.

Palavras-chave: Qualidade de dados; Sensores ambientais; Analítica Visual.

Abstract:

The technological advances has boosted automated and periodic data collections using sensors, such as those performed in Integrated Environmental Monitoring System (SIMA), which dataset is used in this work. Despite of the automation degree, there are some flaws in the dataset caused by problems in the collection, transmission or data storage. Due to the existence of large multivariate temporal data and the possibility of failure, interactive visual representations of the data can improve the study and characterization of the dataset. In this work, Visual Analytics techniques were applied to obtain some characteristics from the environmental dataset, which subsequently can impact the quality of analysis of the associated phenomena. The results showed the benefits of using visual and interactive representations for dataset analysis, which facilitates

the information extraction about dataset characteristics, such as data availability, sensors operation characteristics, and evidences of a failure pattern.

Keywords: Data Quality; Environmental sensors; Visual Analytics.

1. Introdução

A necessidade de registrar, monitorar e entender fenômenos e comportamentos associados ao meio ambiente é cada vez mais importante. Em especial, as características de ambientes aquáticos podem ressaltar importantes fatores que influenciam parâmetros de qualidade da água. Atributos de qualidade da água não definem apenas indicativos sobre o ambiente isoladamente, mas resultam em aspectos que influenciam bens e serviços aplicados sobre outros ecossistemas, como lazer, recreação, propósitos comerciais diversos, bem-estar humano, saúde pública, entre outros (Keeler et al., 2012).

O total entendimento dos processos físicos, químicos e biológicos que agem sobre ambientes aquáticos requer a manipulação de séries temporais com vasto conjunto de atributos meteorológicos e limnológicos (Stech et al., 2011). A coleta automática e periódica de dados utilizando dispositivos autônomos, como as Plataformas de Coleta de Dados (PCD), pode beneficiar o processo de aquisição de dados. O Sistema Integrado de Monitoramento Ambiental (SIMA) é uma das abordagens que se enquadra na definição de PCD. As plataformas do SIMA estão fundeadas em reservatórios e são utilizadas para o monitoramento da hidrosfera, pois são compostas por sensores que permitem a coleta de atributos relacionados ao ar e à água (Stech et al., 2006, Alcântara et al., 2013). Tais plataformas são responsáveis pela coleta e transmissão dos dados via enlace de satélite para que sejam armazenados e processados em um servidor.

Embora a automatização do processo de coleta de dados ofereça benefícios pela resolução temporal da realização das coletas e capacidade de coletar múltiplas variáveis, esta abordagem depende do bom funcionamento de dispositivos eletrônicos frágeis que podem apresentar falhas. Alguns dos fatores que podem influenciar a qualidade dos dados coletados pelas plataformas SIMA podem estar relacionados à: exposição dos sensores às intempéries do ambiente; perdas de dados durante o momento de transmissão para o satélite e a falha na conversão dos sinais elétricos captados pelos sensores em números discretos (Alcântara et al., 2013).

O contínuo avanço de tecnologias que permitam monitoramento remoto em tempo-real (*Real-Time Remote Monitoring* - RTRM) tem impulsionado cada vez mais a utilização de dispositivos autônomos para coleta de dados, resultando em uma importante ferramenta para apoiar o gerenciamento ambiental (Glasgow et al., 2004). Aspectos relacionados à confiabilidade e qualidade dos dados são mencionados em cenários que utilizam dispositivos autônomos para a coleta de dados. A falha de componentes eletrônicos é apontada com uma das fontes que influenciam diretamente a qualidade dos dados coletados por plataformas autônomas (Dubelaar e Gerritzen, 2000; Etcheber et al., 2010).

Algoritmos e rotinas para garantia de qualidade (*Quality Assurance* - QA) podem ser utilizados para evitar o descarte de conjunto de dados e identificar dados anormais com *flags*, inseridas nos metadados, para permitir avaliações sobre a confiabilidade dos dados (Shafer et al., 2000). Após a marcação de dados suspeitos, alguns gráficos – gráficos de linha; gráficos de dispersão; mapas de contorno e séries temporais – podem ser utilizados para observação e validação da

confiabilidade dos dados assinalados, permitindo distinguir entre dados verdadeiros ou falhas do sistema (Hamilton, 1986).

Uma das abordagens para apoiar a análise de dados constantemente em crescimento - como o conjunto gerado por PCD - é por meio de representações visuais e interativas dos dados utilizando técnicas de *Visual Analytics* (VA). Este conceito foi apresentado por Thomas e Cook (2005) no cenário em que apenas as representações visuais não eram suficientes para viabilizar a análise de grandes quantidades de dados. Então, técnicas de interação foram agregadas ao processo de análise para garantir a permanência do analista no centro do processo, de modo que sua capacidade de percepção visual e cognição pudessem ser utilizadas para refinar a exploração dos dados e viabilizar a construção do raciocínio analítico. Em razão do termo *Visual Analytics* não possuir uma tradução adotada em consenso na literatura, neste trabalho será empregado como analítica visual, exploração analítica visual ou análise visual interativa.

A utilização de representações visuais interativas e outros recursos apoiados por computador evidenciam a importância da colaboração entre homem e máquina no processo analítico (Keim et al., 2006). Benefícios podem ser conseguidos utilizando recursos que permitam a interação do usuário com os dados (Ward et al., 2010), dentre os quais, a capacidade de aplicação de filtros nos dados, as operações de seleção dos objetos visuais, navegação e exploração por zoom e reconstrução das representações visuais podem facilitar o entendimento de processos analíticos que muitas vezes podem ser vistos como obscuro por aqueles que não o dominam (Jeong et al., 2009). Tais recursos podem ser amplificados com sua disponibilização e acesso via WEB, cenário em que ferramentas interativas, de representação, exploração e análise de dados são fatores importantes para amplificar a democratização do uso de representações gráficas e visualizações avançadas capazes de facilitar atividades que dependem do acesso e manipulação de conjuntos de dados (Levkowitz e Kelleher 2012).

Diante deste cenário, o objetivo deste trabalho foi o de explorar o conjunto de dados coletados por sensores ambientais na plataforma SIMA utilizando técnicas de Analítica Visual para sua caracterização. A caracterização do conjunto de dados pode ser sumarizada na identificação de alguns fatores principais, tais como: disponibilidade suficiente de dados; intervalo de funcionamento dos sensores; e evidências de falhas sistemáticas ou não, as quais influenciam diretamente a qualidade da análise. O foco de caracterização inicial do conjunto busca viabilizar e facilitar as próximas iterações do especialista no processo de análise dos dados e respectivos fenômenos.

2. Exploração e análise do conjunto de dados

Cada plataforma ou estação SIMA possui sensores meteorológicos, uma cadeia de termistores, capazes de coletar dados de temperatura em diversas profundidades, e uma sonda multiparâmetros que captura atributos de qualidade da água, resultando em aproximadamente 20 variáveis coletadas, dentre elas: clorofila; condutividade; inundação; NH_4+; NO_3-; oxigênio dissolvido; pH; radiação incidente; radiação refletida; temperatura da água; temperatura da sonda; turbidez; CO_2; direção do vento; intensidade do vento; temperatura do ar; umidade relativa do ar; velocidade meridional da corrente; velocidade meridional do vento; velocidade zonal da corrente e velocidade zonal do vento.

Na Figura 1 é possível verificar a disposição dos sensores e painéis solares na estrutura da boia, e notar sua exposição direta ao ambiente, tanto para os sensores em contato com o ar, como aqueles submersos que interagem com a água, fator esse que ressalta sua degradação contínua.

Figura 1: Plataforma SIMA e seus sensores.
Fonte: http://www.dpi.inpe.br/sima/boias.html

As plataformas SIMA realizam a leitura dos sinais dos sensores com a periodicidade padrão de uma hora. Após a leitura, os dados coletados são transmitidos via enlace de satélite para servidores intermediários que são responsáveis por receber os dados e realizar a verificação de erros na transmissão dos sinais. Após a validação, os dados são transmitidos ao servidor no centro de armazenamento, os quais são submetidos ao processo de decodificação, processamento e armazenamento, para posteriormente, ficarem disponíveis em um portal da internet que pode ser utilizado mediante acesso autorizado.

O portal online do SIMA permite a obtenção do conjunto de dados desejado em formato de planilha eletrônica. Uma vez adquiridos os dados de todas as plataformas SIMA, estes 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 com código aberto que possui integração com a extensão espacial PostGIS.

Como o objetivo principal deste trabalho está relacionado com a exploração das capacidades de utilização de técnicas de representações visuais e interativas para apoiar o processo de análise de dados, algumas representações visuais foram adaptadas, implementadas e organizadas em um protótipo de aplicação Web voltado para a exploração e análise dos dados, nomeado como SimaVIS.

O sistema SimaVIS foi concebido em uma arquitetura Cliente-Servidor de forma que o acesso aos dados e processamento estão presentes no ambiente servidor e são utilizados por meio de um *Web Service*. O *Web Service* foi construído considerando a mesma interface de acesso aos dados do padrão SOS (*Sensor Observation Service*), especificado pelo OGC (*Open Geospatial Consortium*), mas com diferenças na modelagem interna dos dados e na codificação de resposta do serviço.

2.1 Identificação do tempo de atividade das plataformas SIMA

A disponibilidade de registros em uma determinada localização e em um específico intervalo de tempo é essencial para o início da análise. Para extrair esta informação a partir de dados tabulares, a dificuldade de interpretação é acrescida de acordo com a quantidade de plataformas analisadas. Na Tabela 1 são apresentadas as informações sobre o tempo de atividade das plataformas. É possível perceber que a extração da informação sobre as plataformas que possuem dados em um determinado instante de tempo não é imediata e exige a comparação individual dos registros da tabela.

Tabela 1: Períodos de atividades de cada plataforma

Plataforma SIMA	Primeiro Registro	Último Registro	Plataforma SIMA	Primeiro Registro	Último Registro
Balbina	16/08/2013	19/01/2015	Itumbiara 2	28/03/2009	19/01/2015
Corumbá	26/01/2005	10/02/2006	Itumbiara 3	18/11/2009	25/09/2011
Curuai	25/04/2004	11/03/2010	Mamirauá	08/06/2009	19/01/2015
Estreito	12/02/2006	30/01/2007	Manso 1	18/01/2004	22/01/2005
Funil 1	08/02/2007	05/09/2011	Manso 2	31/01/2007	06/11/2008
Funil 2	08/02/2007	16/10/2007	Masc. de Moraes	07/02/2006	06/02/2007
Funil 3	24/10/2011	19/01/2015	Segredo	31/01/2013	19/01/2015
Furnas - Embrapa	26/07/2013	19/01/2015	Serra da Mesa 1	12/01/2004	08/06/2010
Furnas 1	13/02/2006	06/02/2007	Serra da Mesa 2	14/12/2011	19/01/2015
Ibitinga 1	13/09/2012	19/01/2015	Três Marias	22/08/2012	19/01/2015
Ibitinga 2	21/03/2013	19/01/2015	Tucuruí 1	27/04/2004	14/02/2010
Ibitinga 3	22/03/2013	19/01/2015	Tucuruí 2	21/11/2012	19/01/2015
Itaipu	20/07/2012	19/01/2015	Xingó	20/10/2012	19/01/2015
Itumbiara 1	23/01/2005	11/02/2006			

Por outro lado, com a utilização de uma representação visual baseada em gráfico de intervalos, é possível obter a mesma informação, porém de uma maneira imediata e intuitiva baseada na percepção visual, como pode ser visto na Figura 2. O gráfico de intervalos pode ser utilizado para exibir o período de tempo de atividade de cada plataforma, que representa o tempo em que as plataformas estiveram coletando e transmitindo dados.

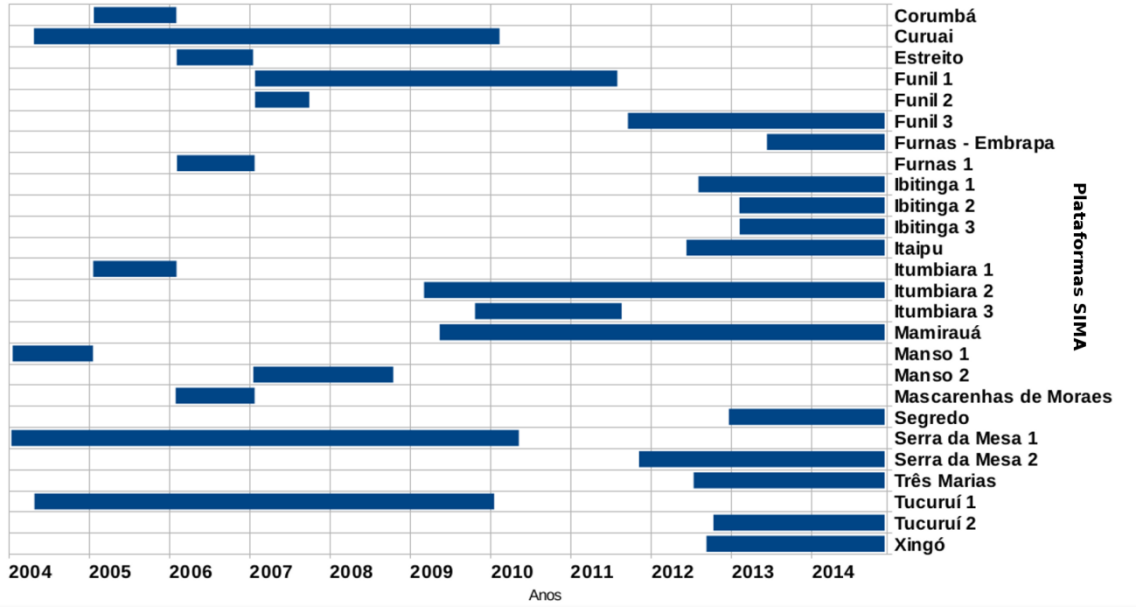

Figura 2: Períodos de atividades das plataformas SIMA

O gráfico apresentado na Figura 2 permite notar a existência de diversas plataformas que produziram dados apenas em um intervalo de tempo de cerca de um ano, como: Corumbá; Estreito; Funil2; Furnas1; Itumbiara1; Manso1 e Mascarenhas de Moraes. Uma possibilidade é que estas plataformas podem ser estações temporárias de coleta que podem ter sido movidas para outros reservatórios, pois, de acordo com o gráfico, esta situação ocorre principalmente no início do projeto.

2.2 Quantificação dos dados faltantes

O gráfico de intervalos não informa se a plataforma esteve realmente coletando e transmitindo dados na plenitude do intervalo, pois são considerados apenas o primeiro e o último dia de registro para defini-lo. Utilizando o gráfico radial, exibido na Figura 3, é possível notar a proporção de dados coletados com sucesso e com falhas. São exibidas as quantidades de registros efetivamente armazenados no banco de dados e a quantidade de coletas que não foram registradas.

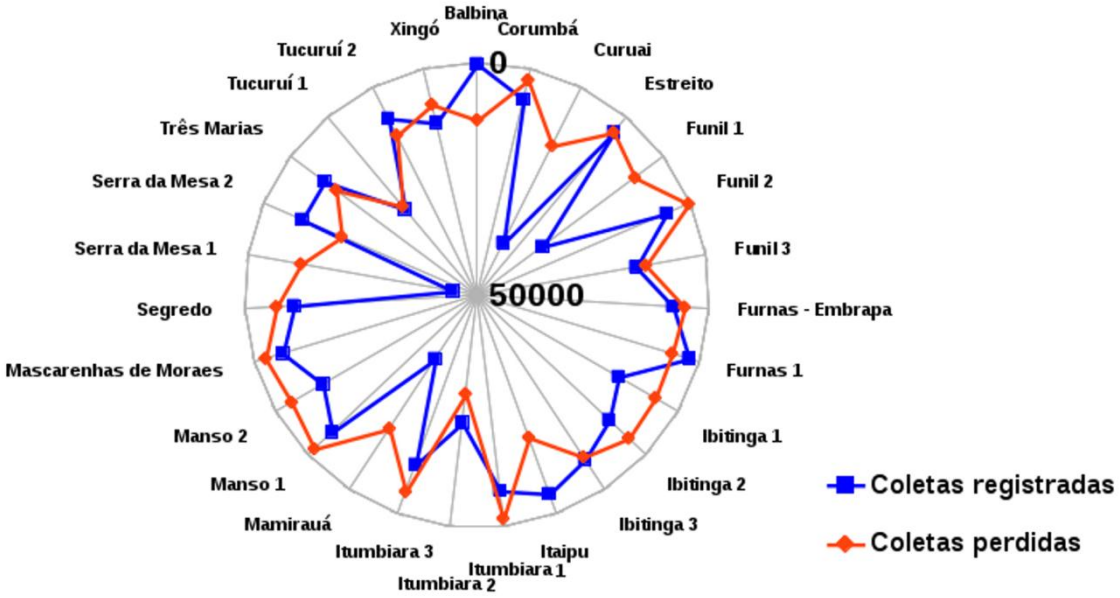

Figura 3: Relação de coletas realizadas pelas plataformas SIMA. Em azul são as coletas registradas no banco de dados e em vermelho as coletas perdidas. Note que o centro corresponde ao valor máximo.

A quantidade esperada de coletas que cada plataforma deve realizar pode ser obtida a partir da quantidade de dias em que a plataforma esteve ativa e o fato de que todas as plataformas devem processar e transmitir 24 coletas diárias. No gráfico, a relação com o percentual de falhas pode ser observada pela distância entre os pontos vermelhos (falhas) e os pontos azuis (sucessos) de cada eixo, na qual diferenças grandes representam um comportamento com poucas falhas ou com muitas falhas, e diferenças pequenas com pontos próximos representam um equilíbrio entre quantidade de falhas e de dados coletados com sucesso. Como indicado na Figura 3: (a) Serra da Mesa 1 é um exemplo de plataforma com poucas falhas; (b) Itaipu apresenta um índice alto de falhas e (c) Tucuruí 1 representa quantidade similar entre falhas e sucessos. Algumas plataformas SIMA possuem índice de falhas superiores a 70% do total coletado, como: Furnas 1 (72,18%); Itaipu (79,45%) e Balbina (98,32%). Este fato aponta que poderiam existir dias

completos sem nenhum dado. No gráfico de barras, visto na Figura 4, podem ser observadas as quantidades de coletas diárias.

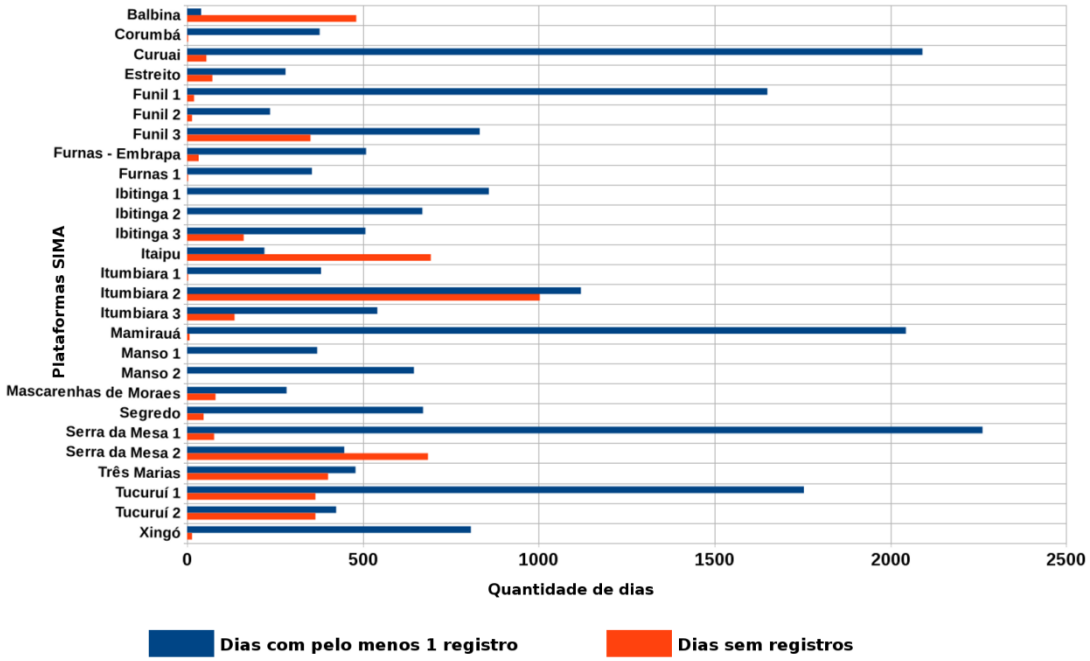

Figura 4: Coletas realizadas por dia. Em azul a quantidade de dias que existem coletas registradas e em vermelho a quantidade de dias que não possuem nenhum registro.

Furnas 1, Itaipu e Balbina indicaram índices de falhas superiores a 70%, mas apenas Furnas 1 apresenta um baixo índice (menor que 1%) de dias completos sem dados. Tal fato pode significar que a plataforma SIMA de Furnas 1 não está programada para realizar coletas em intervalos de 1 hora, como especificado no projeto, mas uma quantidade menor de coletas diárias. Esta justificativa pode ser fortalecida a partir da quantidade média diária de cinco observações encontradas no conjunto de dados em todo período de atividade desta plataforma. A informação da frequência de execução das coletas poderia ser parte integrante dos metadados que definem cada plataforma, reforçando ainda mais a importância da modelagem e organização dos dados.

2.3 Identificação de falhas nos atributos das coletas registradas

Para verificar a integridade e a variação de cada atributo, outras formas de representação visual podem ser consideradas para facilitar a percepção. Uma técnica que pode ser aplicada para a representação do comportamento de dados com atributo temporal é a visualização baseada em pixel, que é útil para observar simultaneamente os valores apresentados por vários atributos na mesma escala de tempo. Com esta representação é possível notar falhas que podem apresentar padrão entre os atributos, como pode ser visto na Figura 5, que mostra um intervalo de dados da plataforma Três Marias, no qual ocorre falha em diferentes grupos de atributos.

A escala de cores linear e contínua utilizada representa os valores de acordo com sua intensidade. São considerados apenas os registros presentes no conjunto de dados, apresentando-os de maneira ordenada no eixo temporal em uma escala não linear, isto é, não contínua. Na Figura 5, é destacado um intervalo que existe um salto no eixo temporal, no qual o registro de

21/12/2012 aparece imediatamente após o registro do dia 27/09/2012, evidenciando um intervalo de tempo em que nenhum dado foi registrado. Esta estratégia foi adotada em razão do objetivo desta representação visual, neste trabalho, ser a identificação de atributos com falhas quando o registro está presente no banco de dados, sem representar o espaço (lacuna) daqueles períodos em que todos os sensores falharam.

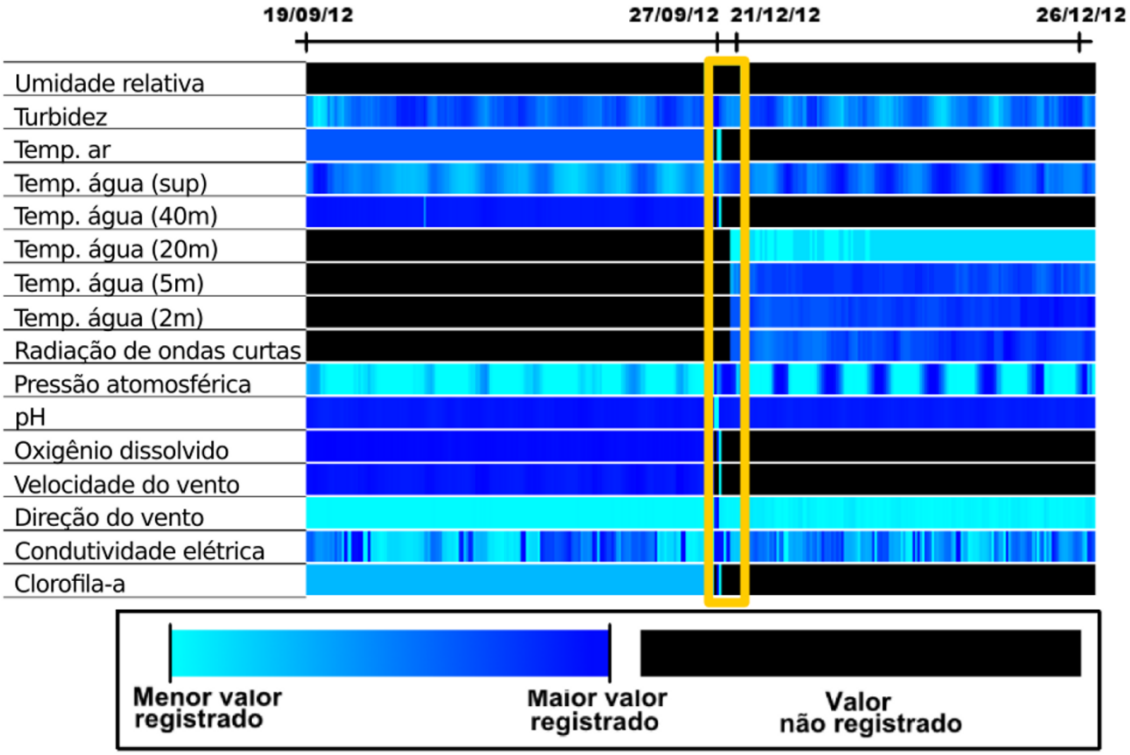

Figura 5: Visualização baseada em pixel de um período de dados da plataforma Três Marias. Em preto são os valores não registrados no banco de dados e o destaque em amarelo ressalta um intervalo sem a existência de nenhum registro no banco de dados

É possível selecionar intervalos de dados com falhas mínimas. Na Figura 6 pode ser visto um intervalo de dados da plataforma SIMA Funil 1, no qual podem ser notadas poucas falhas na captura de alguns atributos, indicada pela quase inexistência de registros com a cor preta. A plataforma Funil 1 possui um dos mais baixos índices de falhas, tanto considerando as falhas em coletas (19,05%) como dias completos sem coleta (1,2%). Além disso, explorando os dados utilizando a representação visual baseada em pixel, é possível notar que também existem poucas falhas nos atributos coletados - pouca existência de lacunas. Tal fato pode indicar uma informação sobre a qualidade dos dados coletados e armazenados.

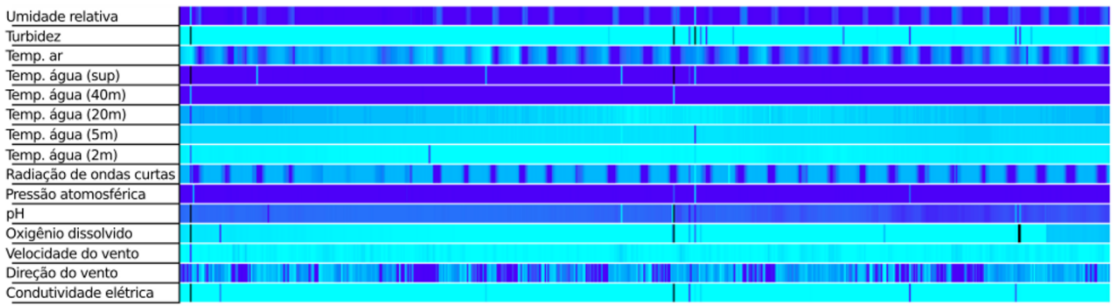

Figura 6: Representação visual baseada em pixel de um intervalo de dados de Funil 1 no qual poucos registros possuem falhas (marcados em preto).

A representação baseada em pixel não considera o eixo temporal como linear e não permite identificar lacunas quando todos os sensores falharam. Para extrair esta informação, a técnica de *Horizon Charts* pode ser utilizada para exibir os valores em uma escala de tempo linear. Na Figura 7 é apresentado o resultado da representação por *Horizon Chart* do mesmo intervalo de dados da plataforma de Funil 1 utilizado na representação baseada em pixel da Figura 6.

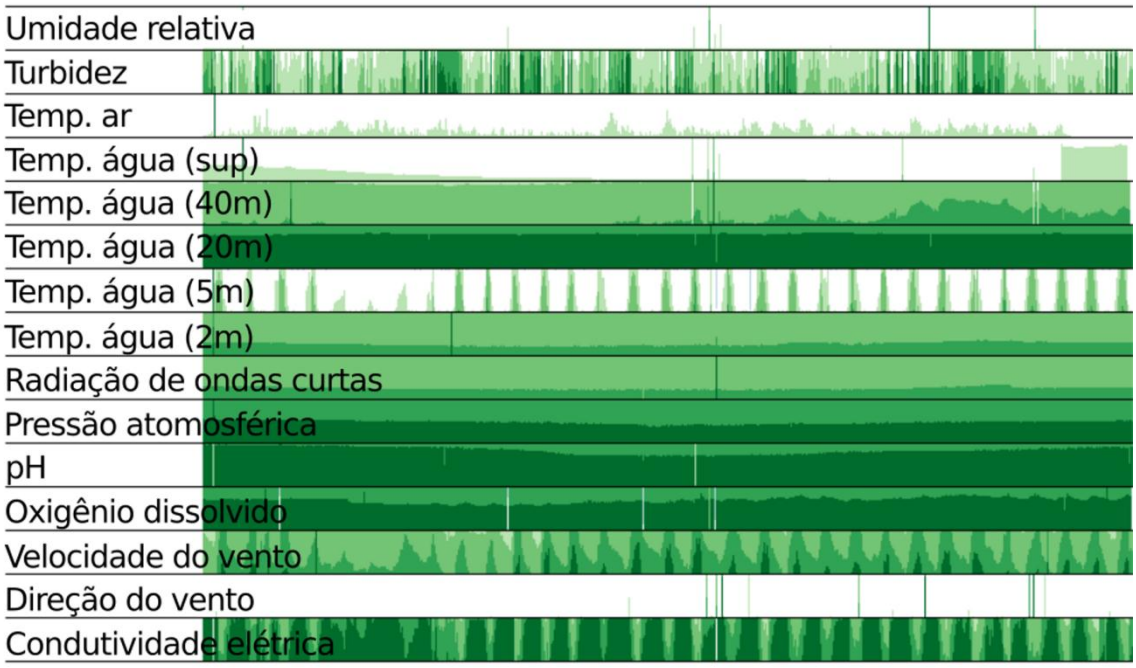

Figura 7: Representação visual utilizando *Horizon Charts* para um intervalo de dados da plataforma Funil 1, no qual poucos registros possuem falhas.

Os resultados apresentados na Figura 6 e Figura 7 poderiam indicar que existe falha em alguns atributos (turbidez, radiações de ondas curtas, oxigênio dissolvido, velocidade do vento e condutividade elétrica), em razão da invariabilidade dos valores e de espaços vazios. No entanto, tal fato ocorre em razão da existência de valores *outiliers* identificados pela análise das estatísticas básicas do conjunto, como pode ser visto na Tabela 2.

Tabela 2: Estatísticas descritivas do conjunto de dados selecionados da plataforma Funil 1. Destacados estão os dados que identificam a presença de *outliers*

Variáveis	Mínimo	Máximo	Variação	Média	Mediana	Variância	Desvio Padrão
Condutividade (mS/cm)	1,00	9935,00	***9934,00***	115,08	***65,00***	313635,93	678,53
Direção do vento (oNV)	0,00	358,59	358,59	152,91	122,34	8765,75	93,90
Velocidade do vento (m/s)	0,00	44,12	***44,12***	1,56	***1,18***	4,38	2,67
Oxigênio dissolvido (mg/l)	0,00	46,09	***46,09***	1,81	***0,54***	10,52	3,68
pH	0,00	13,96	13,96	7,38	7,12	0,86	0,96
Pressão atmosférica (hPa)	800,00	1017,18	217,18	966,02	967,22	144,68	12,19
Radiação de ondas curtas (W/m²)	-355,88	1489,01	***1844,89***	166,31	***7,36***	68767,62	267,17
Temp. água a 4,32m (ºC)	20,40	37,09	16,69	21,04	21,05	0,66	1,02
Temp. água a 9,17m (ºC)	18,46	36,96	18,50	20,78	20,66	0,51	0,95
Temp. água a 19,17m (ºC)	19,75	23,51	3,76	20,48	20,53	0,09	0,32
Temp. água a 39,17m (ºC)	5,26	20,53	15,27	19,72	19,75	1,10	1,05
Temp. água a 4,32m (ºC)	-5,00	22,94	27,94	21,04	21,22	5,08	2,26
Temp. ar (ºC)	11,25	29,33	18,08	18,79	18,16	11,83	3,46
Turbidez (NTU)	0,00	1000,00	***1000,00***	12,51	***2,94***	7077,18	92,54
Umidade relativa (%)	0,78	100,00	99,22	84,60	93,73	343,77	18,54

Existe alta variação entre os valores de mínimo e máximo de alguns atributos, destacados na Tabela 2, que indicam a existência de dados corrompidos (ruídos). Estes mesmos dados

interferem na formação das representações visuais e podem conduzir o analista a interpretações errôneas da informação apresentada. O ruído pode ser eliminado com a aplicação de um filtro, como pode ser visto na Figura 8 que apresenta os dados do atributo Velocidade do Vento utilizando as representações por *Horizon Chart* e baseada em pixel, destacando a posição do valor corrompido e sua composição após aplicação de um filtro para remoção do ruído. A remoção do ruído foi realizada a partir da eliminação do maior valor registrado para o atributo Velocidade do Vento, visto como 44,12 na Tabela 2, aplicando uma restrição a valores menores que o limiar desejado.

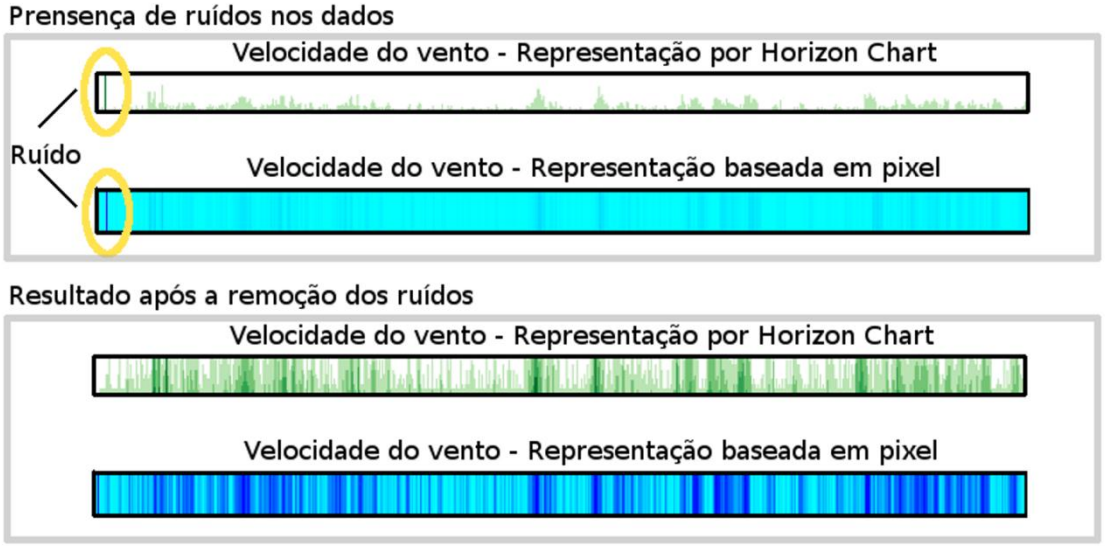

Figura 8: Representações baseadas em pixel e *Horizon Charts* da Velocidade do Vento da plataforma Funil 1, destacando a presença de ruídos (a) e após a remoção destes ruídos (b).

O processo de refinamento da consulta pode ser efetuado quantas vezes forem necessárias na própria interface do SimaVIS, acessada via WEB. Desta forma o analista pode reconfigurar a seleção do conjunto de dados, aplicando filtros combinados entre valores específicos de atributos e adequar o conjunto de amostras às suas necessidades dentro do processo de exploração e análise dos dados.

2.4 Padrões de falhas nos atributos do conjunto de dados

A existência de falhas nos atributos coletados e armazenados nos registros do banco de dados permite explorar uma possibilidade relacionada com a observação de padrões de falhas de grupos de atributos. A visualização por máscara de bits permite explorar a capacidade de identificação de padrões de comportamentos de falhas entre conjuntos de atributos que podem ser representados por valores dicotômicos (Carmo et al., 2013). Para tanto, cada combinação de falha possível para o conjunto de atributos de uma determinada plataforma, é mapeada para o espaço de cores, de forma que cada sequência de falha possa ser identificada e diferenciada através da cor.

Observando paralelamente ambas as representações visuais, baseada em pixel e por máscara de bits, do mesmo intervalo de dados referente a um intervalo de dados da plataforma Três Maria, no qual existem falhas em alguns atributos, é possível notar a relação entre a sequência de bits e os atributos que estão falhando simultaneamente, como pode ser visto na Figura 9, que são concomitantemente diferenciados por meio de sua cor.

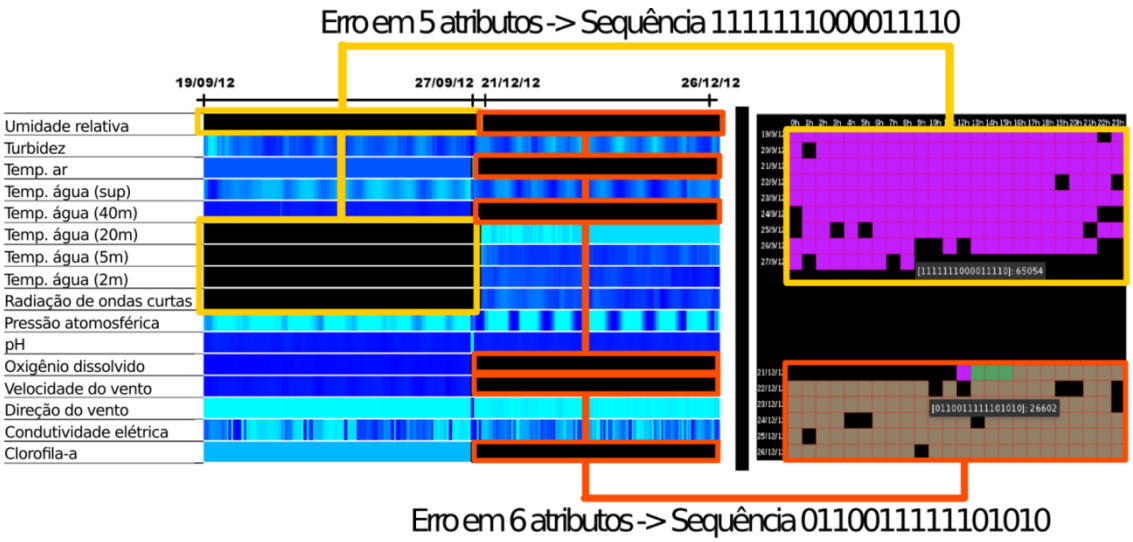

Figura 9: Apresentação simultânea de representação baseada em pixel e por máscara de bits. Em destaque estão os atributos com falhas, que definem a máscara de bits e a cor representativa: Sequência 1111111000011110 indica falha em cinco sensores e 0110011111101010 falha em outros seis sensores.

Na Figura 9, referente ao mesmo intervalo de dados de um intervalo de dados da plataforma Três Maria que existem falhas, estão destacados cinco atributos que possuem falhas no primeiro intervalo de tempo (antes de 27/09/12), e outros seis atributos que falham no segundo período de tempo (após 21/12/12), alterando a sequência de bits e o padrão de cores que identifica o conjunto de atributos com falhas. Adicionalmente, a representação baseada em pixel não registra os intervalos de tempo nos quais não existem dados, enquanto a representação por máscara de bits permite observar também estes intervalos vazios.

Na representação por máscara de bits, apresentada na Figura 9, a ordem dos atributos tem influencia direta no resultado visual, pois cada componente da sequência representa um único atributo, o qual define é utilizado posteriormente para transformação e definição no espaço de cores.

2.5 Características de plataformas SIMA fundeadas no mesmo reservatório

Existem cenários nos quais figuram mais de uma plataforma SIMA fundeada no mesmo reservatório. Alguns reservatórios possuem até três plataformas, como é o caso de Funil. Existe um total de dezessete plataformas SIMA que compartilham reservatórios.

No gráfico de intervalos apresentado na Figura 2, é possível notar dois reservatórios que possuem plataformas com coletas realizadas em um mesmo intervalo: Itumbiara e Ibitinga. Ibitinga possui três plataformas ativas com coletas de dados em intervalo de tempo simultâneo. Itumbiara possui também três plataformas, porém com todas atualmente desativadas e apenas duas em período de coleta de dados concomitantes enquanto estavam ativas.

Exceto Ibitinga 3, todas as outras plataformas em Ibitinga possuem um baixo nível de falhas, tanto referentes às coletas quanto ao acúmulo diário, como mostram os gráficos exibidos na Figura 3 e Figura 4. Observando a representação visual por máscara de bits, que incorpora um

intervalo de dados das três plataformas em Ibitinga, apresentada na Figura 10, é possível notar a existência de quatro padrões de falhas no intervalo compreendido entre março e agosto de 2014, indicados pelas cores das células da matriz: nenhum atributo coletado (cor preta); todos os atributos coletados (cor branca); falha no primeiro atributo (cor cinza) e falha no quinto atributo (cor roxa). Observando os metadados da nova modelagem do sistema SimaVIS, o primeiro atributo se refere à Clorofila A e o quinto atributo ao Oxigênio Dissolvido. A representação por máscara de bits permite observar alguns comportamentos de falhas que podem estar relacionados com a localização das plataformas e a capacidade de transmissão via enlace de satélite. Na Figura 10, para facilitar a leitura e interpretação, são aplicadas algumas marcações de comportamentos que podem ser destacadas.

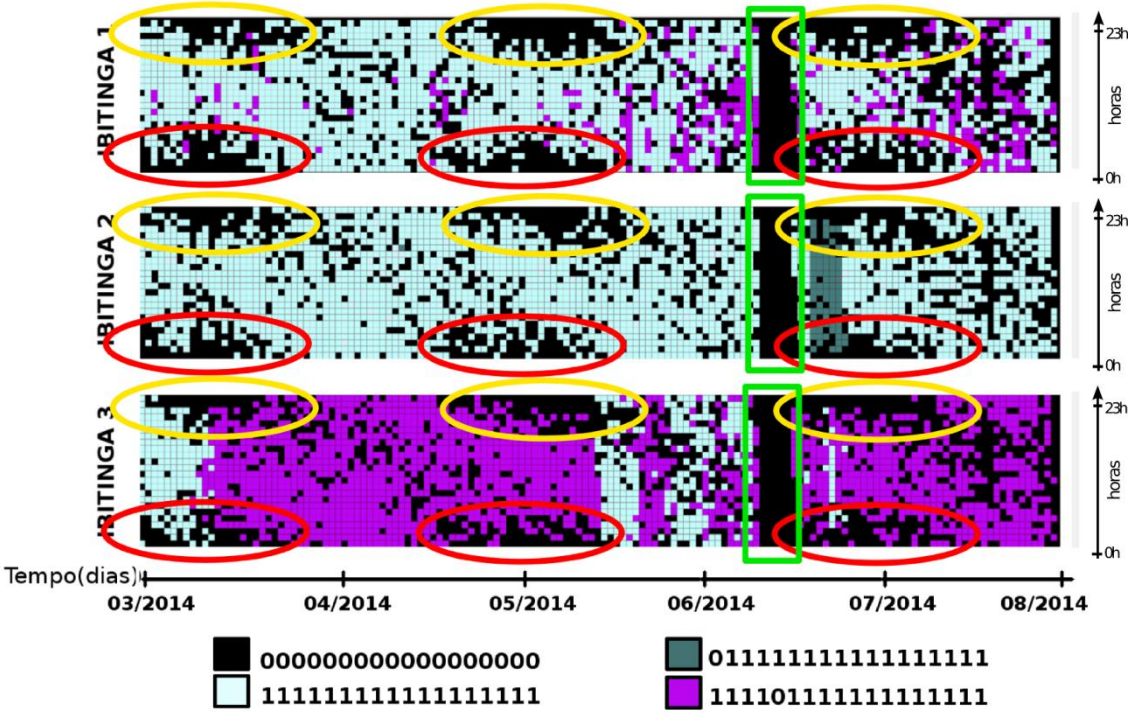

Figura 10: Representação por máscara de bits de um intervalo de dados das três plataformas fundeadas em Ibitinga. Valor 0 significa ausência de dado.

As marcações apontadas na Figura 10 destacam comportamentos que ocorrem simultaneamente de maneira similar nas três plataformas do reservatório de Ibitinga. Com elipses amarelas são destacados momentos de falha nas coletas dos dados que ocorrem, ciclicamente, em intervalos médios de 1,5 meses e em períodos compreendidos entre 17h e 24h do dia. Da mesma forma, nas três plataformas ocorrem falhas simultâneas no início de cada dia, de 0h as 5h, destacados com elipses vermelhas, com um ciclo aproximado de 1,5 meses. Embora no gráfico apareça segmentado, este evento pode ser agrupado em um único intervalo de tempo, considerando a linearidade e continuidade em que ocorre com início às 17h e fim às 5h do dia seguinte (período em que todas as coletas falham). Tal fato pode estar relacionado com problemas durante a transmissão dos dados coletados, já que para que a transmissão seja realizada deve existir um satélite dentro do campo de visão da plataforma, fator diretamente relacionado com a órbita do satélite.

A existência de períodos com falhas nos dados é comum em outros intervalos do conjunto de dados, como apontado por Valéria (2009), a qual utiliza um intervalo de oito meses de dados coletados pela plataforma SIMA de Manso/MT e aplica interpolação e análise por operador de fragmentação assimétrico em razão da existência de lacunas neste conjunto. Alcântara et al. (2013) comentam sobre as possíveis causas das falhas no conjunto de dados gerados pelo SIMA

e apontam duas abordagens aplicadas sobre o conjunto para minimizar o problema de dados faltantes: o processamento com médias móveis para preencher pequenas lacunas; e a computação de série de dados simulados a partir de outros períodos para o caso de longos períodos de mau funcionamento.

Ainda na Figura 10, o destaque com o retângulo verde mostra um intervalo de seis dias, registrados de 04/07/2014 a 10/07/2014, com falhas em todas as coletas, ou seja, nenhum dado foi registrado neste intervalo de tempo. Como é um fato que ocorre exatamente no mesmo período de tempo, tanto para o início como para o fim das falhas, a suposição considerada é de que seja um período em que as três plataformas foram colocadas em manutenção, desta forma, sendo desativadas temporariamente. Esta hipótese poderia ser confirmada mediante consulta a registros e históricos de cada plataforma, caso fossem disponibilizados juntamente com os conjuntos de dados, ressaltando a necessidade fundamental da criação e gerenciamento de metadados associados aos dados coletados.

3. Conclusão

Este trabalho abordou alguns recursos que podem ser utilizados para potencializar o processo de exploração e análise dos dados. A integração entre diferentes abordagens analíticas, técnicas conhecidas de representação visual dos dados, como gráficos interativos, *Horizon Charts* e outras variações de representações baseadas em pixel, se mostraram eficazes como abordagens para facilitar a extração e interpretação de informações, baseadas na capacidade de percepção e cognição associada ao sistema visual humano.

Os gráficos interativos facilitaram a extração de informações de forma rápida e intuitiva, em contraste com o esforço que seria necessário para extrair as mesmas informações a partir de dados tabulares. O período de tempo de operação das plataformas SIMA pôde ser facilmente identificado utilizando o gráfico de intervalos que também é capaz de ressaltar outras informações, como: encontrar as plataformas ativas que estão coletando dados em um determinado instante de tempo; e identificar as plataformas que possuem registros de dados em um mesmo intervalo de tempo.

Um fator que pode ser utilizado para representar a qualidade dos dados coletados, ou a eficiência da realização de coletas pelas plataformas, é o índice de falhas associado às coletas. Os gráficos radiais e por barras permitiram exprimir valores quantitativos relacionados à proporção de falhas e sucessos na realização da coleta, sendo efetivamente aplicados para a comparação entre a eficiência na realização das coletas entre diferentes plataformas.

Quanto à confiabilidade dos dados registrados, as representações visuais baseada em pixel e *Horizon Charts* permitiram observar o comportamento geral dos dados, relacionado à identificação de atributos com falhas no registro e sobre a variação dos dados para identificação de registros espúrios (ruídos). Estas representações facilitam a extração de informações sobre o conjunto de atributos efetivamente coletados em um determinado instante de tempo, bem como a variabilidade dos dados que pode ressaltar algum comportamento anormal ou inesperado.

A representação visual por máscara de bit permitiu a exploração dos dados em busca de comportamentos padronizados na ocorrência de erros entre um determinado conjunto de sensores. A representação visual por máscara de bit (Carmo et al., 2013), permite verificar se o conjunto de atributos com falhas simultâneas apresenta alguma forma de padrão sistemático que pode ser caracterizado e identificado ao longo do tempo de execução das coletas. Tal representação visual aborda um conceito que pode ser utilizado em outros cenários de exploração

e análise de dados, onde exista o interesse em representar comportamentos dicotômicos (dois estados, como: falha e sucesso) de dados multivariados.

As contribuições predominantes deste trabalho podem ser caracterizadas por benefícios a ação e intervenção de especialistas do domínio de aplicação, que podem obter informações preliminares sobre a caracterização do conjunto de dados de forma rápida e intuitiva. Tal conhecimento é de interesse direto de gestores do projeto que podem obter informações sobre o desempenho e qualidade relacionada aos dados coletados pelas plataformas, além de serem úteis para o desenvolvimento de novas plataformas que podem ser aprimoradas baseadas no padrão de falhas encontradas nas plataformas existentes.

Por fim, espera-se que os resultados obtidos e apresentados no decorrer deste trabalho possam fornecer subsídios, aplicáveis tanto para a área de Geociências quanto de Ciência da Computação, para amplificar e incentivar a adoção de padrões para definição de infraestrutura de dados de projeto que utilizem conjuntos de sensores (como o SIMA), bem como de estimular a utilização de técnicas de Analítica Visual como complemento e potencializador de outros procedimentos analíticos no processo de exploração e análise de dados.

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; 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 do SIMA.

REFERÊNCIAS BIBLIOGRÁFICAS

Alcântara, E., Curtarelli, M., Ogashawara, I., Stech, J. e Souza, A. (2013). A system for environmental monitoring of hydroelectric reservoirs in brazil. *Ambiente e Água - An Interdisciplinary Journal of Applied Science 8*(1), 6–17.

Carmo, A. F. C., Shimabukuro, M. H. e Alcântara, E. H. de (2013). Exploração visual interativa de dados coletados pelo sistema integrado de monitoramento ambiental - SIMA. In *XIV Brazilian Symposium on Geoinformatics*, GEOINFO 2013, pp. 127 – 132.

Dubelaar, Gerritzen, G. P. (2000). Cytobuoy: a step forward towards using flow cytometry in operational oceanography. *Scientia Marina 64*(2), 255–265.

Etcheber, H., Schmidt, S., Sottolichio, A., Maneux, E., Chabaux, G., Escalier, J.-M., Wennekes, H., Derriennic, H., Schmeltz, M., Quéméner, L., Repecaud, M., Woerther, P. e Castaing, P. (2011). Monitoring water quality in estuarine environments: lessons from the magest monitoring program in the gironde fluvial-estuarine system. *Hydrology and Earth System Sciences 15*(3), 831–840.

Glasgow, H. B., Burkholder, J. M., Reed, R. E., Lewitus, A. J., e Kleinman, J. E. (2004). Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies. *Journal of Experimental Marine Biology and Ecology 300*(1–2), 409 – 448.

Hamilton, G. D. (1986). National data buoy center programs. *Bulletin of the American Meteorological Society 67*(4), 411–415.

Jeong, D. H., Ziemkiewicz, C., Ribarsky, W. e Chang, R. (2009). Understanding principal component analysis using a visual analytics tool. *2009 US-Korea Conference on Science, Technology and Entrepreneurship*.

Keeler, B. L., Polasky, S., Brauman, K. A., Johnson, K. A., Finlay, J. C., O'Neill, A., Kovacs, K. e Dalzell, B. (2012). Linking water quality and well-being for improved assessment and valuation of ecosystem services. *Proceedings of the National Academy of Sciences 109*(45), 18619–18624.

Keim, D. A., Mansmann, F., Schneidewind, J., e Ziegler, H. (2006). Challenges in visual data analysis. In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pp. 9–16. IEEE.

Levkowitz, H. e Kelleher, C. (2012, Aug). Cloud and mobile web-based graphics and visualization. In *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2012 25th SIBGRAPI Conference on*, pp. 21–35.

Shafer, M. A., Fiebrich, C. A., Arndt, D. S., Fredrickson, S. E., e Hughes, T. W. (2000). Quality assurance procedures in the oklahoma mesonetwork. *Journal of Atmospheric and Oceanic Technology 17*(4), 474–494.

Stech, J., Lima, I., Novo, E., Silva, C., Assireu, A., Lorenzzetti, J., Carvalho, J., Barbosa, C. e Rosa, R. (2006). Telemetric monitoring system for meteorological and limnological data acquisition. *Internationale Vereinigung fur Theoretische und Angewandte Limnologie Verhandlungen 29*(4), 1747–1750.

Stech., J. L., Alcântara, E. H., Lorenzzetti, J. A. e Lima, I. B. T. de (2011). Uso de tecnologia espacial para coleta automática de dados limnologicos e meteorologicos: Aplicações nos reservatórios hidrelétricos de manso e corumbá. In *Novas tecnologias para o monitoramento e estudo de reservatórios hidrelétricos e grandes lagos*, Capítulo 4, pp. 119–162. Parêntese.

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

Valério, A. M (2009). Sensoriamento remoto orbital e de superfície para o estudo do comportamento de corpo de água do reservatório de Manso, MT, Brasil. Dissertação de Mestrado. Instituto Nacional de Pesquisas Espaciais (INPE).

Valerio, A. M., Kampel, M., Stech, J. L. e Assireu, A. T (2011). Variabilidade dos dados da bóia SIMA analisados pelo Operador de Fragmentação Assimétrico . In XV Simpósio Brasileiro de Sensoriamento Remoto (SBSR-2011), p. 5108

Ward, M., Grinstein, G. e Keim, D. (2010). *Interactive Data Visualization: Foundations, Techiniques, and Applications*. A K Peters, Ltd.

Recebido em Dezembro de 2015.

Aceito em Fevereiro de 2016.