

Alisson Fernando Coelho do Carmo

*Exploração e análise de dados coletados pelo
Sistema Integrado de Monitoramento Ambiental
utilizando técnicas de Visual Analytics*

Dissertação de mestrado

Presidente Prudente

2015

Alisson Fernando Coelho do Carmo

**Exploração e análise de dados coletados pelo Sistema
Integrado de Monitoramento Ambiental utilizando
técnicas de Visual Analytics**

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciências Cartográficas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de Presidente Prudente para obtenção do título de Mestre em Ciências Cartográficas.

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”

CÂMPUS DE PRESIDENTE PRUDENTE

FACULDADE DE CIÊNCIAS E TECNOLOGIA

Programa de Pós-Graduação em Ciências Cartográficas

Orientador: Prof. Dr. Milton Hirokazu Shimabukuro

Coorientador: Prof. Dr. Enner Herenio de Alcântara

Presidente Prudente

2015

FICHA CATALOGRÁFICA

C285e Carmo, Alisson Fernando Coelho do.
Exploração e análise de dados coletados pelo Sistema Integrado de Monitoramento Ambiental utilizando técnicas de Visual Analytics / Alisson Fernando Coelho do Carmo. - Presidente Prudente : [s.n.], 2015
121 f.

Orientador: Milton Hirokazu Shimabukuro
Coorientador: Enner Herenio de Alcântara
Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia
Inclui bibliografia

1. Sensores ambientais. 2. Dados temporais. 3. Visual Analytics. I. Shimabukuro, Milton Hirokazu. II. Alcântara, Enner Herenio de. III. Universidade Estadual Paulista. Faculdade de Ciências e Tecnologia. IV. Título.

BANCA EXAMINADORA



Prof. Dr. MILTON HIROKAZU SHIMABUKURO
(ORIENTADOR)



Dr. RAMON MORAIS DE FREITAS
(Camargo-Schubert Engenharia Eólica)



Profa. Dra. MARIA DE LOURDES B. T. GALO
(UNESP/FCT)



ALISSON FERNANDO COELHO DO CARMO

Presidente Prudente (SP), 23 de fevereiro de 2015.

Resultado: APROVADO

À toda minha família e amigos.

AGRADECIMENTOS

Primeiramente, quero agradecer a Deus por permitir a realização deste trabalho por meio do dom da vida que nos presenteia a cada dia.

Gostaria de agradecer também a todas as pessoas de minha família, que jamais pouparam quaisquer esforços que pudessem ser direcionado para ajudar em todos os momentos de minha vida e trajetória acadêmica e, principalmente, compreenderam o sacrifício de tempo que precisei dedicar aos estudos. Em especial agradecer aos meus pais João e Rozangela e aos meus avós Manoel e Izabel pelo apoio incondicional.

Agradeço aos responsáveis por me acompanhar e me guiar durante todo o trabalho, presentes em cada passo dos desafios acadêmicos, meus orientadores Milton e Enner. Devo direcionar menção especial ao professor Milton, o qual sempre esteve ao meu lado, desde o início da graduação e felizmente se transformou em um amigo, professor e orientador.

Não poderia deixar de agradecer a todos meus amigos que compartilharam diversos momentos de alegria e angústia, desespero e alívio, apreensão e recompensa, sem os quais, não tenho dúvida de que toda esta caminhada seria mais difícil, dentre os quais, posso citar os amigos de graduação (Ayrton, Camila, Orlindo, Douglas, Dani, Maila, Ingrid, Arthur e todos os outros), os amigos de projetos (Klaus, Elisa, Dani, Denise, Luiza, Erik, Gabriel, João), os amigos da pós-graduação (Nariane, Fernanda, Luiz, Renato, Ricardo, Bruno), todos os outros amigos que participaram de alguma forma (Nati, João, André, Rafael). Também agradeço à toda equipe do Futebol e Tênis de Mesa que muitas vezes foram os responsáveis pelos momentos de distração e recarga de energias.

De modo geral, agradeço a toda FCT/UNESP que me acolheu e se transformou em minha casa desde 2008, bem como a todo o seu corpo docente, funcionários e direção.

Antecipadamente peço desculpas pelos que não mencionei, mas tenham a certeza de que os trago em meu coração, e os agradeço por estarem comigo nessa jornada.

*“ Os problemas significativos com os quais nos deparamos
não podem ser resolvidos no mesmo nível de pensamento
em que estávamos quando eles foram criados
(Albert Einstein)*

RESUMO

O processo de intervenção do homem no meio ambiente tem grande influência sobre as alterações que nele ocorrem. É importante monitorar características e fenômenos envolvidos neste processo, visando identificar, interpretar e controlar tais alterações. A identificação de fenômenos que variam com rapidez exige dados em alta resolução temporal, fato que pode prejudicar a utilização de Sensoriamento Remoto, pois a frequência de obtenção de dados de uma mesma região pode ser insuficiente. Por outro lado, o estudo de alterações causadas por processos lentos é realizado mediante a análise de longas séries históricas, construídas por meio de aquisições periódicas de dados por um longo período, podendo inviabilizar metodologias de coletas em campo, em razão de dificuldades associadas à logística para execução de tais coletas. O desenvolvimento tecnológico tem impulsionado a utilização de sensores para a realização de coletas automatizadas e periódicas de dados, como o Sistema Integrado de Monitoramento Ambiental (SIMA), cujo conjunto de dados é utilizado neste trabalho. O projeto SIMA é composto por um conjunto de plataformas fundeadas em reservatórios hidrelétricos, capazes de coletar automática e periodicamente atributos do ar e da água. No entanto, podem existir falhas no conjunto de dados causadas por problemas na coleta, na transmissão ou no armazenamento dos registros. Em razão da existência de grande quantidade de dados temporais multivariados e pela possibilidade de falhas, a exploração visual e interativa auxilia na caracterização do conjunto. Neste trabalho, foram investigadas abordagens com a aplicação de técnicas de Visualização de Informação na extração de características do conjunto de dados, as quais impactam a qualidade da análise dos fenômenos associados. Padrões para modelagem e disponibilização dos dados, visando a flexibilidade no desenvolvimento de aplicações e a interoperabilidade entre recursos, também foram avaliados. Os resultados obtidos demonstram a potencialidade da utilização de representações visuais e interativas para a exploração do conjunto de dados, as quais facilitam a percepção de informações acerca das seguintes características: disponibilidade dos dados, funcionamento dos sensores, e evidências de padrões de falhas. As contribuições deste trabalho estão relacionadas com a abordagem para amplificar e facilitar o processo de exploração e análise de conjunto de dados temporais e multivariados, bem como aumentar a interoperabilidade, flexibilidade e reusabilidade de recursos relacionados a conjuntos de dados coletados por sensores. Os resultados alcançados podem subsidiar e motivar a realização de trabalhos futuros para a ampliação da utilização das representações visuais e interativas para a análise dos significados dos dados e dos fenômenos associados, além da caracterização do conjunto de dados, como abordada neste trabalho.

Palavras-chave: Sensores ambientais. Dados temporais multivariados. Visualização Exploratória. Visual Analytics. Padrões para interoperabilidade.

ABSTRACT

The man intervention process in the environment has huge influence on the changes that occur in it. It is important monitoring the features and phenomena involved in this process to identify, interpret and control such changes. The identification of fast varying phenomena requires a high temporal resolution dataset, what can constrain remote sensing utilization, due to insufficient data collection rate. On the other hand, the analysis of changes caused by slow processes is done from long historical records, which come from a periodic data acquisition during a long period of time, which can make traditional technique, such as fieldwork campaigns, an inappropriate approach due to the rate of collection. The technological development has been raising the use of automated sensors to collect data, such as the Environmental Monitoring Integrated System (Sistema Integrado de Monitoramento Ambiental, SIMA) that is the source of dataset used in this project. In the SIMA project, a set of platforms, which are capable of collecting air and water attributes automatically and periodically, are anchored to hydroelectric power dams. However, the dataset may be incomplete because of problems in one or more of the following phases: acquisition, transmission or storing. Due to huge amount of multivariate temporal data and the lack of dataset in some cases, the visual and interactive exploration can support the task of its characterization. In this project, Information Visualization techniques are applied to dataset characterization process, searching for features that may affect the quality of the associated phenomena analysis. Standards for data modelling and availability, aiming the flexibility for application development and interoperability among resources, have also been investigated. The benefits of using visual and interactive representations to explore the dataset are demonstrated by the results achieved in this project. The perception of some features of dataset is facilitated, such as: data availability, sensors operation, and pattern on missing data. The contributions of this project are related to the approach to amplify and facilitate the process of exploration and analysis of temporal and multivariate dataset and increase the interoperability, flexibility and reusability of resources related to data collected by sensors. The achieved results can support and motivate future works to expand the use of visual and interactive representations for the analysis of the meanings of the data and the associated phenomena, besides the characterization of the dataset, as investigated in this project.

Keywords: Environmental sensors. Multivariate temporal data. Exploratory Visualization. Visual Analytics. Standards for interoperability.

LISTA DE ILUSTRAÇÕES

Figura 1 – Grafo LOD em 2007 com 12 conjuntos de dados conectados (SCHMA-CHTENBERG et al., 2014)	33
Figura 2 – Grafo LOD em 2009 com 93 conjuntos de dados conectados (SCHMA-CHTENBERG et al., 2014)	33
Figura 3 – Grafo LOD em 2014 com 570 conjuntos de dados conectados (SCHMA-CHTENBERG et al., 2014)	34
Figura 4 – Arquitetura com os principais serviços OGC. Adaptado de Percivall (2014)	35
Figura 5 – Esquema ilustrativo de integração de sensores no escopo da especificação SWE (REED et al., 2007)	37
Figura 6 – Componentes do <i>framework</i> SWE (BRÖRING et al., 2009)	37
Figura 7 – Elementos envolvidos no processo de requisição SOS, destacando o escopo de requisições do usuário ao <i>Web Service</i> . Adaptado de (REED et al., 2007)	39
Figura 8 – Diagrama de Sequência representando uma requisição de um usuário ao serviço SOS (NETWORK, 2014)	40
Figura 9 – Relação hierárquica entre os componentes SOS. Os componentes da especificações não foram traduzidos. Adaptado de (BRÖRING; STASCH; ECHTERHOFF, 2012)	41
Figura 10 – Tempo consumido para recuperação de observações de acordo com a quantidade de sensores (POORAZIZI; LIANG; HUNTER, 2012)	44
Figura 11 – Espaço de memória demandado para recuperação de observações de acordo com a quantidade de sensores (POORAZIZI; LIANG; HUNTER, 2012)	44
Figura 12 – Modelo de referência para Visualização de Informação. Adaptado de Heer e Agrawala (2007)	47
Figura 13 – Áreas de pesquisas relacionadas com o campo de Visualização de Informação (KEIM et al., 2010)	48
Figura 14 – Relações de <i>Visual Analytics</i> com outros campos de pesquisas (KEIM et al., 2006)	50
Figura 15 – Diagrama de <i>Sense-making Loop</i> ressaltando o ciclo de interatividade. Adaptado de Keim et al. (2006)	50

Figura 16 – Exemplos de técnicas de Visualização de Informação baseadas em pixel: (a) Representação em uma estrutura de calendário (LAMMARSCH et al., 2009); (b) Representação em <i>Circle Segments</i> (KEIM, 2000) e (c) Representação por <i>Heatmap</i> (BORGO et al., 2010).	52
Figura 17 – Construção de um <i>Horizon Chart</i> a partir de um gráfico de linha, ilustrado em três passos. Adaptado de Heer, Kong e Agrawala (2009) .	53
Figura 18 – Apresentação de registros multivariados de uma série temporal, onde cada linha representa uma variável. Adaptado de Aigner et al. (2011) .	54
Figura 19 – Equivalência entre um registro composto por n atributos e sua representação em Coordenadas Paralelas. Adaptado de (INSELBERG; DIMSDALE, 1990)	55
Figura 20 – Padrões comuns apresentados no plano cartesiano e no plano de coordenadas paralelas. Adaptado de (HEINRICH; WEISKOPF, 2012)	56
Figura 21 – Representação visual por Coordenadas Paralelas de um conjunto de dados com informações sobre carros. Adaptado de (YUAN et al., 2009)	56
Figura 22 – Representação de um conjunto de dados composto por cinco variáveis relacionados a carros. A distância entre os eixos representa a correlação entre as variáveis vizinhas (HAUSER; LEDERMANN; DOLEISCH, 2002)	57
Figura 23 – Processo de transformação de cada atributo do conjunto em variáveis dicotômicas	59
Figura 24 – Discretização da sequência binária para o espaço de cores	60
Figura 25 – Exemplo de representação visual por máscara de bits. Na matriz, cada dia é representado por uma linha, cada coluna indica o intervalo de 1 hora e cada cor indica a conversão da sequência de bits para o espaço de cor	61
Figura 26 – Estação SIMA composta por uma boia toroidal de fibra de vidro (INPE, 2014)	64
Figura 27 – Estação SIMA composta por uma plataforma flutuante em alumínio naval (INPE, 2014)	64
Figura 28 – Mapa das localizações das 27 plataformas SIMA, nomeadas como: Balbina, Corumbá, Curuai, Estreito, Funil 1, Funil 2, Funil 3, Furnas - Embrapa, Furnas 1, Ibitinga 1, Ibitinga 2, Ibitinga 3, Itaipu, Itumbiara 1, Itumbiara 2, Itumbiara 3, Mamirauá, Manso 1, Manso 2, Mascarenhas de Moraes, Segredo, Serra da Mesa 1, Serra da Mesa 2, Três Marias, Tucuruí 1, Tucuruí 2, Xingó	65
Figura 29 – Arquitetura do projeto SIMA (INPE, 2014)	66
Figura 30 – Ação do tempo sobre o estado de uma sonda, extraída de uma plataforma SIMA, devido à exposição ao ambiente (INPE, 2014)	66

Figura 31 – Diagrama simplificado deduzido a partir das estruturas de dados que compõem o domínio de informações do SIMA	68
Figura 32 – Arquitetura modularizada do sistema SimaVis, formada por módulo Cliente, Módulo servidor e Bases de Dados	71
Figura 33 – Interface Cliente do sistema SimaVis, na qual é possível: (a) selecionar as plataformas que se deseja analisar; (b) escolher os atributos/variáveis de interesse; (c) limitar o intervalo de tempo para consultar os dados; (d) aplicar filtros de atributos; (e) definir qual a representação visual que se deseja utilizar e (f) exportar relatórios contendo as representações visuais e comentários	72
Figura 34 – Modelo EAV para os dados importados do projeto SIMA	74
Figura 35 – Lista de tabelas que compõem o Schema criado automaticamente pelo servidor 52°North-SOS para receber os dados	75
Figura 36 – Fragmento da requisição XML gerada pelo 52°North-SOS-Importer para importar dados de arquivos CSV no servidor 52°North-SOS	76
Figura 37 – Diagrama Entidade Relacionamento que define o modelo de dados utilizado. As plataformas SIMA são vistas como múltiplas fontes de dados e podem haver quantas forem necessárias	77
Figura 38 – Mapa de localização das plataformas SIMA. A seleção das plataformas é sincronizada com a composição dos filtros de seleção dos dados. Em destaque estão as plataformas selecionadas, tanto pelo mapa, quanto pelos filtros	80
Figura 39 – Gráficos construídos no SimaVis, gerados a partir de dados simulados: (a) gráfico de barras; (b) gráfico radial e (c) gráficos de intervalos. O gráfico de intervalos considera o primeiro e o último dia de coleta de cada plataforma, enquanto o gráfico radial e de barras representam a proporção, respetivamente, de falhas ocorridas nas coletas e de dias sem coleta de dados	81
Figura 40 – Visualização baseada em pixel implementada pelo SimaVis, gerada com dados simulados	81
Figura 41 – Visualização <i>Horizon Charts</i> utilizado no SimaVis, gerada com dados simulados	82
Figura 42 – Visualização de Coordenadas Paralelas no SimaVis, gerada com dados simulados. As cores representam o atributo tempo definido de forma relativa no primeiro eixo	83
Figura 43 – Visualização por máscara de bit implementada no SimaVis, gerada com dados simulados	83
Figura 44 – Períodos de atividades das plataformas SIMA	84

Figura 45 – Identificação das plataformas SIMA que estiveram ativas simultaneamente no início de 2012. Em amarelo são destacadas as plataformas que atendem a esta restrição	85
Figura 46 – 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.	86
Figura 47 – Quantidade de falhas das plataformas SIMA: (a) Serra da Mesa 1 com poucas falhas; (b) Itaipu com muitas falhas; (c) Tucuruí 1 com quantidade similar entre falhas e acertos	88
Figura 48 – 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.	88
Figura 49 – Visualização baseada em pixel de um período de dados da plataforma Três Marias. Em preto encontram-se os valores dos atributos não registrados no banco de dados (valores nulos) e o destaque em amarelo ressalta um intervalo sem a existência de nenhum registro no banco de dados	89
Figura 50 – Representação visual baseada em pixel de um intervalo de dados de Funil 1 no qual poucos registros possuem falhas (marcados em preto) .	90
Figura 51 – Representação visual, utilizando <i>Horizon Charts</i> , para um intervalo de dados da plataforma Funil 1 no qual poucos registros possuem falhas. .	91
Figura 52 – Presença de ruído nos valores dos atributos da velocidade do vento, identificáveis nas representações baseada em pixel e <i>Horizon Charts</i> . .	92
Figura 53 – Representações visuais baseada em pixel e <i>Horizon Charts</i> do atributo Velocidade do Vento da plataforma Funil 1 após a remoção do ruído . .	92
Figura 54 – Visualização por máscara de bits do início do intervalo de dados da plataforma Três Marias. O padrão de cores predominante (1111111000011110) indica erro em cinco sensores	93
Figura 55 – Visualização por máscara de bits do fim do intervalo de dados da plataforma Três Marias. O padrão de cores predominante (0110011111101010) indica erro em outros seis sensores	93
Figura 56 – 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 5 sensores e 0110011111101010 falha em outros 6 sensores	94
Figura 57 – Intervalos de tempo ativos das plataformas que compartilham reservatórios. Em destaque estão aquelas plataformas do mesmo reservatório que produzem dados em um mesmo intervalo de tempo	95
Figura 58 – Horas de atividade das plataformas	99

Figura 59 – Dias ativos das plataformas	99
Figura 60 – Representação por máscara de bits de um intervalo de dados das três plataformas fundeadas em Ibitinga. Valor 0 significa ausência de dado.	100
Figura 61 – Série temporal dos dados de temperatura da água. A proximidade das curvas indica possível alta correlação entre os dados	103
Figura 62 – Scatter Plot dos dados de temperatura da água. Linearidade do espalhamento indica possível alta correlação entre os dados	103
Figura 63 – Representação visual por Coordenadas Paralelas. A simetria entre as arestas que conectam os eixos confirma a alta correlação entre a variabilidade dos dados. A cor está relacionada com o atributo temporal, exibido no primeiro eixo	104
Figura 64 – Representação visual por Coordenadas Paralelas exibindo todos os atributos coletados pela plataforma Itaipu sem ordenação dos eixos, dificultando a interpretação	104
Figura 65 – Componentes Principais dos dados da temperatura da água de um conjunto de dados extraído de Itaipu	105
Figura 66 – Comportamento do novo fator temperatura da água, obtido com a aplicação da técnica PCA, de um conjunto de dados extraído de Itaipu	105
Figura 67 – Exemplo de percurso sobre o dígrafo representado pela matriz de correlação para definir a ordem dos atributos na representação por Coordenadas Paralelas	107
Figura 68 – Representação em Coordenadas Paralelas dos dados selecionados, após sintetização dos atributos de temperatura da água e reordenação dos eixos de acordo com a matriz de correlação	107

LISTA DE TABELAS

Tabela 1	– Tabela contendo as funções oferecidas por cada componente SOS . . .	42
Tabela 2	– Variáveis capturadas pelo conjunto de plataformas SIMA	64
Tabela 3	– Operações implementadas no <i>Web Service</i> SimaVis	78
Tabela 4	– Tabela contendo os períodos de atividades de cada plataforma	85
Tabela 5	– Tabela contendo a quantidade relativa de falhas das coletas	87
Tabela 6	– Estatísticas descritivas do conjunto de dados selecionados da plataforma Funil 1. Destacados estão os dados que identificam a presença de outliers (Variação discrepante em relação às medidas de tendência central) . . .	92
Tabela 7	– Reservatórios com mais de uma plataforma SIMA	95
Tabela 8	– Atributos coletados e respectivas plataformas SIMA. Marcados com X indica a presença na plataforma. Aqueles destacados em vermelho indicam a inexistência em uma das plataformas deste reservatório . . .	97
Tabela 9	– Estatísticas descritivas para o mesmo intervalo de dados das plataformas SIMA do reservatório de Ibitinga	101
Tabela 10	– Métricas da estatística descritiva sobre o conjunto de dados da plata- forma Itaipu	102
Tabela 11	– Matriz de correlação dos dados de um conjunto extraído de Itaipu . . .	106
Tabela 12	– Síntese das características das técnicas de representações visuais utilizadas	110

LISTA DE ABREVIATURAS E SIGLAS

AJAX	Asynchronous Javascript and XML
API	Application Programming Interface
AS	Abstract Specifications
CGI	Common Gateway Interface
CSS	Cascading Style Sheets
CSV	Comma-separated values
D3	Data-Driven Documents
DER	Diagrama Entidade Relacionamento
DM	Data Mining
DOM	Document Object Model
EAV	Entity-Attribute-Value
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
INPE	Instituto Nacional de Pesquisas Espaciais

IPCA	Interactive Principal Component Analysis
IS	Implementation Standards
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
LOD	Linked Open Data
O&M	Observations and Measurements
OGC	Open Geospatial Consortium
PCA	Principal Component Analysis
PCD	Plataformas de Coletas de Dados
RDF	Resource Description Framework
RGB	Red Green Blue
SAS	Sensor Alert Service
SDI	Spatial Data Infrastructure
SensorML	Sensor Model Language
SGBD	Sistema Gerenciador de Banco de Dados
SIMA	Sistema Integrado de Monitoramento Ambiental
SOS	Sensor Observation Service
SPS	Sensor Planning Service
SR	Sensoriamento Remoto
SSW	Semantic Sensor Web
SWE	Sensor Web Enablement
TC	Technical Committee
TML	Transducer Markup Language
UNIVAP	Universidade do Vale do Paraíba
URI	Uniform Resource Identifier

URL	Uniform Resource Locators
VA	Visual Analytics
VDM	Visual Data Mining
WNS	Web Notification Service
XML	eXtensible Markup Language

SUMÁRIO

1	INTRODUÇÃO	21
1.1	Descrição do Problema e Justificativa	23
1.2	Hipótese	25
1.3	Objetivos	25
1.3.1	Objetivo Geral	25
1.3.2	Objetivos Específicos	26
1.4	Organização do Documento	26
2	AQUISIÇÃO E GERENCIAMENTO DE DADOS AMBIEN- TAIS OBTIDOS POR SENSORES	27
2.1	Abordagens para coleta de dados ambientais	28
2.2	Publicação e interoperabilidade de conjuntos de dados e serviços	30
2.3	<i>Open Geospatial Consortium</i> - OGC	35
2.4	OGC <i>Sensor Web Enablement</i> - SWE	36
2.4.1	<i>Sensor Observation Service</i> - SOS	38
2.4.2	<i>Frameworks</i> e servidores compatíveis com SOS	41
3	EXPLORAÇÃO E ANÁLISE VISUAL DE DADOS	46
3.1	<i>Visual Analytics</i> e conceitos relacionados	47
3.2	Representação visual de dados temporais	51
3.2.1	Representação visual por máscara de bits	57
4	ESPECIFICAÇÕES DO CONJUNTO DE DADOS	62
4.1	Sistema Integrado de Monitoramento Ambiental - SIMA	63
4.1.1	Características do processo de coleta dos dados	65
4.1.2	Estrutura dos dados coletados	67
5	RESULTADOS E DISCUSSÕES	69
5.1	Arquitetura do SimaVis	70
5.1.1	Definição e implementação da infraestrutura de manipulação do conjunto de dados	72
5.1.1.1	Modelo Entidade-Atributo-Valor - EAV	73
5.1.1.2	Padrão OGC <i>Sensor Observation Service</i> - SOS	74
5.1.1.3	Modelagem compatível entre conjunto de dados heterogêneos	77
5.1.2	Recursos de exploração e análise de dados implementados no software Cliente	78
5.1.3	Técnicas de Visualização de Informação implementadas	79

5.2	Tempo de atividade das plataformas SIMA	83
5.3	Quantidade de dados faltantes	86
5.4	Falhas nos atributos das coletas registradas	89
5.5	Características de plataformas SIMA fundeadas no mesmo reserva- tório	94
5.6	Métodos estatísticos complementados por representações visuais . .	102
5.7	Integração dos resultados das técnicas de representações visuais . .	108
6	CONCLUSÕES	111
	Referências	114

CAPÍTULO 1

INTRODUÇÃO

A necessidade de registrar, monitorar e entender os fenômenos e os comportamentos associados ao meio ambiente sempre esteve presente no cotidiano do homem e é um de seus objetos de estudo. Fenômenos naturais e, principalmente, ações antrópicas sobre o meio ambiente têm grande influência sobre seu comportamento. A necessidade do registro e observações das características ambientais é cada vez maior para que seja possível interpretar as mudanças ocorridas e controlá-las prontamente. Tal tarefa tem se beneficiado do desenvolvimento tecnológico, principalmente relacionado à evolução de mecanismos para comunicação e dispositivos sensores utilizados para extração de dados associados a objetos de interesse.

Para monitorar o comportamento de qualquer cenário é preciso realizar observações por meio de coletas de atributos capazes de descrever suas características. Existem diversos atributos que podem ser medidos e monitorados para identificar padrões e anomalias que podem afetar o cenário ambiental. As características de ambientes aquáticos oferecem um grande conjunto de variáveis representativas que podem responder direta ou indiretamente sobre o estado de uma determinada região e integrar recursos para o monitoramento ambiental.

Existem diferentes alvos de estudo que podem ser observados para extrair informações sobre o ambiente desejado. A metodologia utilizada para a extração das informações também pode ser variada e depende, sobretudo, do alvo a ser observado.

Atualmente, em razão do desenvolvimento da área de Sensoriamento Remoto (SR), as imagens produzidas por satélites orbitais e fotografias geradas por outros dispositivos imageadores aéreos, têm se consolidado como recursos importantes para a obtenção de dados referentes às características da superfície. Os sensores remotos oferecem uma alternativa

para a análise que, agregada às metodologias de coletas clássicas de levantamentos e amostragem locais (*in situ*), representam as principais abordagens de obtenção de dados.

A utilização de sensoriamento remoto e a coleta local de dados apresentam aspectos que podem interferir em seu potencial de análise. No escopo de análise de ambientes aquáticos, Stech. et al. (2011) afirmam que para o total entendimento dos processos físicos, químicos e biológicos que agem sobre estes ambientes é necessário a manipulação de séries temporais tão grandes quanto possíveis e com vários atributos meteorológicos e limnológicos. Para permitir a identificação de determinados fenômenos que acontecem rapidamente, é necessário uma alta resolução temporal de dados, ou seja, coletas efetuadas frequentemente de maneira constante. Já para analisar fenômenos mais lentos, é necessário um longo registro histórico que permita monitorar sua ação gradativa.

Motivada pela evolução dos dispositivos sensores, infraestruturas de processamento e comunicação, uma abordagem que está sendo crescentemente utilizada é a coleta automática e periódica de dados utilizando plataformas autônomas. O Sistema Integrado de Monitoramento Ambiental (SIMA) é uma das abordagens que utiliza Plataformas de Coleta de Dados. Tais plataformas estão fundeadas em reservatórios hidrelétricos para a coleta de diversos atributos relacionados ao ar e à água, constituído por um conjunto de tecnologias aplicadas à coleta de dados e monitoramento da hidrosfera (INPE, 2014). As plataformas SIMA realizam a leitura dos sinais dos sensores periódica e constantemente ao longo do dia. Após a leitura, os dados coletados são primeiramente armazenados em um *buffer*¹ local e posteriormente são transmitidos, via enlace de satélite, para servidores em estações terrestres. Os dados são então analisados para verificação da existência de erros na transmissão dos sinais, passando pelo processo de decodificação, processamento e armazenamento, para ficarem disponíveis, mediante acesso autorizado, em um portal da internet.

A periodicidade de obtenção de dados oferece uma nova vertente para análise, possibilitando a integração do atributo tempo com os valores coletados, originando longas séries temporais que representam o comportamento histórico de diversas variáveis. 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. Neste sentido, os dados coletados pelo SIMA fornecem informações importantes para o monitoramento ambiental, principalmente em razão da constante coleta dos dados, que podem caracterizar o contexto de localização de cada plataforma.

O resultado da análise dos dados pode ser enriquecido de acordo com a dimensão do conjunto utilizado e a variedade de atributos manipulados; por outro lado, a capacidade de extração e interpretação de informações presentes em grandes conjuntos de dados, prin-

¹ Buffer é uma memória de tamanho reduzido, geralmente utilizada para armazenamento temporário de dados para acesso imediato

principalmente multivariados, pode exigir processamentos mais robustos e outras abordagens para viabilizar a exploração. Diante desses fatos, a análise exige a utilização de recursos computacionais para suportar a tarefa de extração de novas informações e facilitar a interpretação e descoberta de conhecimento. Estes recursos computacionais para a descoberta de conhecimento em bases de dados, referenciado na literatura como *Knowledge Discovery in Databases* (KDD), dispõem de técnicas para a busca e processamento de dados e se configuram como ferramentas essenciais para viabilizar a exploração e análise dos dados.

Com o apoio computacional, os dados podem ser processados, sintetizados, analisados, comparados e interpretados empregando diversas técnicas. Uma das abordagens para apoiar a esse processo é por meio de representações visuais dos dados, utilizando técnicas de Visualização de Informação combinadas com algoritmos de Mineração de Dados, abordagem conhecida como Mineração Visual de Dados (*Visual Data Mining*). A Mineração de Dados é responsável pela busca e análise dos dados primitivos, com o objetivo de extrair informações dos mesmos. A partir das informações extraídas, a análise pode ser potencializada com a utilização de recursos visuais representativos, por meio de técnicas de Visualização de Informação, e com uma interface rica e interativa para facilitar o processo analítico.

Um assunto que vem sendo discutido e desenvolvido constantemente é a utilização de recursos de interatividade que priorizem a presença do analista no processo de exploração e análise visual, configurando-se como uma das formas de potencializar o processo de análise por meio de técnicas de *Visual Analytics* (VA) que integram as possibilidades de Visualização de Informação com os ambientes ricos e interativos. Abordagens como capacidade de aplicação de filtros nos dados, operações de seleção dos objetos visuais, navegação e exploração por zoom e reconstrução das representações visuais, podem amplificar as possibilidades das tarefas analíticas. Tais técnicas podem ser utilizadas para complementar outras abordagens de análise, pois facilitam a identificação de comportamentos dos valores dos dados e aproveitam a capacidade de percepção visual humana.

1.1 Descrição do Problema e Justificativa

Diferentes abordagens podem ser utilizadas para o levantamento e coleta de informações relacionadas às alterações que ocorrem no meio ambiente. A coleta de dados por amostragem local, por meio de levantamentos *in situ*, com uma grande periodicidade se torna altamente custosa e trabalhosa, podendo inviabilizar sua execução. Devido a resolução temporal, espacial e espectral de alguns sensores, a utilização de tecnologias de sensoriamento remoto pode se tornar limitada para monitoramento de recursos que requerem observações e coletas frequentes de diversos atributos integrados, como citado por Ritchie, Zimba e Everitt (2003), que concluem que a integração entre diferentes tecnologias,

como dados de sensoriamento remoto, GNSS (*Global Navigation Satellite System*) e GIS (*Geographic Information System*) é imprescindível.

Os sistemas aquáticos possuem comportamento extremamente dinâmico, ou seja, alterações significativas em seu estado podem acontecer em um intervalo de tempo curto (horas). Para identificar tais alterações e viabilizar tomadas de decisões baseadas em dados é necessária a realização de coletas com frequência mais próxima possível do tempo real, inviabilizando outras formas de logística para levantamento dos dados, que não seja um mecanismo de coleta automática e periódica.

Mecanismos que permitam a coleta automática e periódica de características ambientais, como o SIMA, são valiosos recursos que podem ser utilizados para monitoramento constante, pois coletam um número significativo de atributos com uma frequência que permite múltiplos registros diários. No entanto, dois aspectos devem ser considerados com essa abordagem: a quantidade crescente de dados produzidos e a possibilidade de existência de falhas, originadas pelo processo de coleta automatizado – mediante utilização de sensores embarcados nas plataformas – fatores que podem influenciar a qualidade e integridade dos dados produzidos.

A possibilidade de ocorrência de falhas na realização das coletas dos dados é um fator de fundamental importância a ser considerado, pois influencia diretamente o processo de análise dos dados. Tal fato justifica a necessidade da caracterização prévia do conjunto de dados, para que o analista tenha conhecimento sobre o comportamento geral do conjunto e seja capaz de decidir sobre estratégias para extrair informações de qualidade. No âmbito deste trabalho, as principais características de interesse no conjunto de dados podem ser sintetizadas entre: a disponibilidade de dados relacionadas ao intervalo de tempo de existência de determinados atributos em localizações específicas; funcionamento dos sensores no que diz respeito ao sucesso na coleta e armazenamento das variáveis e, principalmente, evidências de falhas que podem apresentar padrões sistemáticos de ocorrências, os quais interferem diretamente na qualidade dos dados utilizados.

Embora a disponibilidade de grande quantidade de dados e variáveis relacionados ao componente temporal seja de suma importância para a análise, a interpretação desses dados pode requerer processamentos mais robustos, principalmente no que diz respeito à manipulação de conjuntos de dados temporais multivariados. A constante aquisição de dados faz com que o volume coletado permaneça em crescimento frequente. Para que o vasto conjunto possa ser analisado e interpretado, é necessária a utilização de recursos computacionais capazes de processar e sintetizar o conjunto.

Um dos recursos que podem potencializar a análise é a utilização de técnicas de Visualização de Informação, que buscam representar visualmente os dados para facilitar a interpretação. Estas técnicas podem ainda serem integradas com outros recursos que se beneficiem da percepção do analista durante a exploração dos dados e garantam a

interação durante a exploração dos dados, por meio de técnicas de *Visual Analytics*, que abrange a utilização de representações visuais em conjunto com recursos de interatividade. Keim et al. (2006) salientam a importância da colaboração entre homem e máquina no processo analítico. Também neste contexto, Ward, Grinstein e Keim (2010) ressaltam os benefícios que podem ser conseguidos utilizando recursos que permitam a interação do usuário com os dados.

A representação gráfica das informações permite que o conjunto de dados seja analisado de maneira mais rápida e intuitiva, favorece a observação de padrões e comportamentos e se beneficia do nível de detalhamento que pode ser empregado na representação visual, em razão da capacidade de percepção e cognição associadas ao sistema visual humano. Além disso, o resultado do processo de análise é diretamente dependente da qualidade dos dados coletados e armazenados. A padronização dos dados e existência de dados com erros ou que não foram coletados são fatores que podem ser problemáticos e interferir no resultado final da análise. Tais situações devem ser tratadas de maneira específica para, a princípio, garantir a integridade dos dados. Esta interferência na integridade dos dados pode ser consequência de diversos fatores relacionados à: degradação dos sensores pela exposição à intempéries do ambiente; falhas na comunicação e transmissão dos dados via enlace de satélite ou erro na conversão para grandezas físicas dos sinais coletados.

1.2 Hipótese

A exploração visual e interativa potencializa a análise dos dados coletados pelo Sistema Integrado de Monitoramento Ambiental e facilita a observação de características do conjunto de dados, tais como: disponibilidade de dados; funcionamento dos sensores; evidências de falhas sistemáticas, as quais influenciam diretamente a qualidade da análise.

1.3 Objetivos

Este trabalho é multidisciplinar e trata da integração entre tópicos das diferentes áreas de Geociências e Ciência da Computação. Neste sentido, os objetivos apresentados são relacionados com aspectos referentes à exploração e análise visual e interativa de dados presentes no campo de *Visual Analytics*, bem como às características ambientais que podem ser ressaltadas pelos dados coletados por meio de sensores das plataformas que integram o projeto SIMA.

1.3.1 Objetivo Geral

O objetivo geral é identificar técnicas de análise exploratória interativa visual de dados mais adequadas para a caracterização dos conjuntos de dados obtidos por múltiplos

sensores integrados, direcionados para o monitoramento ambiental.

Neste trabalho, são utilizados os dados coletados em plataformas do projeto SIMA, no qual registram conjuntos temporais de atributos do ar e da água coletados em reservatórios hidrelétricos. A análise dos fenômenos associados aos significado dos dados não corresponde a um objetivo deste trabalho, pois o foco está na caracterização inicial do conjunto de dados – caracterização prévia relacionada à disponibilidade de dados, funcionamento dos sensores e evidências de falhas sistemáticas, assim como descrito na seção anterior – para viabilizar e facilitar as próximas iterações do especialista com o conjunto de dados abordado.

1.3.2 Objetivos Específicos

São objetivos específicos deste trabalho:

- Selecionar técnicas de Visualização de Informação que possam contribuir para a exploração e análise de conjuntos de dados temporais multivariados, para facilitar a caracterização dos dados, no que diz respeito à identificação de possíveis relações entre o comportamento dos dados coletados e identificação de falhas existentes no conjunto de dados;
- Identificar recursos de interação que podem ser inseridos no processo de exploração e análise de dados para permitir que o analista possa construir e refinar suas análises;
- Integrar as abordagens de representações visuais selecionadas e o conjunto de dados SIMA em um protótipo de aplicação Web para apoiar a exploração e análise dos dados, utilizando abordagens que aumentem a capacidade de reusabilidade, flexibilidade e interoperabilidade, os quais podem ser concretizadas pela adoção de padrões.

1.4 Organização do Documento

Este documento é composto por outros cinco capítulos específicos. Os Capítulos 2 e 3 compreendem a fundamentação teórica com a revisão literária que sustentam o desenvolvimento deste trabalho, e tratam dos assuntos, respectivamente, relacionados às possibilidades de análise de dados ambientais juntamente com os padrões de modelagem da infraestrutura de dados e; aos conceitos e aspectos relacionados à Visualização de Informação. No Capítulo 4 são apresentadas as definições e características do conjunto de dados a ser analisado, bem como as características do projeto SIMA. Os resultados obtidos são descritos no Capítulo 5 para subsidiar as considerações finais e conclusões apresentadas no Capítulo 6.

CAPÍTULO 2

AQUISIÇÃO E GERENCIAMENTO DE DADOS AMBIENTAIS OBTIDOS POR SENSORES

O monitoramento ambiental requer a realização de coletas constantes e frequentes de diversos atributos para a detecção de comportamentos repentinos ou um longo registro histórico para identificação de processos lentos, fator que pode influenciar diretamente a escolha metodológica para a realização das amostragens dos dados (RITCHIE; ZIMBA; EVERITT, 2003). Neste sentido, diferentes técnicas de coleta de dados podem ser integradas para beneficiar a análise dos dados, tais como: coleta de dados *in situ*; produtos de sensoriamento remoto e coletas por mecanismos automáticos.

Uma grande influência nos aspectos ambientais de uma determinada região é a construção e operação de reservatórios hidrelétricos. Inicialmente, a construção dos reservatórios era motivada pela geração de energia renovável em hidrelétricas ou para a reserva de água. No entanto, com o passar do tempo, tais reservatórios começaram a ser aplicados para múltiplos usos, como irrigação, criação e reprodução de espécies, recreação. As consequências do uso múltiplo dos reservatórios, podem amplificar ainda mais o impacto no clima da região em que se localiza, relacionados a alterações da existência de espécies locais, principalmente de espécies de peixes (AGOSTINHO; JR; BORGHETTI, 1992; ARAYA; AGOSTINHO; BECHARA, 2005), qualidade da água (WIATKOWSKI, 2011), hidrodinâmica (LIN; WANG; HU, 2010), padrões de precipitação (DEGU et al., 2011) e até mesmo quanto a emissão de gases do efeito estufa (FEARNSIDE, 2002). Esta característica transforma os corpos d'água que constituem os reservatórios em importantes

objetos de estudos e se configuram como fontes ricas de informações que podem integrar diferentes conjuntos de dados a serem aplicadas em diversas finalidades.

A existência de diferentes bases de dados, geradas por diferentes processos, armazenadas em diferentes fontes, sugere a necessidade de integração destas instâncias de forma que possam enriquecer o processo de análise por meio da integração destes recursos. Para viabilizar a integração de diferentes fontes, é necessário que exista um método de comunicação que permita o compartilhamento de recursos e integração de resultados. Neste sentido, os incentivos para adoção de padrões é um esforço que se faz cada vez mais presente para garantir a configuração das infraestruturas de dados que possam estar em conformidade com outros cenários de utilização.

A especificação e desenvolvimento de padrões ocorre em vários escopos de alcance - locais, regionais, nacionais e internacionais, pois é uma importante ferramenta que permite a organização dos conceitos e produtos técnicos, como dados digitais, e que pode agregar valor sobre o sistema econômico, social e legal (KNOOP, 2000).

Antes dos atuais padrões internacionais, diversos padrões nacionais foram produzidos e especificados visando atender a necessidade de seu próprio escopo, contendo uma definição vasta e abrangente de componentes e funções. Porém, a principal limitação desses padrões residia no seu alto grau de especificidade, ou seja, cada um desses padrões definiam sua própria estrutura de modelagem e comunicação para um cenário específico, na qual a qualidade dos dados estava presente de maneira discreta e secundária (DEVILLERS; JEANSOULIN et al., 2006).

Neste contexto, este capítulo descreve algumas possibilidades para a realização da coleta de variáveis ambientais, sobretudo relacionados aos sistemas aquáticos, bem como algumas abordagens possíveis para a exploração, análise e interpretação dos dados coletados. Neste escopo, também são abordadas especificações reconhecidas como padrão para gerenciamento e acesso a conjunto de dados obtidos por sensores.

2.1 Abordagens para coleta de dados ambientais

Existem diversas metodologias que podem ser aplicadas para a extração e coleta de dados do meio ambiente. Especificamente sobre os ambientes aquáticos, Tundisi (1999) afirma que para a compreensão do funcionamento e estado da vizinhança de um sistema aquático, como os reservatórios hidrelétricos, é necessário observar diferentes variáveis climáticas (temperatura e ar), hidrológicas (fluxo de escoamento) e limnológicas (turbidez, pH, clorofila). Tais variáveis podem ser registradas por meio de diversas metodologias, como medições *in situ*, produtos de sensoriamento remoto e coletas automáticas por estações de sensores especializados.

Dependendo do objetivo da análise e consequente necessidade da registro de determinado parâmetro, um determinado método de coleta será mais propício para obtenção dos resultados. A escala temporal referente à possibilidade de repetições frequentes de amostragens, a escala espacial das regiões que devem ser realizada a cobertura, e o tipo de atributo desejado são fatores primordiais que devem ser levados em consideração no planejamento da coleta dos dados.

A integração das metodologias de coleta pode ser motivada pela necessidade de validação dos dados das fontes utilizadas (DEKKER; VOS; PETERS, 2001), complementação do conjunto capturado (WANG et al., 2012) ou criação de novos modelos de dados (MORADKHANI, 2008; GLENN et al., 2007). Existe também a possibilidade de integração com outras tecnologias para complementar e amplificar as capacidades de análise, como GPS e GIS (RITCHIE; ZIMBA; EVERITT, 2003; HINTON, 1996), possibilitando, respectivamente, a identificação da localização geográfica do ponto de amostra e posterior exibição de informações integradas.

Uma das abordagens para permitir o levantamento de características dos cenários locais é a coleta de dados por levantamento em campo. Os levantamentos em campo possuem, geralmente, um alto custo relacionado referente ao fator financeiro necessário para realização, incluindo gastos com transporte, equipamentos e pessoal qualificado, além do tempo despendido para a execução (JENSEN et al., 2009). Em relação à extração de parâmetros da água, técnicas convencionais de medidas *in situ* dependem da amostragem pontual realizada em vários locais, para serem posteriormente analisadas e extrapoladas para atender a grandes áreas, dificultando o monitoramento efetivo e tornando esse processo oneroso para as instituições de pesquisa e órgãos responsáveis pelo monitoramento da qualidade da água (HARTMANN et al., 1990).

O desenvolvimento de sensores e tecnologias empregadas na construção de sensores orbitais tem permitido um grande avanço na capacidade de investigação e execução de tarefas por meio de soluções de sensoriamento remoto. As imagens multiespectrais geradas por câmeras alocadas em satélites fornecem subsídios importantes para permitir o estudo e interpretação de alvos de interesse. A utilização de tecnologias atuais e que estão em constante desenvolvimento, como aquelas aplicadas no contexto de sensoriamento remoto podem oferecer recursos relevantes para apoiar o processo de extração de informações e análise (PELLEC-DAIRON, 2012).

A principal vantagem na utilização dos produtos de sensoriamento remoto está na captura periódica e programada de informações em uma visão ampla e sinóptica dos alvos de interesse, permitindo a cobertura de grandes áreas. No entanto, a frequência de obtenção de dados de uma determinada região pode não ser suficiente para atender a necessidade de certas análises que necessitam de resolução temporal mais refinada. Por outro lado, a obtenção de dados mediante levantamentos em campo tem um alto custo

associado, o qual é diretamente proporcional ao nível de resolução temporal (frequência de execução das coletas) e espacial (quantidade de pontos coletados em uma área).

A abordagem para coleta de dados utilizando plataformas autônomas de coleta consideradas como estações ambientais automáticas ou Plataformas de Coletas de Dados (PCD), foi impulsionada pelo desenvolvimento de dispositivos sensores, comunicação e infraestrutura de controle. Uma PCD pode ser definida como uma estação que automaticamente coleta dados do meio ambiente e os transmite para estações terrestres por telemetria – usando os satélites (ALONSO, 1982). Estas plataformas de coleta podem ser caracterizadas com a associação das qualidades de outras metodologias de coleta de dados: a automatização do processo de obtenção de dados por sensoriamento remoto (ganho de resolução espacial) e capacidade de se obter resolução temporal de levantamentos em campo conforme necessidade ou disponibilidade.

A utilização de mecanismos que permitam a aquisição automática e periódica de conjuntos de atributos, como as PCD, viabilizam uma alta resolução temporal, ou seja, cada atributo coletado integra o registro histórico de conjuntos de dados temporais. As séries temporais constituem uma importante configuração de registro de dados, as quais os representam os valores e variações 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.

Existem 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 (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 Publicação e interoperabilidade de conjuntos de dados e serviços

As principais organizações que definem padrões internacionais relacionados a dados geoespaciais são *International Organization for Standardization* (ISO), por meio do comitê técnico 211 (ISO/TC211) e *Open Geospatial Consortium* (OGC) (GIANNECCHINI et al., 2006). Ambos os padrões, OGC e ISOTC211, são especificações internacionais, porém os padrões ISO são desenvolvidos e aprovados por membros especialistas que compõem o interesse de diferentes nações, enquanto o OGC é visto como um padrão

industrial desenvolvido por comitês de especialistas compostos pelas partes interessadas na padronização (DI, 2008). Uma vez que as experiências e recursos dos membros integrantes do OGC e ISO são diferentes, a ênfase e características das especificações também são diferentes, pois os padrões OGC são focados tanto em especificações abstratas como em implementações técnicas de modelagens e serviços, já as definições ISO são principalmente focadas em especificações de alto nível a partir de uma perspectiva institucional (PENG; TSOU, 2003).

Os comitês da ISO/TC211 e OGC foram estabelecidos quase simultaneamente na década de 90, com o mesmo objetivo de desenvolver padronizações internacionais para manipulação de informações geográficas e foram se desenvolvendo por meio de diferentes experiências. Em razão da intersecção dos objetivos de ambas as iniciativas, em 1998 foi firmado um acordo formal de cooperação para permitir que o comitê ISO/TC211 pudesse adotar as especificações OGC e publicá-las compartilhadamente como padrões ISO, bem como que o OGC tivesse a permissão de implementar e publicar padrões ISO perante as especificações OGC (KRESSE; FADAIE, 2004). De maneira geral, ISO possui objetivos mais amplos com um nível de abstração maior que as especificações OGC, fazendo com que ambos os esforços sejam complementares e necessários (PENG; TSOU, 2003).

As diretrizes que compreendem o padrão *Sensor Web Enablement* (SWE), elaborado pelo OGC, permitem que os desenvolvedores possam tornar todos os tipos de sensores, transdutores e repositórios de dados do sensor capazes de serem descobertos, acessados e utilizados através da Web. O conceito de sensor se expande a todos os processos capazes de gerar dados que podem ser descritos, definidos sensores virtuais, como resultados de processamento, modelos e algoritmos.

O avanço de tecnologias aplicadas à construção de sensores, computadores e dispositivos de comunicação em rede estão em progresso juntamente com a crescente demanda de conectar e integrar diferentes sistemas. Diferentes fontes de dados são consumidas por diversos cenários, como: infraestrutura de segurança; monitoramento ambiental; meteorologia; avaliação de situações de risco, entre outras.

Em muitos casos, políticas locais para garantir acesso livre às informações públicas motivam a adoção de estratégias compatíveis com padrões internacionais. Moumen et al. (2014) apresentam uma plataforma interoperável como solução de baixo custo, compatível com padrões internacionais, que habilitam o fornecimento e compartilhamento de dados sobre o nível e qualidade de águas subterrâneas de Morocco, localizada na região norte da África.

Moumen et al. (2014) utilizam soluções do servidor 52°North, que implementam recursos do SWE, integrado com o geoportal *deegree* e geocatálogos para construir uma plataforma compatível com serviços descritos pelo OGC, tais como: *Sensor Observation Service* (SOS); *Web Feature Services* (WFS); *Web Map Service* (WMS) e *Catalogue Service*

for the Web (CSW). Os autores concluem que o compartilhamento dos dados dos recursos de água subterrânea de uma região particular facilita a tomada de decisão de administradores e traz benefícios importantes para a população dessa região.

Projetos governamentais que realizam coleta de dados são fontes importantes de informação e geralmente possuem o interesse comum em publicar estes dados, tornando-os encontráveis, conhecidos e reutilizáveis. O projeto de pesquisa e desenvolvimento iniciado por *Nuclear Regulation Authority* (NRA) e *Japan Atomic Energy Agency* (JAEA) sobre responsabilidade de execução do *Advanced Industrial Science and Technology* (AIST) é um exemplo prático de interesse neste sentido (KOJIMA et al., 2014). Tal projeto visa o desenvolvimento e implementação de um framework ¹ baseado em padrões de Linked Open Data (LOD) e serviços OGC para disponibilizar os dados de níveis de radiação coletados em Fukushima. Este projeto, lida com grande quantidade de dados coletados em tempo real, previamente armazenados em arquivos CSV, contando com mais de 17 milhões de registros já existentes. Ao final, os autores comentam sobre algumas questões e desafios que merecem atenção, como: desempenho da realização das consultas; grande número de registros no banco de dados e sobre a duplicidade da base de dados (uma para oferecer os dados de acordo com os padrões OGC-SOS, outra intermediária para ser processada e importar os dados originais em formato CSV).

Um cenário que exemplifica o esforço envidado para permitir o compartilhamento de recursos e integração de diferentes fontes de dados digitais é o avanço relacionado à expansão da quantidade de dados disponibilizados sobre o formato *Linked Open Data* (LOD). Qualquer fonte de informação publicada na internet pode ser disponibilizada de acordo com as especificações LOD para fazer parte da rede interconectada de dados, nomeada como *Web of Data* ou *Web of Things*. A *Web of Data* é formada por um agrupamento de nodes conectados que representam conjuntos de dados estruturados disponibilizados seguindo as boas práticas de definição de LOD, formando um novo espaço global de dados interconectados (BIZER; HEATH; BERNERS-LEE, 2009).

O primeiro passo em direção ao conceito de LOD surgiu com a definição da sintaxe *Resource Description Framework* (RDF) por Berners-Lee e Connolly (1998). Alguns anos depois, Masinter, Berners-Lee e Fielding (2005) propõem a definição de *Uniform Resource Identifier* (URI), que promove uma forma mais genérica que *Uniform Resource Locators* (URL) para identificação de qualquer entidade que existe na Web. Com isso, Berners-Lee (2006) definem posteriormente algumas regras para publicação de dados na Web, de forma que todos os dados publicados possam fazer parte de um único espaço global de dados, conceituando o LOD.

Com a definição do conceito de LOD (BERNERS-LEE, 2006), incentivando a

¹ framework pode ser definido como um arcabouço formado por definições abstratas que descreve uma funcionalidade de propósito geral

Figura 1 – Grafo LOD em 2007 com 12 conjuntos de dados conectados (SCHMACHTENBERG et al., 2014)

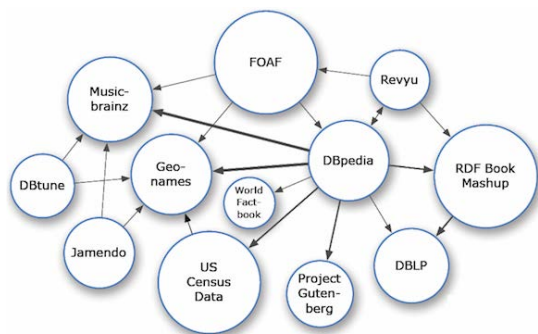
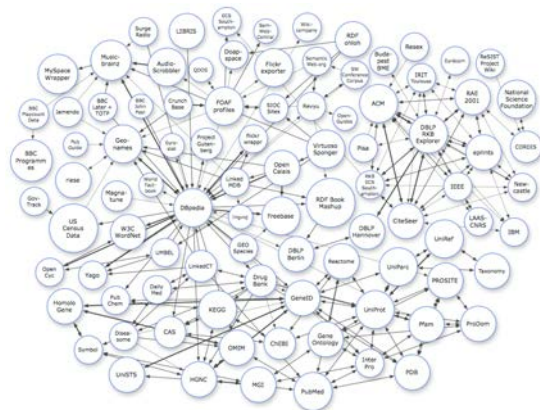


Figura 2 – Grafo LOD em 2009 com 93 conjuntos de dados conectados (SCHMACHTENBERG et al., 2014)



integração e interoperabilidade entre diferentes conjuntos de dados, a quantidade de conjunto de dados que foram definidas de acordo com esse conceito passou a ser crescente. Schmachtenberg et al. (2014) acompanham a popularização do conceito e oferecem dados estatísticos sobre a evolução da adoção e incremento desse cenário. Na Figura 1 pode ser encontrado o grafo que representa os conjuntos de dados conectados em 2007, no início da disseminação do conceito LOD, contando com apenas 12 conjuntos conectados. Uma visão intermediária em 2009 pode ser vista na Figura 2, com um avanço significativo contendo 93 conjuntos conectados.

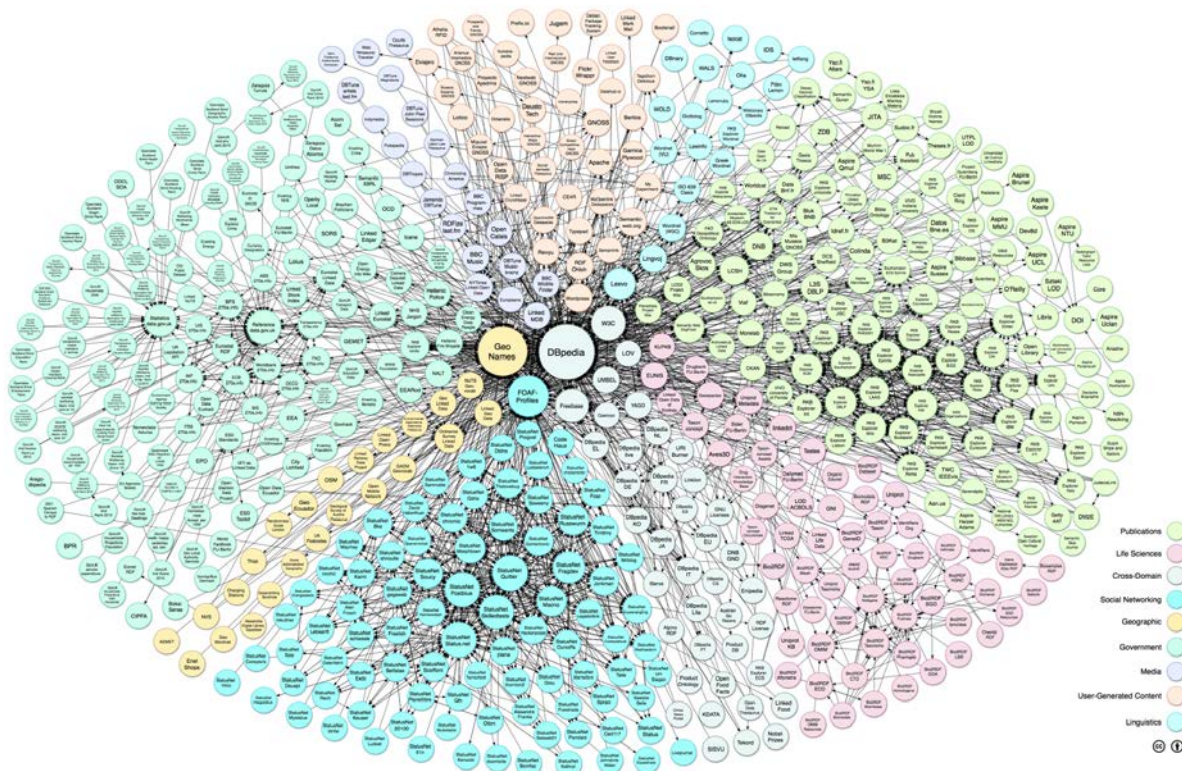
Em 2014, existiu um aumento significativo da popularização do conceito LOD, como pode ser visto na Figura 3, no qual o grafo apresenta uma forma mais densificada, contendo 570 vértices representando os conjuntos de dados conectados.

Na Figura 3 também é possível notar diversas categorias de conjunto de dados, diferenciadas pela cor, como: Publicações; Ciências da vida; Domínios interligados; Redes Sociais; Informações Geográficas; Projetos Governamentais; Arquivos de mídia; Conteúdo gerados por usuário e assuntos linguísticos.

Schmachtenberg, Bizer e Paulheim (2014) comparam o cenário em 2011 e 2014, de conjuntos definidos em conformidade com LOD e ressaltam a existência de quase o dobro de conjuntos disponíveis e conectados, de 295 em 2011 para 570 em 2014. Os autores observam também a tendência para adoção de vocabulários padronizados por um número cada vez maior de conjunto de dados em contraste com a diminuição significativa do uso de vocabulários proprietários de definição de dados - 64,41% em 2011 para 23,08% em 2014.

Os padrões OGC e o conceito de LOD, para construção e expansão da rede que forma a *Web of Things*, possuem o mesmo objetivo em comum relacionado à publicação e

Figura 3 – Grafo LOD em 2014 com 570 conjuntos de dados conectados (SCHMACHTENBERG et al., 2014)



compartilhamento de recursos digitais, de forma que possam ser descobertos, reutilizados e integrados em outro contexto. Os padrões OGC garantem a interoperabilidade de dados e serviços com outros sistemas que também atendem ao padrão OGC. O conceito de *Web of Things* pode ser visto como uma rede em que diferentes recursos digitais são interconectados e relacionados de forma direta e explícita. Neste sentido, a junção entre os dois escopos pode trazer benefícios significativos na direção da interoperabilidade, pois publicando dados sobre sensores e suas observações - como a especificação SOS - de acordo com o conceito de LOD tornaria possível encontrá-los e acessá-los fora do contexto da Infraestrutura de Dados Espaciais (*Spatial Data Infrastructure* - SDI) (KESSLER; JANOWICZ, 2010).

Tal abordagem representa uma possibilidade constantemente relacionado à atribuição de significados aos conteúdos publicados e disponíveis na Web, escopo que define a visão de Web Semântica. Nesse sentido, um novo conceito vêm sendo assunto de discussões e aborda esta possibilidade de fusão entre os ambientes de Sensor Web e Web Semântica, definido como *Semantic Sensor Web* (SSW).

SSW aumenta a capacidade de interoperabilidade entre aplicações avançadas que utilizam dados de fontes de dados heterogênea de sensores. Para isto, tal conceito se baseia na inserção de novas anotações semânticas nas estruturas de linguagem existentes que são utilizadas na especificações SWE (SHETH; HENSON; SAHOO, 2008). Dessa forma, as

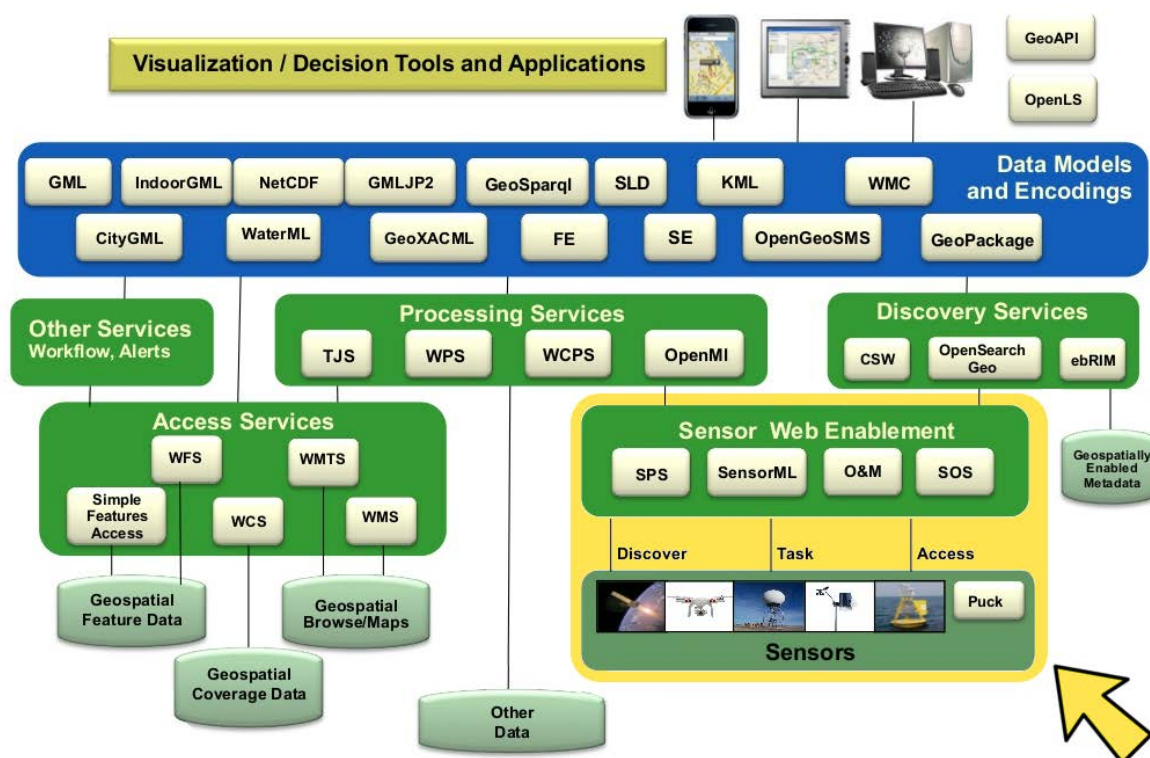
anotações semânticas permitem um maior nível de detalhamento do significado sobre o conteúdo, que pode agir como um mecanismo de conexão entre os metadados definidos no escopo SWE e os padrões de metadados definidos no cenário de Web Semântica.

2.3 Open Geospatial Consortium - OGC

Membros do OGC, composto por usuários e provedores de tecnologias, constroem colaborativamente padrões abertos de interface e respectivos padrões de codificação, que auxiliam desenvolvedores a criar sistemas de informações capazes compartilhar dados geoespaciais e instruções com outros sistemas.

OGC trabalha na produção e manutenção de especificações que descrevem detalhadamente aspectos de engenharia e normas para a implementação de interfaces ou estruturas de codificação que resolvem um problema específico de interoperabilidade de dados geoespaciais (LONGHORN, 2005). A Figura 4 apresenta um diagrama contendo os principais serviços definidos pelo OGC agrupados por similaridade de significados, com destaque para o grupo de especificações SWE utilizados neste trabalho.

Figura 4 – Arquitetura com os principais serviços OGC. Adaptado de Percivall (2014)



Atualmente, o OGC é composto por 509 membros, entre agências governamentais, universidades e companhias industriais, que contabilizam um total de 41 especificações de padrões publicadas. Cada padrão é composto por um subconjunto de documentos que descreve detalhadamente cada componente presente em uma determinada especificação.

A lista atualizada de padrões é disponível publicamente para acesso no portal OGC ², incluindo as versões, número de documentos, situação atual e outras notificações.

Os documentos que compõem os padrões OGC podem ser categorizados em dois conjuntos: Especificações Abstratas (*Abstract Specifications* - AS) e Padrões de Implementação (*Implementation Standards* - IS). Os documentos AS contemplam a especificação de um modelo abstrato que é independente de tecnologia, já os documentos IS são diretamente dependentes de determinadas tecnologias e especificam interfaces, codificações baseadas nas especificações AS. Apenas as especificações IS podem ser submetidas e se tornar padrão ISO.

2.4 OGC *Sensor Web Enablement* - SWE

O OGC define o SWE como uma interface de serviços que permite a interoperabilidade entre os recursos dos sensores, de modo a ocultar as especificidades de cada sensor, permitindo a manipulação e o acesso aos dados a partir de protocolos padronizados e acessíveis na Web (BRÖRING et al., 2011).

Padrões SWE são desenvolvidos e mantidos por membros do OGC que participam do Grupo de Trabalho *Sensor Web Enablement*. O conceito de *Sensor Web* se refere à uma rede de sensores conectados juntamente com seus dados registrados que podem ser descobertos e acessados por meio de padrões e diferentes tipos de *Application Program Interfaces* (API) (REED et al., 2007). A especificação SWE oferece padrões abertos que permite a integração de diversos tipos de sensores, como pode ser visto na Figura 5, na qual diferentes sensores, que coletam dados de diferentes formatos - imagens de satélite, vídeos de câmeras de vigilância, radares, entre outros - mas estão integrados e permitem a fusão dos dados para serem consumidos pelos usuários tomadores de decisão.

Todos os sensores presentes no cenário SWE, ilustrado na Figura 5, possuem a componente espacial informando sua posição geográfica, além de um conjunto de metadados que os definem detalhadamente. Em razão de todos os sensores serem conectados à internet, os dados podem ser acessados via Web e alguns deles estão preparados para serem controlados remotamente, por exemplo, uma câmera de vigilância pode ser manipulada pelo usuário para alterar o campo de visada.

O *framework* SWE pode ser decomposto em recursos relacionados ao modelo de serviços e ao modelo de informações, como esquematizado na Figura 6.

Cada um dos componentes do SWE aborda serviços e funções específicas, compreendendo um subconjunto de especificações. Cada componente do SWE é descrito brevemente a seguir:

² Padrões OGC publicados: <https://portal.opengeospatial.org/files/?artifact_id=55370>

Figura 5 – Esquema ilustrativo de integração de sensores no escopo da especificação SWE (REED et al., 2007)

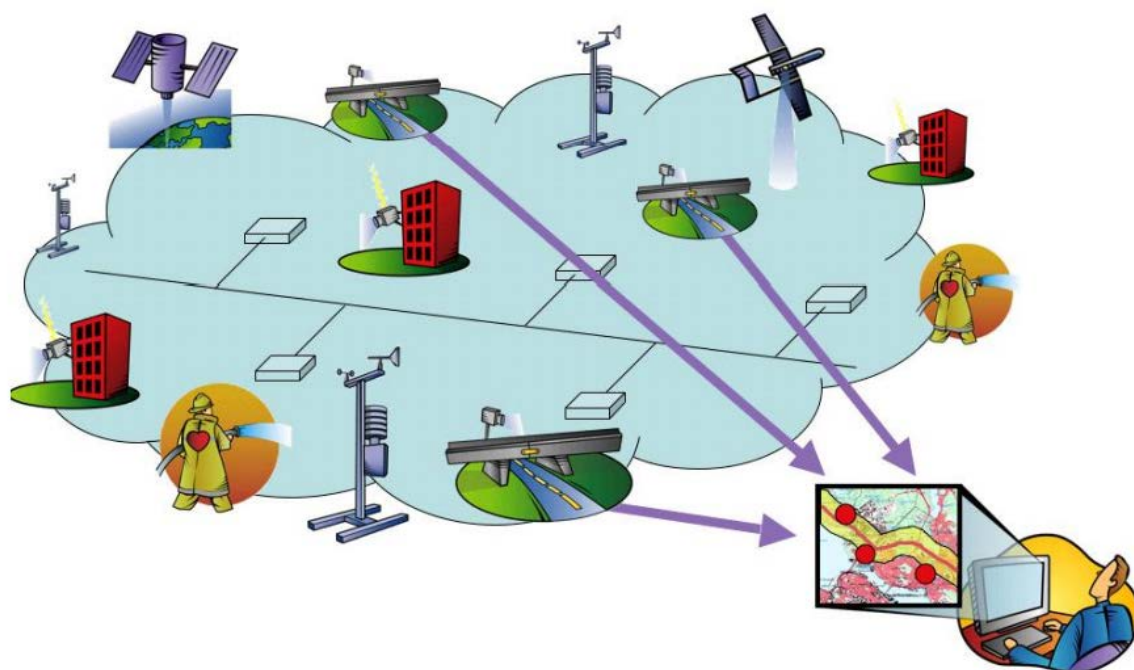
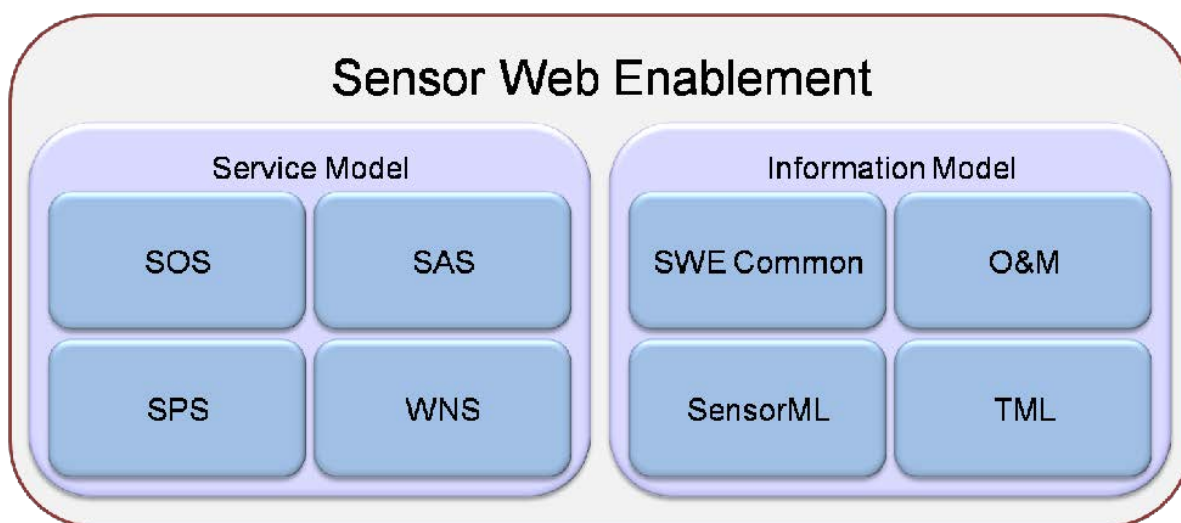


Figura 6 – Componentes do *framework* SWE (BRÖRING et al., 2009)



- *Sensor Observation Service* (SOS): Interface aberta em arquitetura de *Web Service* que permite obter observações e descrições de sensores;
- *Sensor Alert Service* (SAS): Serviço para permitir o disparo de alertas caso seja atingido um determinado comportamento do sensor definido pelo usuário;
- *Sensor Planning Service* (SPS): Uma interface aberta para um *Web Service* através do qual o cliente pode determinar a viabilidade da coleta de dados a partir de um ou mais sensores, bem como enviar solicitações de coleta aos dispositivos;

- *Web Notification Service* (WNS): Servidor que habilita a comunicação assíncrona entre os outros componentes SWE;
- *SWE Common Data Model*: Define os modelos de dados de baixo nível para a troca de dados de um determinado sensor com todos os outros componentes SWE;
- *Observations and Measurements* (O&M): Modelos gerais para codificação XML das observações e medições armazenadas;
- *Sensor Model Language* (SensorML) Modelos de estrutura XML para descrever os processos internos do processamento de sensores e observações;
- *Transducer Markup Language* (TML): Codificação híbrida para integração de dados e metadados de sensores otimizada para transmissão de fluxos de dados.

De acordo com a necessidade do projeto, um ou mais padrões podem ser integrados. A escolha pode ser restrita tanto entre os próprios padrões que definem o escopo SWE como qualquer outra especificação definida pelo OGC. Por se tratarem de padrões, os outros serviços podem ser integrados gradativamente, aumentando a capacidade oferecida pelo sistema.

Neste projeto, os recursos das especificações SOS são utilizados como possibilidade para disponibilização e compartilhamento dos dados dos sensores utilizados.

2.4.1 *Sensor Observation Service* - SOS

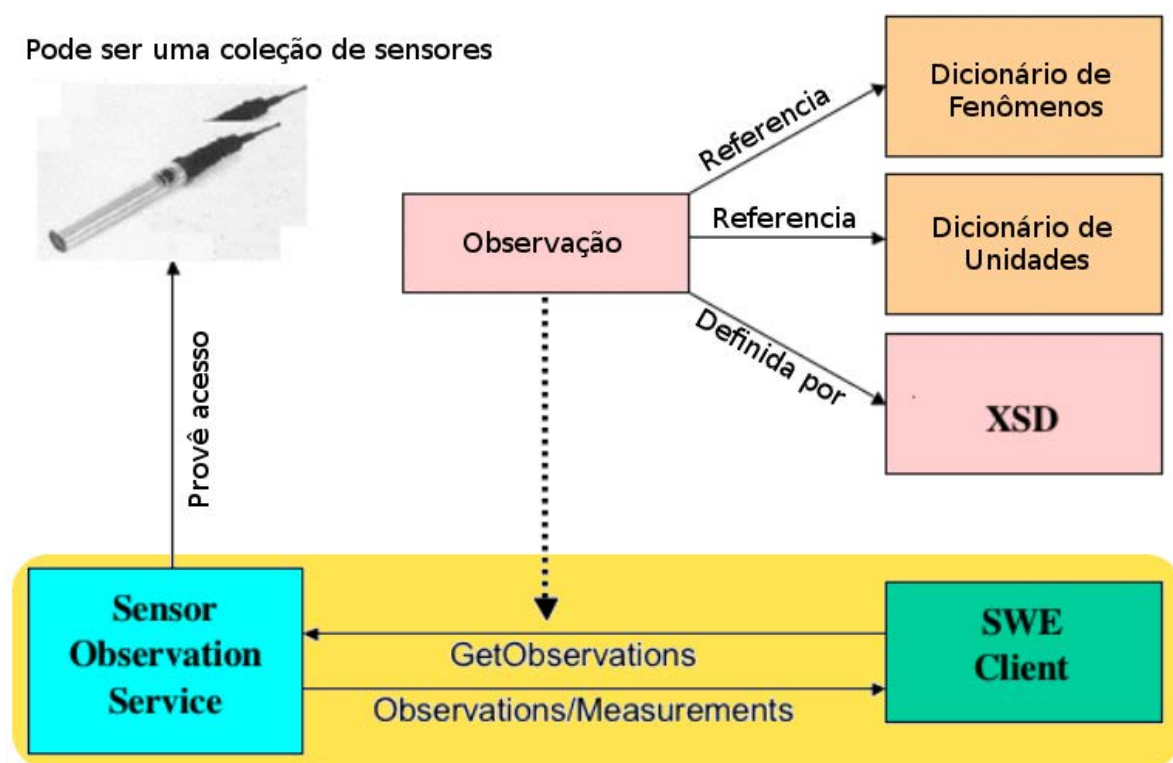
De modo geral, o objetivo do SOS é oferecer acesso às observações de sensores de maneira padronizada, consistente e suficientemente genérica para atender à todos os tipos de dados de sensores (REED et al., 2007).

Padrões SOS são aplicáveis em casos nos quais os dados do sensor tem de ser gerido de forma interoperável, ou seja, com capacidades para integração com outros sistemas e serviços. Esta norma define uma interface de *Web Service*³ que permite a obtenção de observações, metadados do sensor, bem como representações de características observadas, como pode ser visto na Figura 7, na qual apresenta alguns conceitos relacionados e envolvidos no processo de requisição de uma observação, como o dicionário de dados que é sempre consultado e reflete as capacidades consumidas pelos clientes. Nessa Figura, o escopo de iteração e troca de mensagens com o usuário está destacado.

Cada artefato produzido por um servidor SOS é resposta de uma requisição enviada ao *Web Service*. Tanto a estrutura da requisição como a codificação da resposta estão descritas detalhadamente na especificação do padrão (BRÖRING; STASCH; ECHTERHOFF,

³ *Web Service* pode ser caracterizado como um conjunto de funções (serviços) que são providos na Web em uma localização específica

Figura 7 – Elementos envolvidos no processo de requisição SOS, destacando o escopo de requisições do usuário ao *Web Service*. Adaptado de (REED et al., 2007)

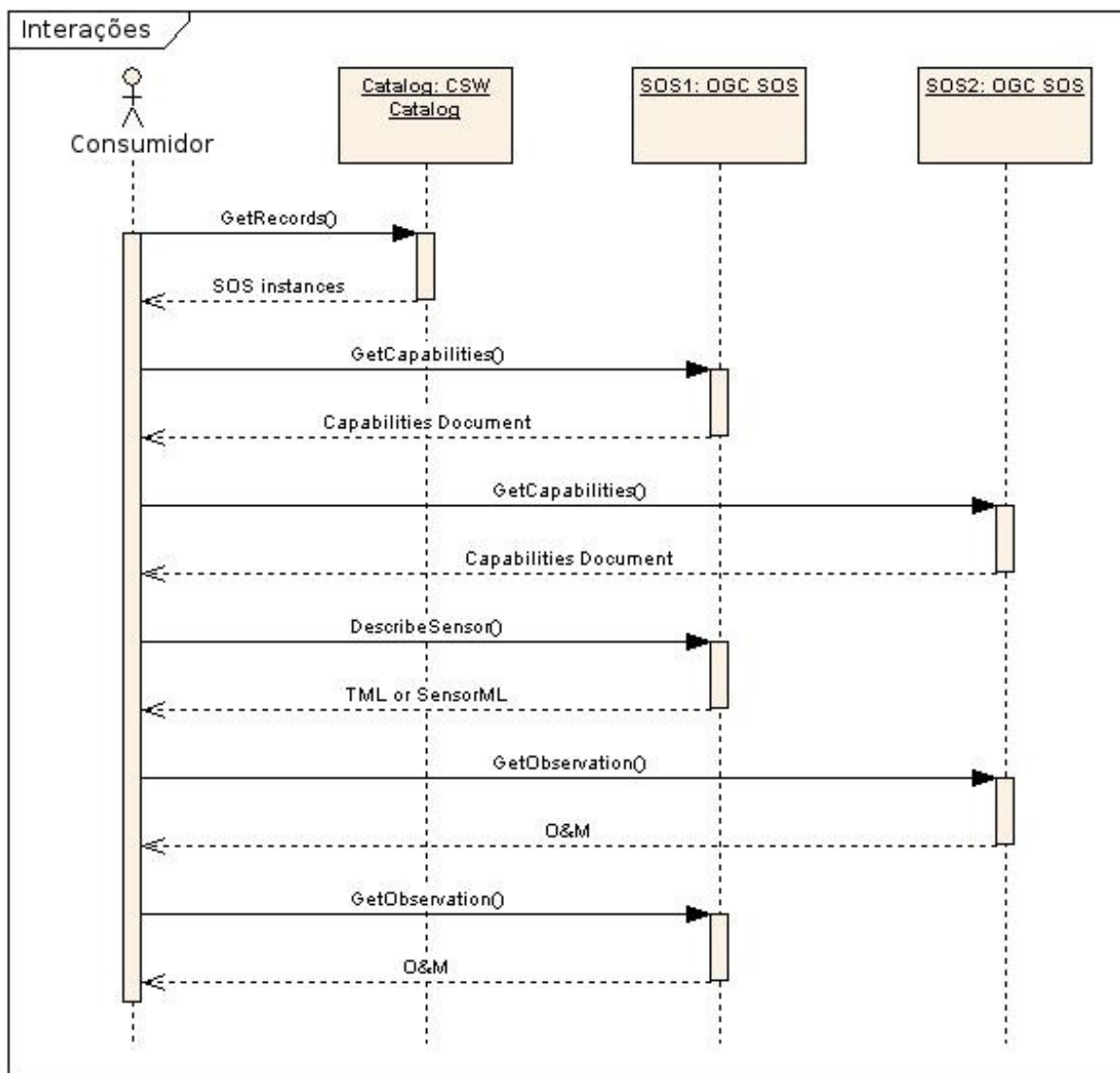


2012). Um caso de uso representando o fluxo geral das mensagens trocadas em uma requisição do usuário pode ser visto no diagrama de sequência apresentado na Figura 8. Nesse diagrama, o usuário realiza a requisição paralela a dois servidores SOS, os quais respondem alternadamente devolvendo a resposta para o usuário que pode utilizar esses dados para realizar outras consultas.

A primeira etapa para utilização dos serviços SOS é a ação necessária para descoberta do serviço. No diagrama da Figura 8, o sistema faz a requisição a um servidor de catálogos para obter as definições e localizações dos serviços SOS. Na sequência, o usuário faz a requisição *GetCapabilities* para duas instâncias de servidores SOS, os quais respondem com o documento geral que define a estrutura e os recursos suportados pelo servidor. Com estes dados, o usuário pode recuperar informações relacionadas aos servidores e às observações, codificadas em linguagem *SensorML* e *O&M*, respectivamente.

As requisições que o usuário pode enviar a um servidor SOS e utilizar determinadas funções é diretamente dependente dos recursos habilitados no *Web Service*. Existe uma hierarquia de dependência entre cada componente que define quais funções estarão disponíveis no servidor SOS, como pode ser visto na Figura 9. O núcleo (*Core*) do modelo SOS é construído baseado em especificações de operações de serviços (*OWS Common*, *SWE Service Model* e *Filter Encoding*) e no formato padrão de resposta sobre observações (*O&M*).

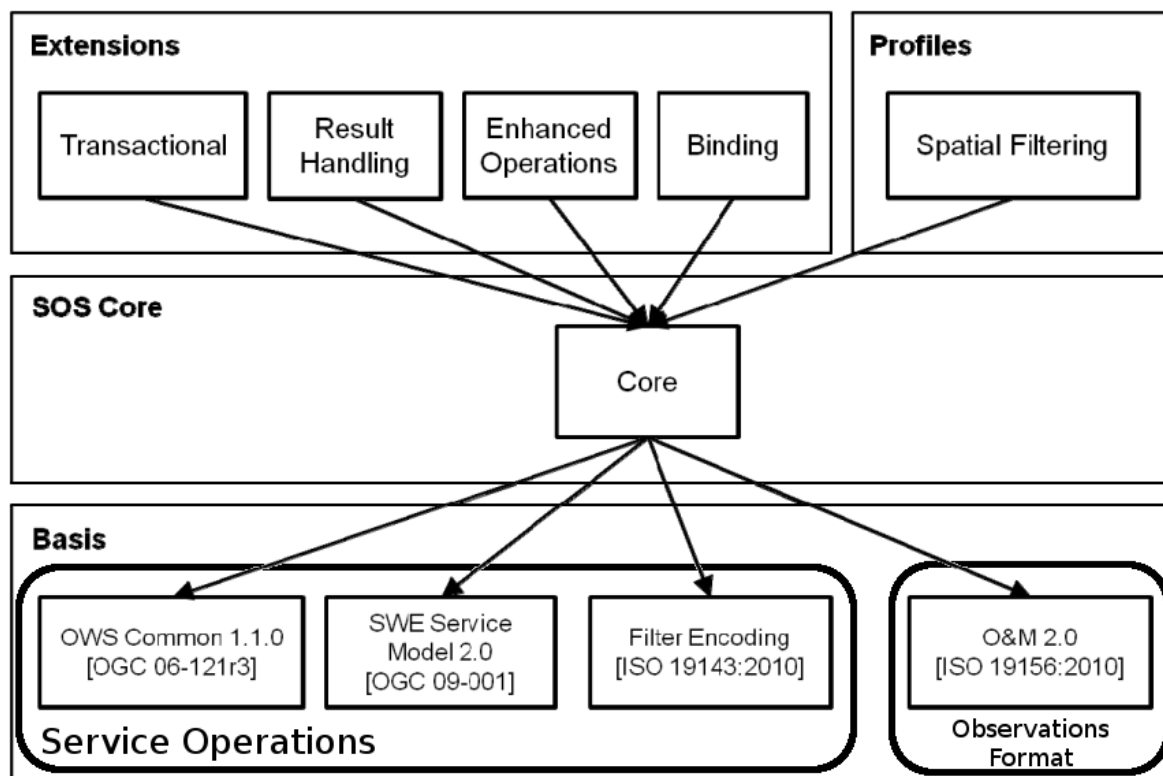
Figura 8 – Diagrama de Sequência representando uma requisição de um usuário ao serviço SOS (NETWORK, 2014)



Qualquer instância de servidor SOS deve implementar no mínimo o componente núcleo da arquitetura, pois todos os outros componentes são diretamente dependentes das funções do núcleo. A partir do núcleo da arquitetura SOS, outras extensões podem ser adicionadas para expandir as capacidades do servidor. Internas a cada um desses componentes, existem funções específicas que devem ser implementadas. A Tabela 1 apresenta as funções presentes em cada um dos componentes com uma breve descrição de sua funcionalidade, desde o componente núcleo obrigatório até as extensões opcionais.

Outros perfis podem ser adicionados ao *Web Service* SOS implementado, tendo em vista o cenário de utilização em que o mesmo será aplicado. Dependendo do domínio da origem dos dados, determinadas funções, não previstas nas especificações do padrão, podem ser implementadas e oferecidas pelo *Web Service*. A decisão de quais perfis e recursos serão atendidos é de responsabilidade dos desenvolvedores dos frameworks que implementam

Figura 9 – Relação hierárquica entre os componentes SOS. Os componentes da especificações não foram traduzidos. Adaptado de (BRÖRING; STASCH; ECHTERHOFF, 2012)



os serviços SOS. Desde que as especificações do núcleo sejam atendidas, o *framework* ou sistema construído pode ser considerado como compatível com o padrão SOS.

2.4.2 Frameworks e servidores compatíveis com SOS

Todas as especificações do OGC, inclusive o padrão SOS, estão definidas em conjuntos de documentos que detalham as capacidades, estruturas e modelos que devem ser considerados para tornar os sistemas compatíveis e em conformidade com tal padrão. No entanto, a OGC se responsabiliza por conceber, definir, atualizar, avaliar e publicar o conjunto de especificações que definem estes modelos, a implementação desses serviços especificados em cada padrão é uma iniciativa pública de outras organizações que optam pela implantação de infraestruturas em conformidade com os padrões definidos.

Em razão da vasta modularização dos padrões definidos pelo OGC, existe uma flexibilidade maior do processo de implementação destes padrões, pois cada desenvolvedor pode escolher quais componentes serão utilizados, viabilizando a construção de um sistema modular, que pode futuramente, atender a outros padrões. O mesmo ocorre com o padrão SWE, no qual cada subcomponente pode ser disponibilizado em uma infraestrutura de dados de forma particionada e articulada.

Recentemente a infraestrutura SWE vem sendo aplicada em diferentes projetos

Tabela 1 – Tabela contendo as funções oferecidas por cada componente SOS

Componente	Função	Descrição
Core	GetCapabilities	permite acesso aos metadados sobre as funções disponíveis pelo servidor SOS
	DescribeSensor	recupera metadados sobre os sensores disponíveis pelo servidor SOS
	GetObservation	permite acesso às observações utilizando filtros espaciais, temporais e temáticos
Enhanced Operations Extension	GetObservationByID	permite acesso às observações utilizando o identificador da mesma
	GetFeatureOfInterest	provê acesso às “Features of Interest” nas quais o servidor oferece observações
Transactional Extension	InsertSensor	permite o registro de novos sensores e respectivos metadados
	DeleteSensor	permite a exclusão de sensores e todas as suas observações
	InsertObservation	permite a inserção de observações e metadados no servidor SOS
Result Handling Extension	InsertResult	permite a inserção direta de observações, sem metadados (template predefinido)
	InsertResultTemplate	permite a inserção de template contendo os metadados e estrutura das observações
	GetResultTemplate	provê acesso ao template utilizado para estruturação do resultado
	GetResult	provê acesso a um resultado de observação, sem informações de metadados

- demonstrando a praticidade e adequabilidade dos padrões SWE - e um dos principais serviços SWE utilizados, percebido por meio de levantamento não exaustivo de projetos e aplicações, é o padrão SOS (BRÖRING et al., 2011).

Existe uma vasta gama de iniciativas que implementam as especificações do OGC em frameworks e servidores que são disponíveis publicamente. Em razão da flexibilidade oferecida pelas especificações no que diz respeito à escolha das tecnologias, as implementações são realizadas a partir de diferentes linguagens de programação e arquiteturas de execução. Dentre os projetos que oferecem ferramentas e aplicações que implementam as especificações SOS, pode-se citar os projetos *open source*: 52°North-SOS; MapServer-SOS e deegree SOS.

A iniciativa informal 52°North foi fundada em 2004 pelo *Institute for Geoinformatics of the University of Muenster* (IGFI) para incentivo de software geoespaciais livres e *open source*. Com a junção e colaboração de outras organizações, em 2006, representantes destas instituições estabeleceram a fundação da companhia *52° North Initiative for Geospatial Open Source Software GmbH*, conferindo força para a realização de pesquisa e desenvolvimento e transformando-a em uma entidade confiável e bem estabelecida na área de geoinformática (52NORTH, 2015b). Diversos softwares geoespaciais foram desenvolvidos colaborativamente pela iniciativa 52°North e disponibilizados sobre a licença *open source*,

incluindo uma suíte de implementações do padrão SWE. O servidor 52°North-SOS (Versão 4.1.0) é uma arquitetura robusta, bem documentada e suportada pela implementação do padrão SOS 2.0, construída em Java e disponibilizada como um *Servlet*⁴. É o único servidor que implementa todas as funções de todos os perfis e extensões da especificação SOS 2.0 (52NORTH, 2015a).

MapServer é uma plataforma *open source* para publicação de dados espaciais e aplicações interativas de mapas na Web que, por ser implementado como uma aplicação CGI (*Common Gateway Interface*), é capaz de ser executado em diferentes plataformas e acessado, via MapScript, em vários ambientes de programação, como PHP, Python e .NET (MAPSERVER, 2015b). O suporte ao padrão SOS está disponível no MapServer desde a versão 4.10.0 (atualmente 6.4.1) e segue a mesma linha de configuração para publicação de dados geoespaciais, utilizando o arquivo mapfile. As funcionalidades do núcleo do serviço SOS são implementadas sobre a especificação SOS (OGC 05-088r1, Version 0.1.1) (MAPSERVER, 2015a).

O *framework* deegree é um software *open source* construído sobre os pilares dos padrões OGC e ISO/TC211, voltado para infraestrutura de dados espaciais que inclui componentes para o gerenciamento de dados geoespaciais para acesso, visualização, descoberta e segurança de dados (DEEGREE, 2015a). Assim como a solução 52°North, o *framework* deegree SOS é implementado em Java e executado no servidor como um *Servlet*. As funções do núcleo da especificação SOS 1.0 são suportadas com o acréscimo da função *GetFeatureOfInterest* pertencente à extensão *Enhanced* (DEEGREE, 2015b).

Em razão da existência de várias iniciativas de implementação dos padrões OGC, alguns estudos comparativos entre os recursos disponíveis em cada servidor podem ser encontrados na literatura. Poorazizi, Liang e Hunter (2012) apresentam uma discussão baseada em avaliação quantitativa do desempenho de três servidores SOS - 52°North-SOS; deegree SOS e MapServer SOS - em três diferentes cenários, relacionados à observação de tempo e espaço necessário para: recuperar informações de observações de sensores; aplicação de filtro espacial por *Bounding Box* e aplicação de filtro temporal.

Um dos resultados dos testes realizados por Poorazizi, Liang e Hunter (2012), referente ao desempenho na recuperação de observações a partir de um conjunto com diferentes quantidade de sensores, pode ser visto na Figura 10 e 11, respectivamente, sobre o tempo e espaço consumido no processo.

Os autores concluem, baseado também nas respostas dos outros experimentos, que a diferença entre o desempenho dos três servidores pode estar relacionada com as modelagens dos dados que cada um utiliza. Este fato justifica o espaço de memória maior demandado pelo servidor 52°North e o aumento exponencial no tempo de execução de

⁴ Programa em Java que é executado como um *Web Server*

Figura 10 – Tempo consumido para recuperação de observações de acordo com a quantidade de sensores (POORAZIZI; LIANG; HUNTER, 2012)

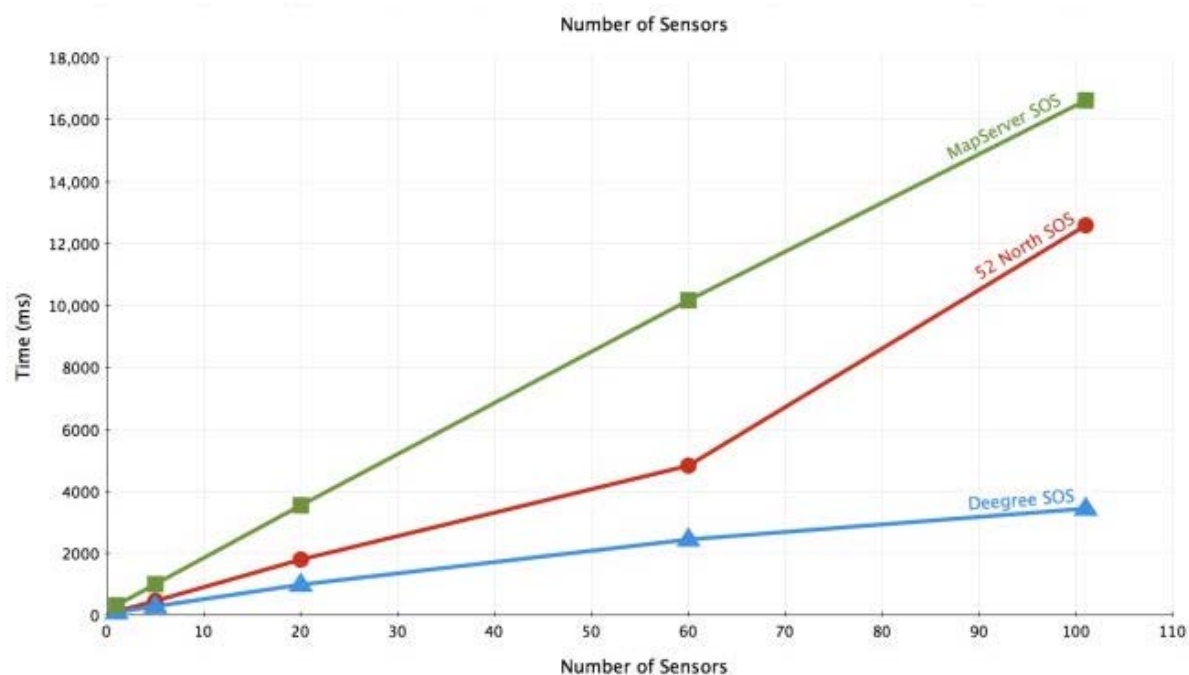
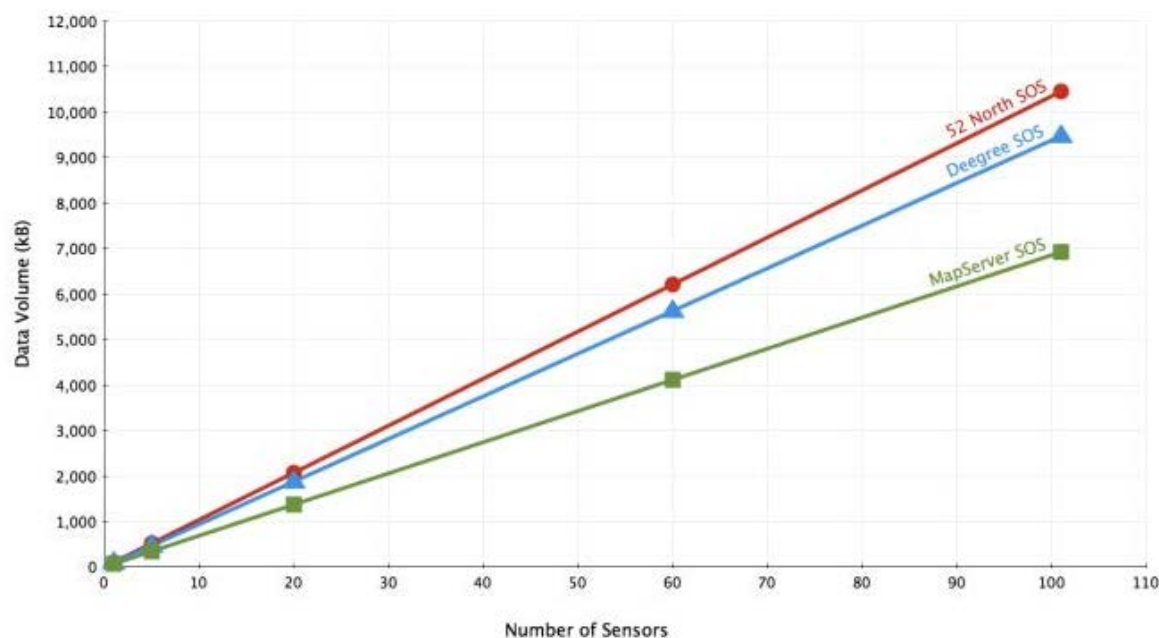


Figura 11 – Espaço de memória demandado para recuperação de observações de acordo com a quantidade de sensores (POORAZIZI; LIANG; HUNTER, 2012)



acordo com a quantidade de sensores, pois dentre os três, é o que possui o modelo de dados mais estruturado e complexo, exigindo a consulta integrada à diferentes entidades do banco de dados (POORAZIZI; LIANG; HUNTER, 2012).

McFerren et al. (2009) realizaram um estudo comparativo entre quatro servidores que implementam o SOS: 52°North-SOS; PySOS; MapServer SOS e deegree SOS. Os autores

avaliam qualitativamente cada uma das implementações em três cenários de utilização de conjuntos de dados específicos: observações de incêndio por imagens do satélite MODIS; dados meteorológicos coletados *in situ* por estações automáticas e parâmetros de qualidade da água.

Ao final, McFerren et al. (2009) concluem que as principais diferenças entre as implementações estão relacionadas com a maneira em que são configuradas, bem como o nível de liberdade para a modelagem do banco de dados. MapServer, deegree e PySOS possuem uma flexibilidade maior para trabalhar com bases de dados já existentes e previamente preenchidas com dados, enquanto 52°North possui uma modelagem de dados mais complexa em razão de utilizar um modelo de banco de dados próprio, o qual os autores indicam para casos em que não exista dados previamente armazenados no banco de dados.

CAPÍTULO 3

EXPLORAÇÃO E ANÁLISE VISUAL DE DADOS

A complexidade envolvida no processo de exploração e análise de dados aumenta com: o volume de dados; a quantidade de parâmetros a serem interpretados; e a integridade e qualidade dos dados. Técnicas de análises e recursos computacionais, como aquelas presentes no processo de *Visual Analytics*, podem oferecer as ferramentas necessárias para viabilizar essa tarefa e torná-la mais natural e intuitiva.

O termo *Visual Analytics* foi apresentado por Thomas e Cook (2005) no cenário em que a representação visual utilizando técnicas de Visualização de Informação não era suficiente para viabilizar a análise de grandes quantidades de dados. Técnicas de interação e manipulação direta 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 dados e elaboração do raciocínio analítico.

Manipulação direta foi um conceito introduzido por Shneiderman (1992 apud AHLBERG; SHNEIDERMAN, 1994), o qual descreve a capacidade de percepção do usuário sobre uma determinada ação de maneira progressiva. Princípios como a possibilidade de obtenção imediata e contínua do resultado de suas intervenções, que devem ser realizadas de maneiras naturais e intuitivas, possibilitam que estas sejam refinadas e direcionadas para outra ação, ou conclusão da análise. De forma geral, a manipulação direta pode ser vista como a capacidade que o sistema de análise tem de exibir ao analista todas as transições das alterações aplicadas, de forma que ele possa acompanhar as modificações de maneira fluída.

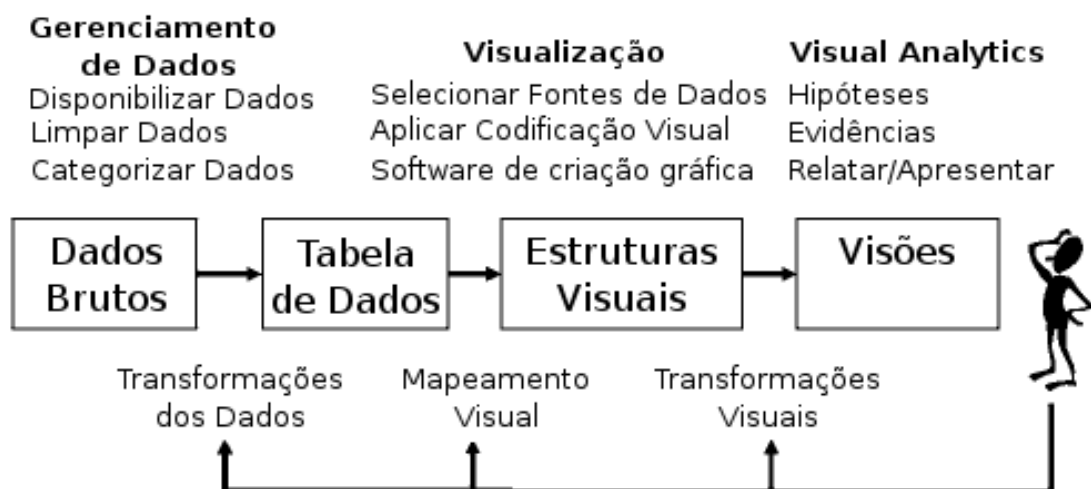
Este capítulo apresenta alguns assuntos relacionados ao cenário de Visualização de Informação e *Visual Analytics*, que são necessários para compreender o processo de desenvolvimento do trabalho.

3.1 *Visual Analytics* e conceitos relacionados

Em 1996, discussões sobre um conjunto de teorias, ferramentas e utilização de recursos computacionais para auxiliar no processo de extração de informações úteis a partir de volumes de dados em constante crescimento dão origem ao conceito de Descoberta de Conhecimento em Bases de Dados (KDD - *Knowledge Discovery in Databases*) (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996). O conceito de KDD trouxe consigo algumas abordagens para complementar o processo de exploração e análise de base de dados, dentre elas a Mineração de Dados e Visualização de Informação.

Card, Mackinlay e Shneiderman (1999) definem Visualização de Informação como a utilização do computador para construção de representações visuais interativas de dados abstratos para amplificar a capacidade cognitiva. Os autores estabelecem um modelo de referência para Visualização de Informação, ilustrado na Figura 12, para descrever as etapas desse processo. Dados brutos são pré-processados, transformados e organizados adequadamente em tabelas. Em seguida, é realizado o mapeamento para estruturas visuais, as quais são manipuladas para a exibição gráfica para o usuário. Esse modelo, bastante difundido, é utilizado por Heer e Agrawala (2007) para mostrar que a atividade colaborativa e interação com o usuário está presente em várias fases do processo de visualização.

Figura 12 – Modelo de referência para Visualização de Informação. Adaptado de Heer e Agrawala (2007)



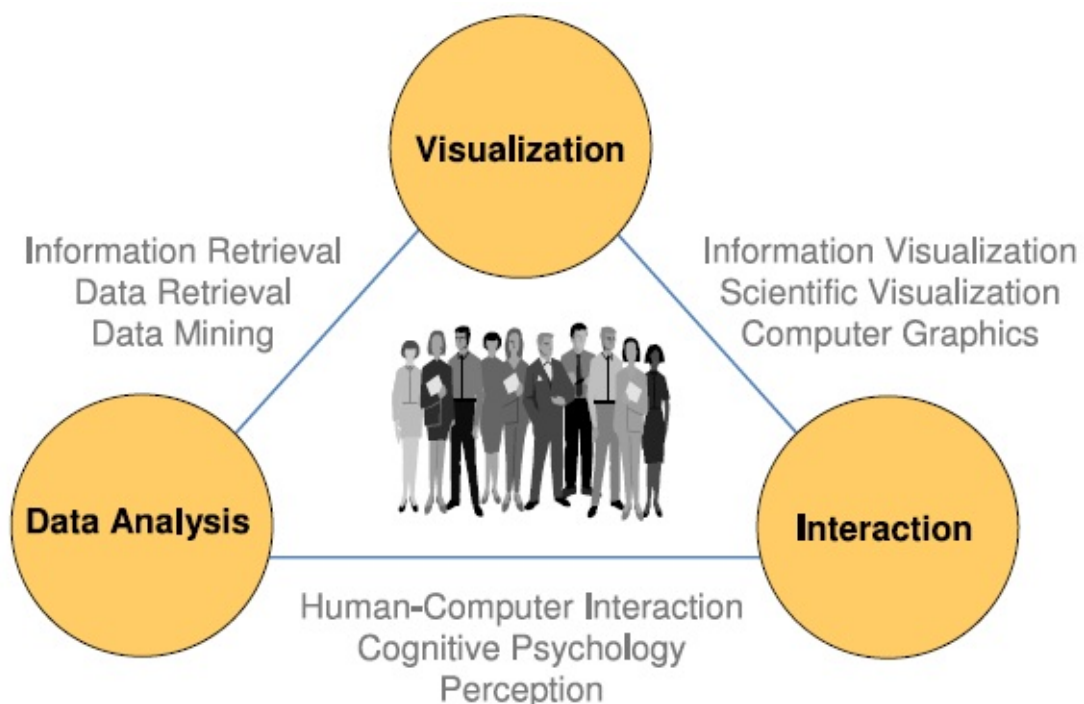
Nesse conceito de Visualização de Informação, o analista detém papel primordial no processo de exploração e análise dos dados a partir da interação e manipulação das representações visuais. De acordo com o *Information Seeking Mantra*, definido por

Shneiderman (1996), a exploração visual dos dados segue um processo composto por três etapas básicas: visão geral do conjunto de dados; aplicação de zoom e filtros; e realização do detalhamento conforme necessário. Tais etapas salientam a intervenção do usuário no processo analítico, configurando um ambiente semiautomático que se difere da metodologia automática aplicada ao escopo de áreas como Mineração de Dados (DM - *Data Mining*).

Tratando da convergência entre Mineração de Dados e Visualização de Informação, Ankerst (2001) define Mineração Visual de Dados (VDM - *Visual Data Mining*) como uma etapa no processo de extração de conhecimento (KDD) que utiliza a visualização como uma forma de comunicação entre computador e usuário para apoiar a identificação de padrões interpretáveis. A utilização de recursos para mineração visual pode ser justificada em razão de que o processo automático de mineração conduzido por processamentos computacionais é difícil e nem sempre produz resultados diretamente interpretáveis.

Para a representação visual de dados são utilizadas técnicas que transformam os dados em elementos visuais significativos. O campo de Visualização de Informação abrange outros assuntos relacionados ao escopo de análise e interpretação de dados, configurando-se como uma vertente multidisciplinar. Dentre as áreas que possuem aspectos relacionados pode-se citar: Visualização de Informação, Visualização Científica, Visualização de Software, Visualização Geográfica, Visualização de Redes, Realidade Virtual, Realidade Aumentada, entre outros. Uma relação entre algumas áreas de pesquisas relacionadas à Visualização de Informação pode ser vista na Figura 13, na qual destaca novamente o papel dos analistas no centro do processo de exploração e análise de dados (KEIM et al., 2010).

Figura 13 – Áreas de pesquisas relacionadas com o campo de Visualização de Informação (KEIM et al., 2010)



O número de sistemas que utilizam técnicas de Visualização de Informação para suporte às tarefas analíticas é cada vez maior. Uma das formas para explorar os dados e amplificar a capacidade analítica dos especialistas é o uso de tecnologias apropriadas para desenvolvimento de recursos que permitam amplificar a capacidade de intervenção no processo de análise, cuja interatividade torna o analista apto a explorar a informação por meio das representações visuais (STASKO; GÖRG; LIU, 2008). A necessidade de integração entre técnicas de Visualização de Informação com recursos amplos de interatividade origina o termo *Visual Analytics* (VA).

Thomas e Cook (2005) definem VA como a ciência do raciocínio analítico facilitado por interfaces gráficas interativas. VA oferece automatização no processo de sintetização dos dados e representação gráfica dos mesmos. No entanto, continua sendo extremamente importante e necessária a intervenção do usuário para composição da análise e validação do resultado.

Considerando a cronologia do surgimento das áreas de pesquisas relacionadas com a utilização de representações visuais, *Visual Analytics* pode ser vista como uma evolução baseada na Visualização de Informação e Visualização Científica (KEIM et al., 2006). Com isto, da mesma forma que a Visualização de Informação está diretamente relacionada com diversas outras áreas, *Visual Analytics* também está relacionada com outros campos referentes à interação, percepção e cognição, como pode ser visto no diagrama ilustrado na Figura 14.

Keim et al. (2006) salientam que a visualização é uma etapa do processo analítico semiautomático, no qual humanos e máquinas cooperam usando suas respectivas capacidades associadas à percepção durante a interpretação (humano) e à capacidade de armazenamento e processamento (máquina), para gerar resultados mais eficazes. Icke (2009) reforça esse conceito definindo VA como o trabalho colaborativo entre homem e máquina no processamento e análise dos dados para permitir a tomada de decisão, identificando três aspectos distintos relevantes neste processo: o sistema, o usuário, e a ação efetiva empregada na colaboração homem-máquina.

Na Figura 15, são representadas todas as etapas envolvidas no processo de VA, definidas por Wijk (2005), e adaptada por Keim et al. (2006) como *Sense-making Loop*. O diagrama representa todas as etapas compreendidas no processo de exploração e análise de dados visuais, desde os dados armazenados até a resposta do usuário à visualização. Por meio desse diagrama fica claro a influência fundamental do humano no processo, e a importância da interação do usuário com o sistema de visualização para um refinamento gradual do processo de exploração e análise.

Heer e Agrawala (2007) e Card, Mackinlay e Shneiderman (1999) discutem as fases presentes no ciclo de construção do conhecimento existente durante o processo de exploração e análise por meio de representações visuais, e apresentam aspectos a serem

Figura 14 – Relações de *Visual Analytics* com outros campos de pesquisas (KEIM et al., 2006)

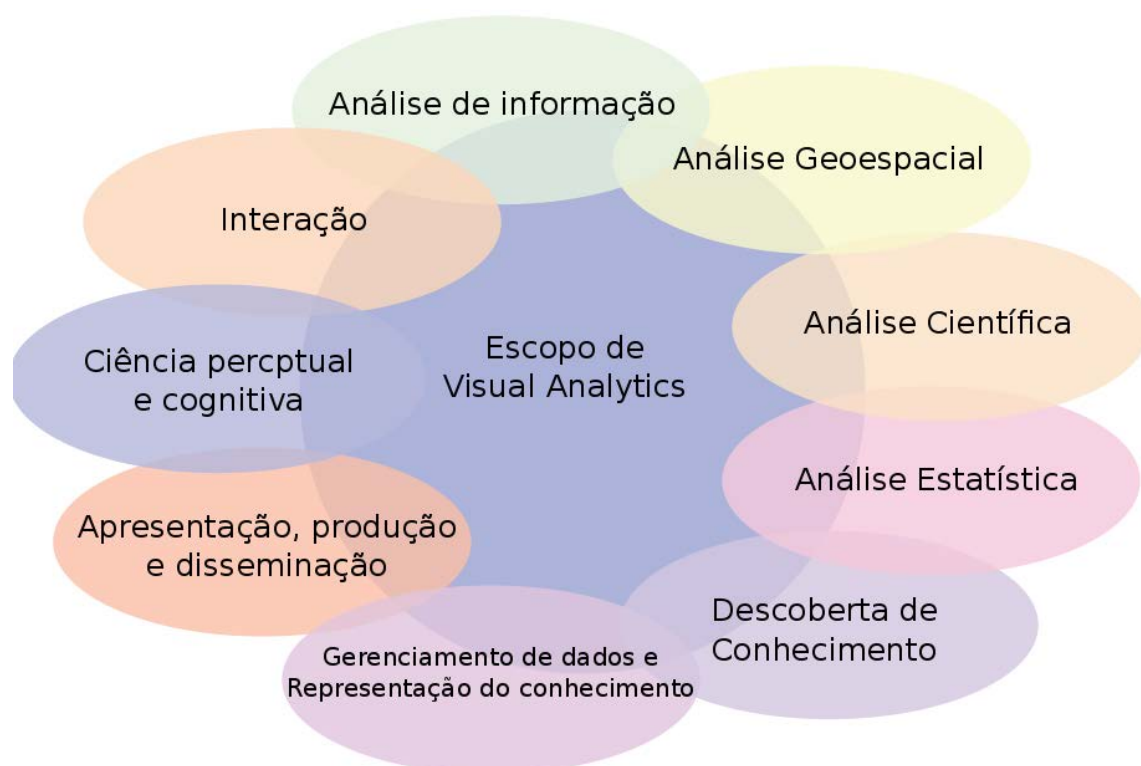
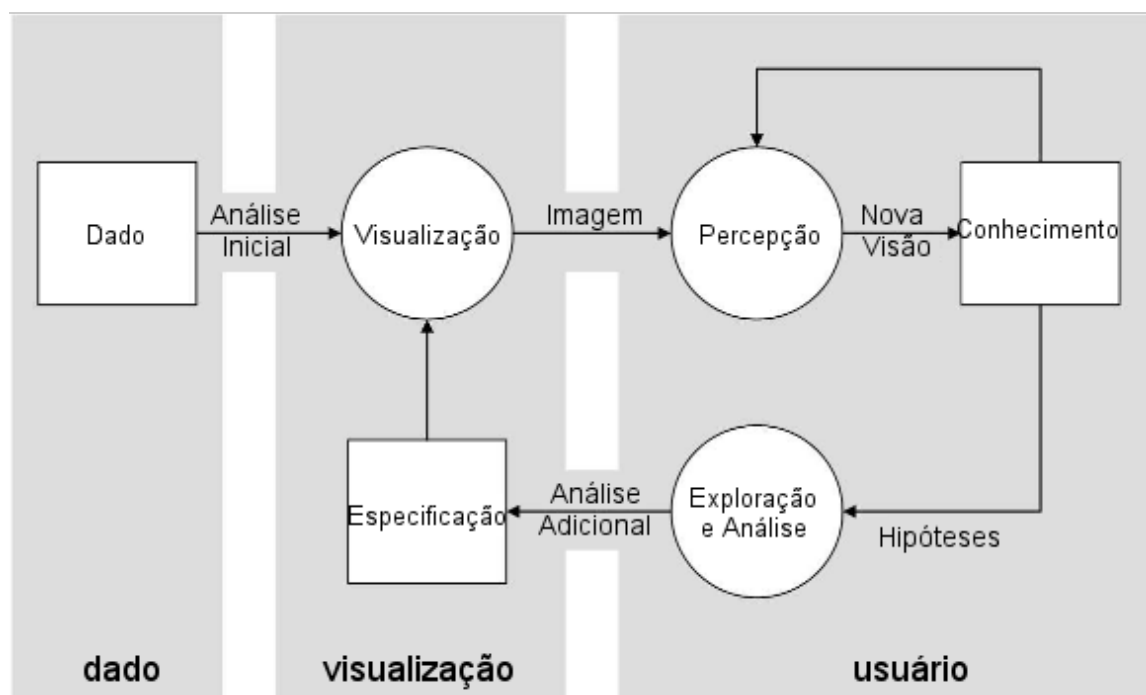


Figura 15 – Diagrama de *Sense-making Loop* ressaltando o ciclo de interatividade. Adaptado de Keim et al. (2006)



considerados para a construção de ambientes para a análise visual. Ward, Grinstein e Keim (2010) ainda abordam de maneira clara e explicativa os aspectos e técnicas envolvidas na visualização de dados, sobretudo com recursos e abordagens interativas. No mesmo escopo,

Thomas e Cook (2005) evidenciam os desafios e metas da aplicação de técnicas de VA, visando incentivar e disseminar sua utilização.

Especificamente no cenário de análise de dados obtidos por sensores, a componente temporal pode oferecer outras oportunidades de análise, como apresentado por 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. Neste sentido, atenção especial deve ser direcionada para a utilização de recursos e técnicas de Visualização de Informação, de acordo com a estrutura do conjunto de dados a ser analisado.

3.2 Representação visual de dados temporais

Na literatura existem diversos modelos para representação visual de dados, cada um com características específicas que direcionam o uso à uma determinada estrutura de dados. A estrutura dos dados é o principal fator, que aliado ao objetivo da análise, ajuda a definir as técnicas de representação visual mais propícias para obtenção do resultado esperado e capaz de amplificar o processo de exploração e análise dos dados.

Existem diversos cenários de aplicação que utilizam conjuntos de dados que possuem componentes temporais e espaciais associados a cada registro, relacionados respectivamente, ao momento em que o dado foi obtido e à localização de obtenção do mesmo. No entanto, dados espaciais e temporais raramente são amostrados com uma escala ótima, resultando na inserção de incertezas sobre a qualidade, associadas ao processo de transformação das amostragens discretas para um possível escopo de sistema contínuo (CRESSIE; WIKLE, 2011). Neste sentido, técnicas de representação visual das informações, considerando originalmente os componentes espaciais e temporais, podem oferecer informações importantes sem a necessidade de realização de interpolações e outros métodos que podem aumentar as incertezas dos dados.

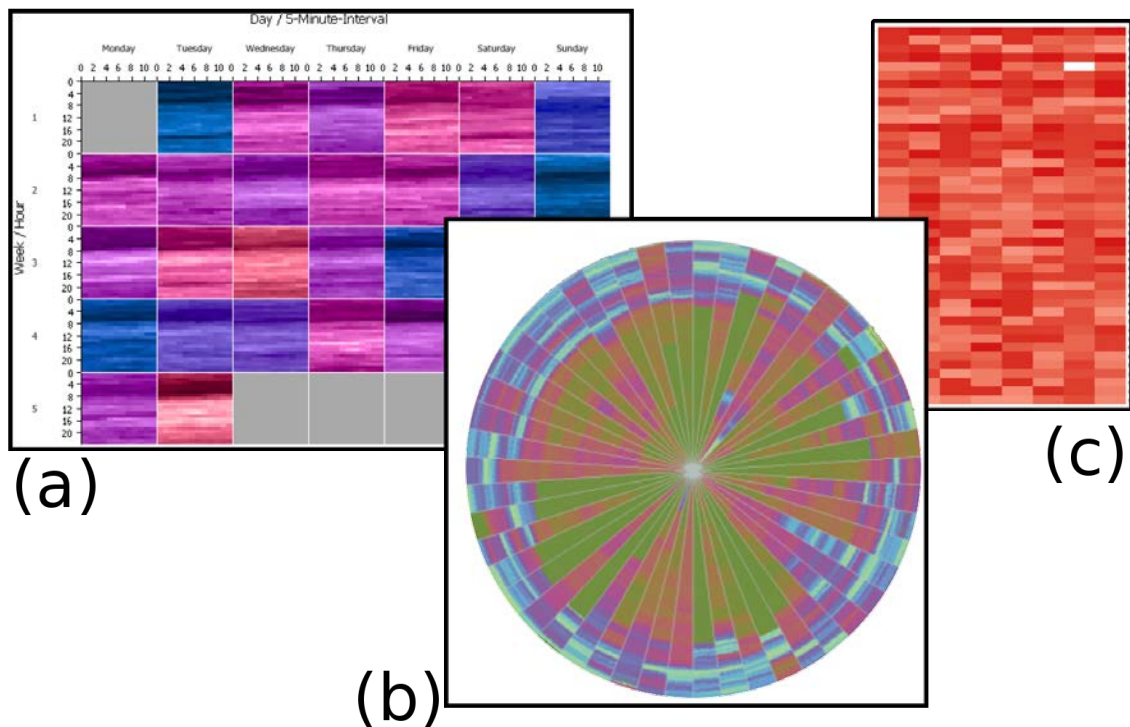
A referência temporal pode trazer benefícios associados ao registro histórico do comportamento dos dados multivariados, que pode ser de fundamental importância para o monitoramento e análise dos dados coletados. Algumas técnicas específicas para a manipulação de dados com referência temporal podem ser utilizadas e outras que não possuem essa característica podem ser adaptadas e aproveitadas. Técnicas de Visualização de Informação ainda podem ser utilizadas em conjunto com outros métodos de análise de dados e abordagens estatísticas, para facilitar e complementar o entendimento das representações visuais.

Uma importante classe de técnicas de Visualização de Informação voltadas para a análise de dados multivariados são as representações visuais baseadas em pixel (*pixel based*), que buscam representar a maior quantidade possível de dados simultaneamente,

mapeando cada valor do conjunto de dados para um único pixel da tela, de forma que esses pixels possam ser distribuídos e rearranjados para ressaltar características do conjunto (KEIM, 2000). O processo de construção desta classe de representação visual pode ser resumido à atribuição de uma cor para cada pixel que representa um determinado valor do conjunto de dados. Com isso, em um monitor com resolução de 1900x1080 pixels seria possível representar no máximo 2.052.000 de registros simultaneamente.

As representações visuais baseadas em pixel representam uma classe de técnicas que podem ser adaptadas e utilizadas de diversas maneiras. A concepção de sua implementação é diretamente dependente do objetivo da análise. Algumas abordagens que utilizam técnicas de Visualização de Informação baseadas em pixel em diferentes conjuntos de dados podem ser vistas na Figura 16.

Figura 16 – Exemplos de técnicas de Visualização de Informação baseadas em pixel: (a) Representação em uma estrutura de calendário (LAMMARSCH et al., 2009); (b) Representação em *Circle Segments* (KEIM, 2000) e (c) Representação por *Heatmap* (BORGIO et al., 2010).



Todas as representações visuais apresentadas na Figura 16 utilizam a técnica baseada em pixel para representar um grande volume de dados. Cada representação tem o objetivo de ressaltar uma determinada característica do conjunto de dados. A representação 16(a) apresenta dados temporais dispostos em uma matriz na qual cada célula indica os registros de ocorrências policiais realizados em um dia (azul para valores baixos e vermelho para valores altos). Em 16(b) são exibidos 20 anos de índices diários de 50 ações da bolsa de valores de *Frankfurt*, utilizando a construção espiral da representação, em que valores altos são apresentados pelas cores mais brilhantes. No Gráfico 16(c) é

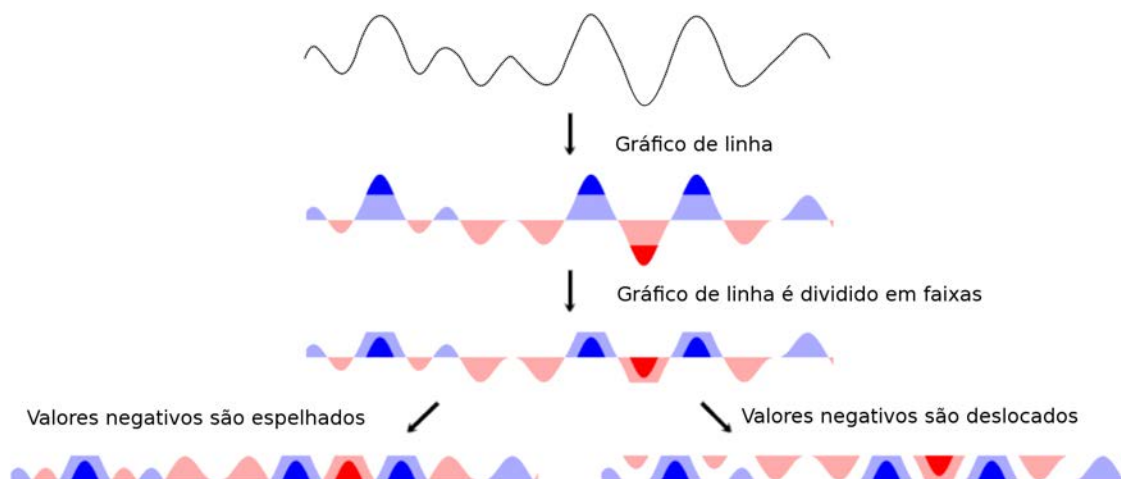
exibido um mapa de valores de estatísticas da bolsa de valores de Londres, na qual a intensidade da cor indica a amplitude do valor.

A capacidade de adaptação de técnicas baseadas em pixel para serem aplicadas em diferentes cenários, com diferentes volumes e estruturas de dados, confere uma maior flexibilidade que pode favorecer a exploração e análise dos dados.

Outra técnica comumente utilizada e difundida na literatura é a técnica baseada em *Horizon Charts*. Esta técnica foi introduzida por Heer, Kong e Agrawala (2009) para utilizar o espaço mais eficientemente do que simples gráficos de linha e permitir a coexistência de mais gráficos simultaneamente, viabilizando a comparação entre os valores de registros multivariados que possuem referência temporal.

A construção dos *Horizon Charts* se baseia em dois principais aspectos: o aproveitamento de espaços não utilizados por gráficos de linha, eliminando qualquer existência visual de lacuna e a utilização de cores para representar a intensidade dos valores em diferentes níveis (crescimento positivo ou negativo em relação à uma base de referência). Na Figura 17 pode ser observado o processo de construção de *Horizon Charts* a partir de um gráfico de linha.

Figura 17 – Construção de um *Horizon Chart* a partir de um gráfico de linha, ilustrado em três passos. Adaptado de Heer, Kong e Agrawala (2009)

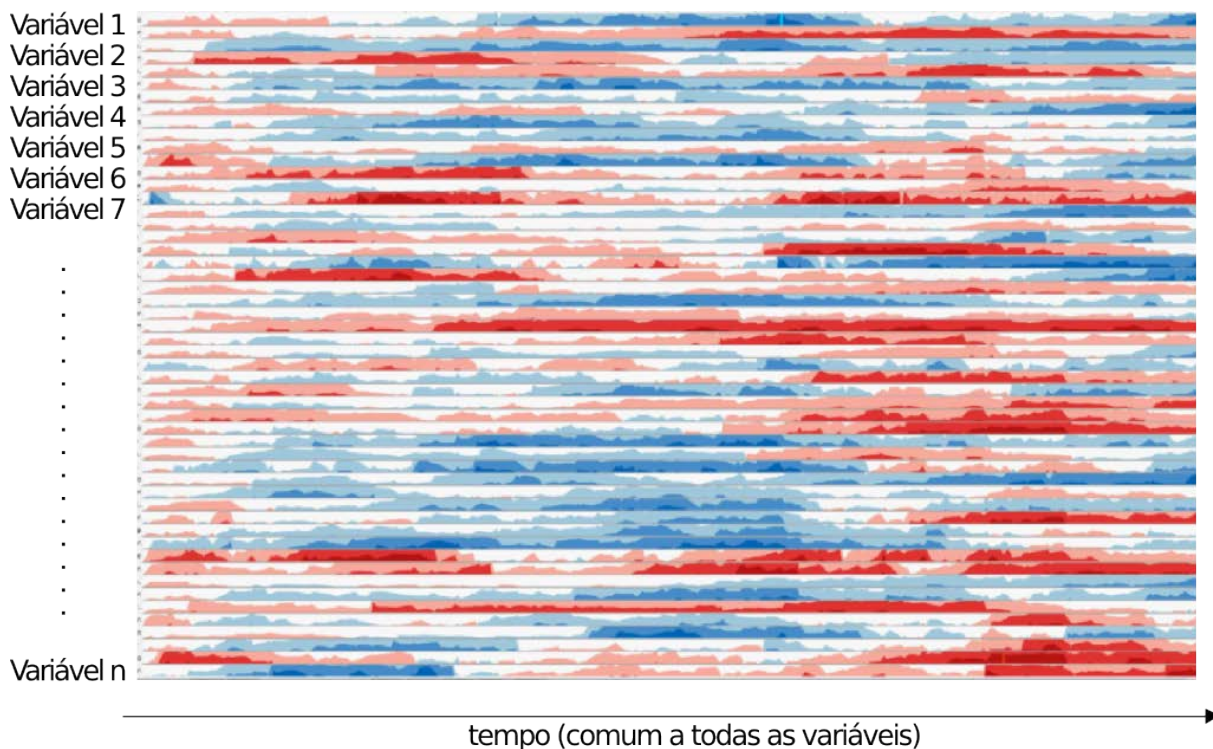


A representação ilustrada na Figura 17 é elaborada em três etapas básicas: segmentação da curva em níveis previamente determinados de acordo com a intensidade do valor; deslocamento dos picos de cada subdivisão em direção ao eixo de origem; espelhamento dos blocos negativos ou deslocamento do bloco para o eixo de origem de modo a formar uma única faixa com os dados representados de maneira visual.

A abordagem com os *Horizon Charts* se torna útil em razão da facilidade de percepção do comportamento da série temporal a partir do aspecto geométrico (curva dos blocos) e perceptivo (cores relacionadas às intensidades), sobretudo na comparação com

outra variável no mesmo intervalo de tempo. O aproveitamento do espaço na apresentação de diversas variáveis simultaneamente pode ser significativo na manipulação de dados multivariados, como pode ser visto no exemplo apresentado na Figura 18, no qual exibe simultaneamente mais de cinquenta variáveis que correspondem a um conjunto de dados de teste.

Figura 18 – Apresentação de registros multivariados de uma série temporal, onde cada linha representa uma variável. Adaptado de Aigner et al. (2011)

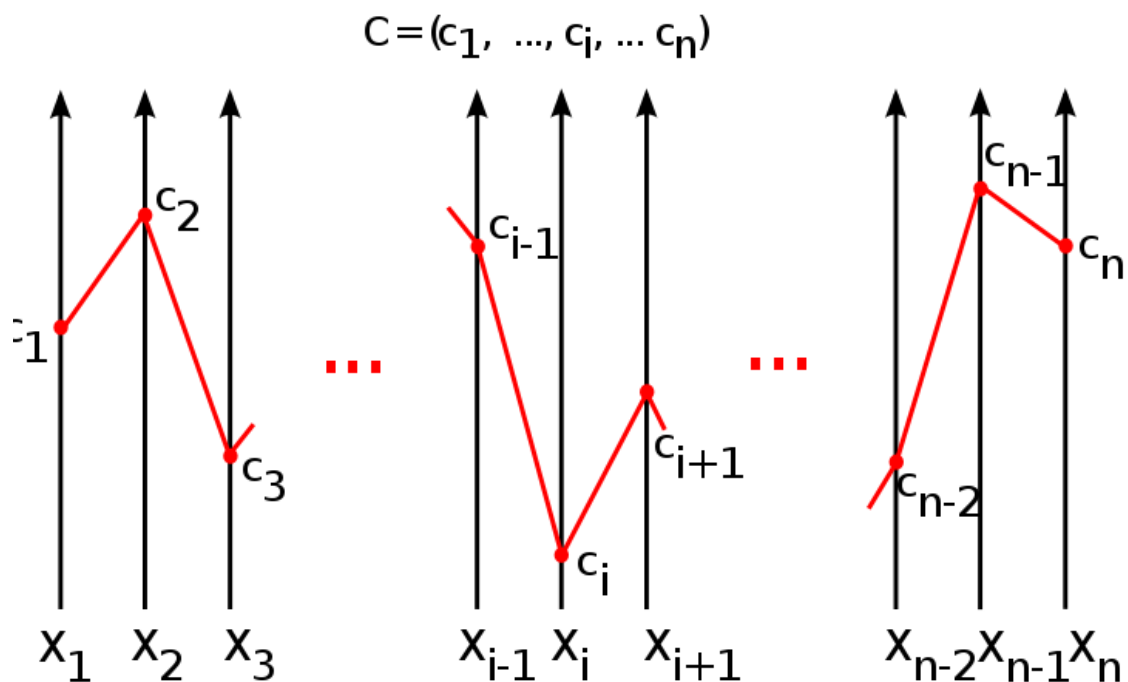


A representação visual por *Horizon Chart*, como visto na Figura 18, permite exibir paralelamente um conjunto de variáveis em diferentes escalas de valores e unidades de medida, mas que compartilham o mesmo eixo temporal. As cores aplicadas na construção do gráfico facilitam a percepção da variabilidade dos dados e ressaltam algumas possíveis relações entre os valores apresentados por diferentes variáveis. No exemplo da figura, áreas em vermelho representam valores negativos e áreas em azul indicam valores positivos. A intensidade da cor aponta a amplitude do determinado valor.

Cada variável é escalonada de forma que o valor apresentado esteja de acordo com um valor de referência, ou seja, cada atributo é identificado por meio da variação positiva ou negativa em relação ao valor de referência. Dependendo do problema ou do objetivo da análise, esta referência pode ser diferente, por exemplo, caso o interesse seja a observação do desvio em relação à uma medida central dos valores do conjunto, a média de cada variável poderia ser tomada como referência; ou caso seja necessário observar a existência de valores positivos e negativos, o valor zero pode ser atribuído como referência.

Outro tipo de representação visual que pode ser destacada no que se refere ao aproveitamento do espaço é a técnica de Coordenadas Paralelas que é específica para ser utilizadas para a exploração e análise de dados compostos por múltiplas variáveis. Esta técnica foi concebida por Inselberg e Dimsdale (1990) como uma metodologia utilizada para análise visual de dados representados no R^N sem a necessidade de mapeamento com projeção n-dimensional para bidimensional, de forma que uma hipersuperfície ¹ possa ser apresentada em uma imagem planar (duas dimensões), como pode ser visto na Figura 19.

Figura 19 – Equivalência entre um registro composto por n atributos e sua representação em Coordenadas Paralelas. Adaptado de (INSELBERG; DIMSDALE, 1990)

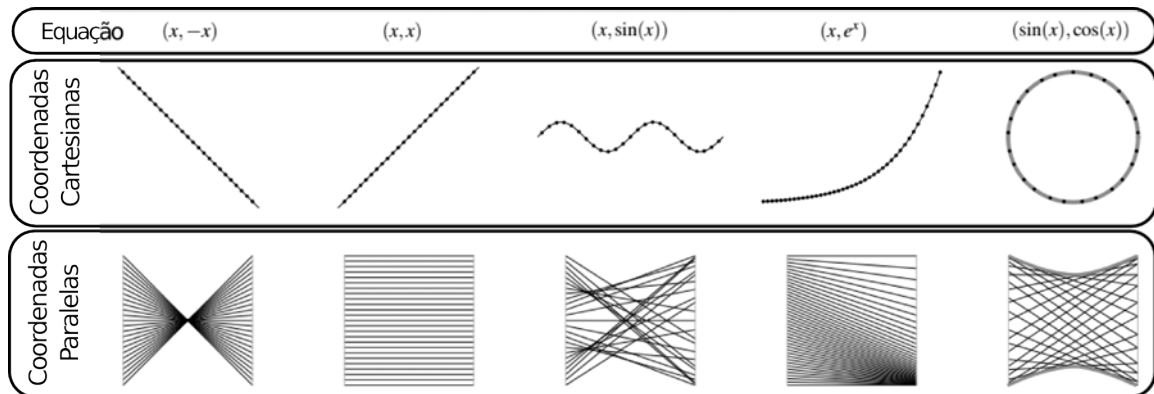


Com a construção das Coordenadas Paralelas, cada registro do conjunto de dados não representa apenas um ponto, como em um plano cartesiano, mas sim uma polilinha que intercepta cada eixo das variáveis em seu respectivo ponto significativo. A observação dos comportamentos das diferentes linha conectados entre os eixos permite extrair informações sobre padrões de comportamentos e relações entre os dados. A relação entre o comportamento dos dados no plano cartesiano e no plano de coordenadas paralelas é explorada por Heinrich e Weiskopf (2012) e apresentada na Figura 20.

Na Figura 20 são definidos comportamentos que podem ser observados a partir da disposição das retas que interceptam os eixos das variáveis. Relações inversas, lineares, cíclicas e exponenciais podem ser percebidas por inspeção visual e permitem a compreensão da variabilidade dos dados. Essas relações podem ser observadas em diferentes conjunto de dados, desde que os eixos das coordenadas sejam apresentados de maneira que proporcione

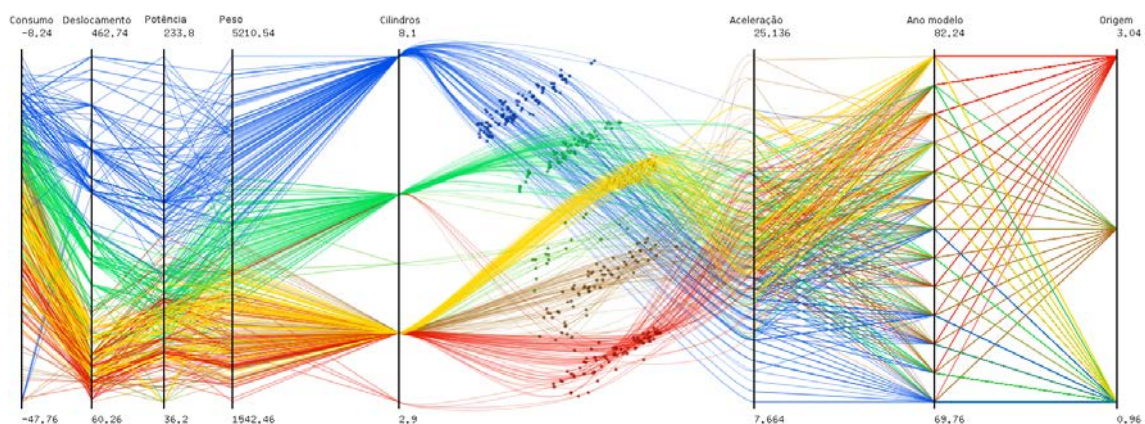
¹ Hipersuperfície é um objeto gráfico que representa uma superfície n-dimensional

Figura 20 – Padrões comuns apresentados no plano cartesiano e no plano de coordenadas paralelas. Adaptado de (HEINRICH; WEISKOPF, 2012)



esta percepção, pois para serem observadas, os eixos relacionados devem ser vizinhos. Um exemplo de algumas variáveis correlacionadas pode ser visto na Figura 21 que apresenta uma representação de um conjunto de dados sobre modelos de carros, contendo 392 registros e 7 variáveis (YUAN et al., 2009).

Figura 21 – Representação visual por Coordenadas Paralelas de um conjunto de dados com informações sobre carros. Adaptado de (YUAN et al., 2009)

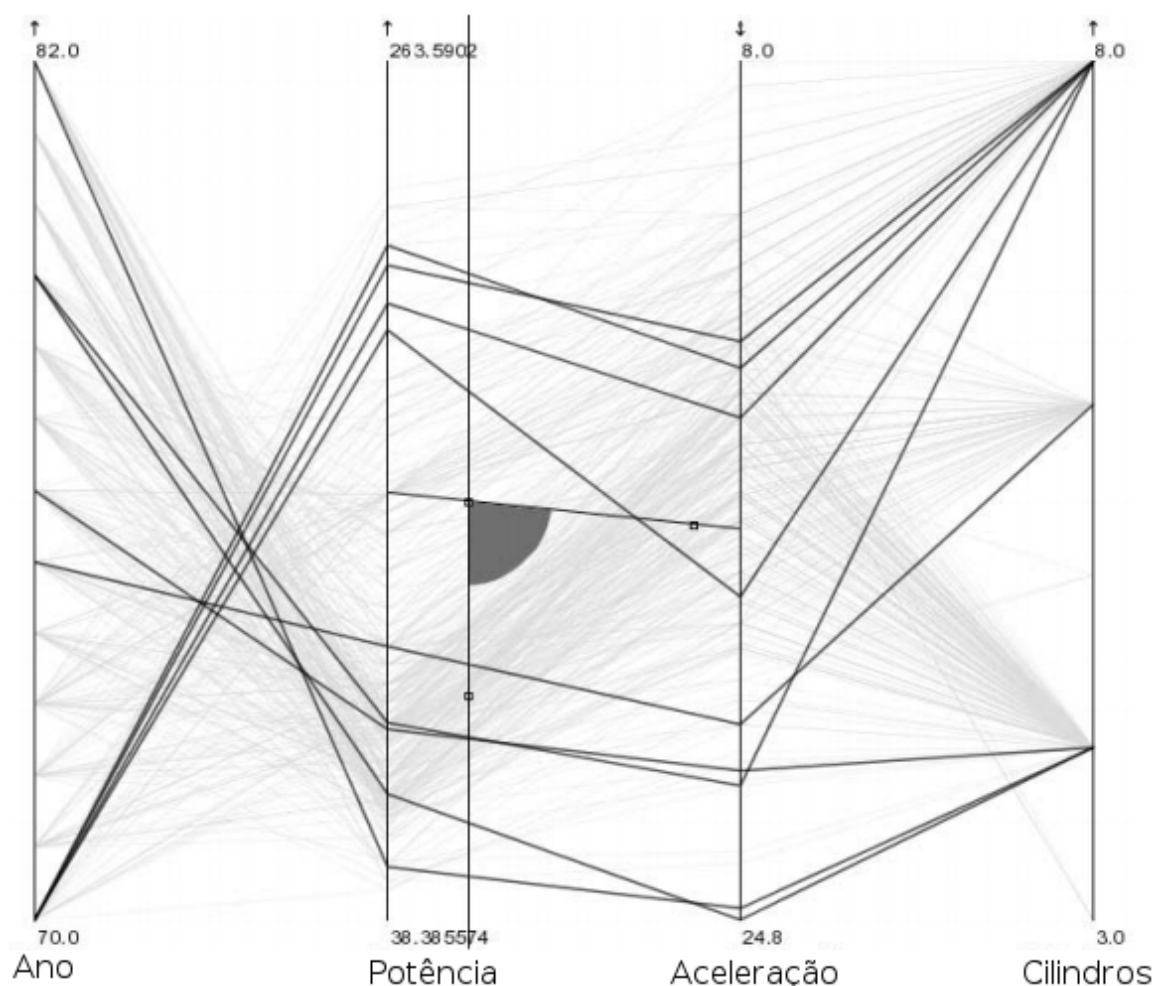


Na representação por Coordenadas Paralelas, exibida na Figura 21, os registros foram diferenciados pelas cores por meio do atributo relacionado à quantidade de cilindros. Esta disposição dos eixos das variáveis, permite a observação de alguns padrões visíveis, como: relação direta entre o consumo, deslocamento do motor, potência, peso e número de cilindros e relação inversa entre número de cilindros e aceleração. Outras relações entre os dados não podem ser facilmente percebidas sem a reorganização dos eixos das variáveis.

A representação por Coordenadas Paralelas auxilia o analista na percepção conjunta e simultânea do comportamento dos atributos do conjunto de dados e se beneficia da capacidade de representar um espaço n-dimensional sem a necessidade de redução da dimensionalidade dos dados por meio de técnicas de projeção ou redução de dados. Existem algumas variações desta técnica, como pode ser visto na Figura 22, onde a distância entre

cada eixo das variáveis representa a linearidade e correlação entre a variabilidade dos atributos (HAUSER; LEDERMANN; DOLEISCH, 2002).

Figura 22 – Representação de um conjunto de dados composto por cinco variáveis relacionados a carros. A distância entre os eixos representa a correlação entre as variáveis vizinhas (HAUSER; LEDERMANN; DOLEISCH, 2002)



Apesar da representação visual por Coordenadas Paralelas não ser diretamente voltada para a representação de dados temporais, adaptações podem ser feitas para adicionar essa percepção e aproveitar a capacidade de integração com a representação por cor, de modo que a variável tempo seja mapeada para o espaço de cores.

3.2.1 Representação visual por máscara de bits

Atributos temporais vem se tornando cada vez mais presentes e importantes em aplicações de conjunto de dados, trazendo a característica de ser relevante para que possa ser compreendida por pessoas com diferentes níveis de experiências e necessidades (SILVA; CATARCI, 2000). As representações visuais não devem tratar a componente temporal como um parâmetro quantitativo comum, assim como outros atributos, pois dados dependentes

do tempo necessitam de um tratamento diferenciado para que o significado temporal associado seja percebido (MULLER; SCHUMANN, 2003).

Atenção especial deve ser direcionada para a construção de representações visuais de dados que possuem o atributo temporal e são multivariados (vários atributos). Uma abordagem comumente utilizada é a associação de características visuais, como cor, tonalidade, posicionamento e tamanho, a cada um dos possíveis parâmetros (dimensão) dos dados (HEALEY; BOOTH; ENNS, 1995).

Dependendo do contexto de aplicação da exploração e análise, informações sobre as características do conjunto de dados podem fornecer subsídios importantes para complementar a análise do significado dos dados. Características relacionadas à variáveis dicotômicas ² dos registros do conjunto de dados, como integridade (registrado ou ausente), consistência (válido ou inválido), qualidade (confiável ou não) e outras particularidades implícitas, podem auxiliar na identificação de padrões de comportamentos no conjunto de dados analisado.

Diante destes aspectos, é possível implementar uma representação visual que ressalte determinada característica dos registros de um conjunto temporal e multivariado de dados. O objetivo principal dessa visualização é possibilitar a identificação de padrões na exibição simultânea de variáveis dicotômicas associadas a cada atributo do conjunto de dados multivariado e com referência temporal.

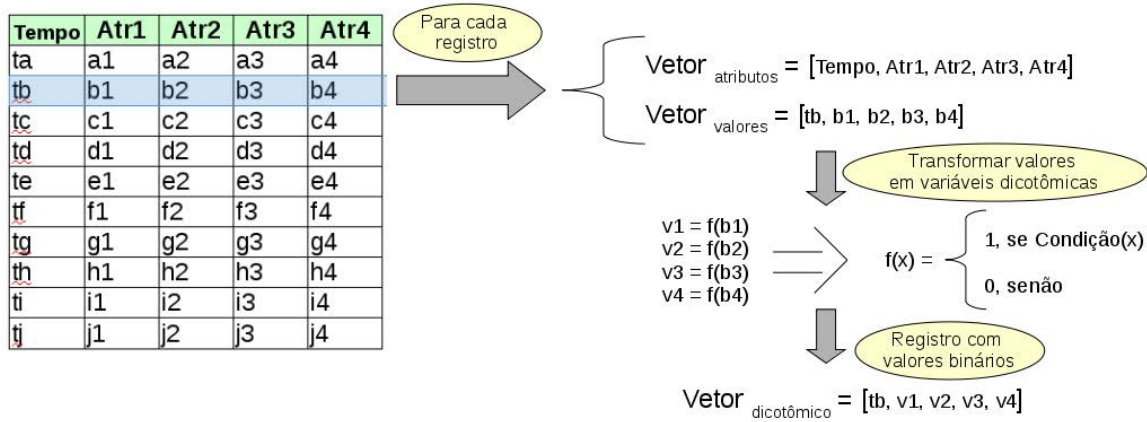
Por se tratar de conjuntos de dados multivariados, cada elemento do conjunto pode ser visto como um vetor de atributos no qual, originalmente, cada posição representa uma característica associada a um valor, seja ele quantitativo ou qualitativo. No entanto, o interesse dessa representação visual não é relacionar informações sobre o significado dos valores dos atributos, mas sim relações entre variáveis dicotômicas associadas ao estado de cada atributo do conjunto. Nesse sentido, os atributos de cada registro devem ser convertidos para os significados dicotômicos e mapeados para os valores binários (0 e 1), como pode ser visto na Figura 23.

O processo de transformação dos dados em variáveis dicotômicas e, em seguida, para valores binários, é realizado para todos os atributos de cada registro do conjunto de dados, exceto para o atributo que traz a informação sobre o tempo, pois esse segue em formato original para ser utilizado na composição da representação visual. Uma vez realizada essa conversão, os dados não representam mais o fenômeno indicado por cada atributo, mas sim um evento, definido pela função de conversão, na qual o comportamento de todos os atributos do vetor pode ser identificado.

Por exemplo, em uma fábrica existem quatro geradores auxiliares de energia que possuem sensores que registram a capacidade de potência gerada por hora e registram estes

² Variáveis dicotômicas são atributos qualitativos nos quais só existem dois valores possíveis, como sim ou não, ligado ou desligado, entre outros

Figura 23 – Processo de transformação de cada atributo do conjunto em variáveis dicotômicas



valores constantemente no banco de dados. A realização da conversão destes valores em variáveis dicotômicas poderia ser baseada na utilização de um limiar: se o gerador registra um valor de potência inferior a um certo limiar, significa que o mesmo está operando abaixo da capacidade por alguma falha; caso contrário está funcionando corretamente. A falha ou sucesso pode ser então convertida para os valores binários (0 para "falha", 1 para "funcionando"). Dessa forma, um registro contendo os valores [1, 1, 0, 0] indicaria o evento de falha no terceiro e quarto gerador do conjunto.

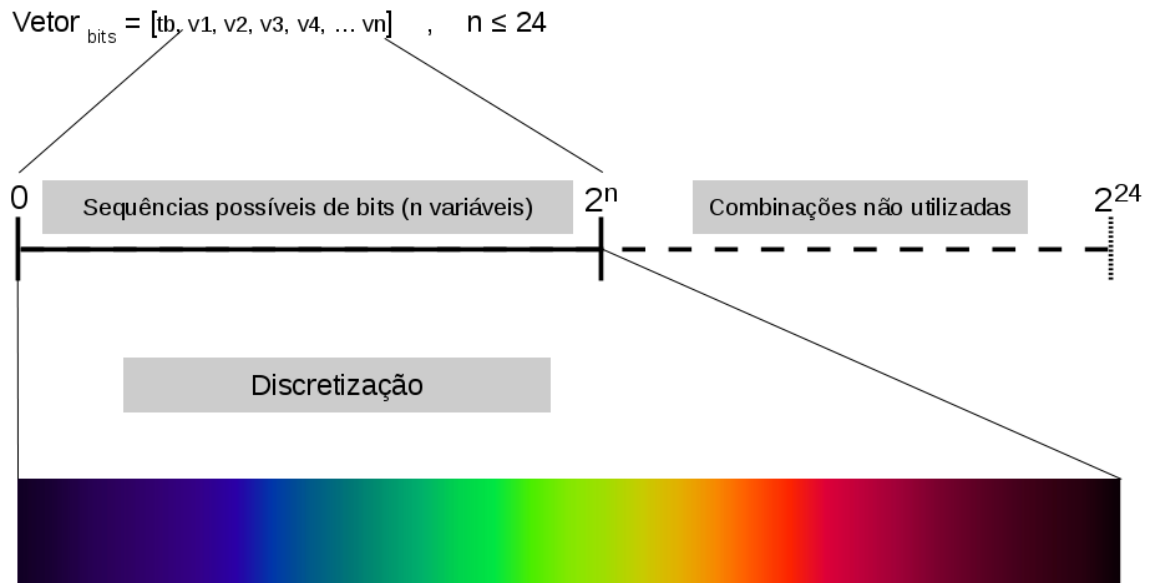
Para permitir a representação da informação relacionada ao evento identificado pelo vetor de valores binários, esta sequência é então convertida para o espaço de cores, de forma que cada possibilidade de evento seja mapeada para uma determinada cor que permita a diferenciação entre os eventos. O espaço de cores RGB é formado por três canais de cores, na qual cada um pode ser escrito por oito bits, ou seja, 256 (2^8) níveis em cada canal. Ao todo, os 24 bits dos três canais são capazes de gerar 16.777.216 (2^{24}) valores distintos.

Para aumentar a capacidade de diferenciação entre as cores que representam os eventos, principalmente quando o número de variáveis do conjunto de dados for significativamente menor que as possibilidades do espaço de cores (sequência de 24 bits), o mapeamento é realizado por meio da discretização do espaço representado pela sequência de bits para o espaço total de cores (Figura 24). Com esta abordagem, todo o espaço de cores poderá ser contemplado, mesmo quando existirem poucas variáveis.

O espectro de cores utilizado na Figura 24 é apenas ilustrativo para facilitar o entendimento do processo, pois o intervalo de cores utilizado para processamento e configuração da representação visual se estende a todo o espaço de cores RGB, contendo 16.777.216 valores.

O processo de transformação em atributos dicotômicos, transformação em variáveis binárias e discretização para o espaço de cores, converte os vetores de valores contendo

Figura 24 – Discretização da sequência binária para o espaço de cores



n variáveis de cada registro em um vetor contendo apenas duas posições: tempo e cor. Com os dados do evento estudado sintetizados em apenas duas variáveis, a forma de representação visual pode ser explorada. Visando facilitar a identificação de padrões de comportamentos dos eventos representados pelas cores no decorrer do tempo, foi utilizada a metáfora semelhante aquela aplicada à técnica *Calendar View* (WIJK; SELOW, 1999).

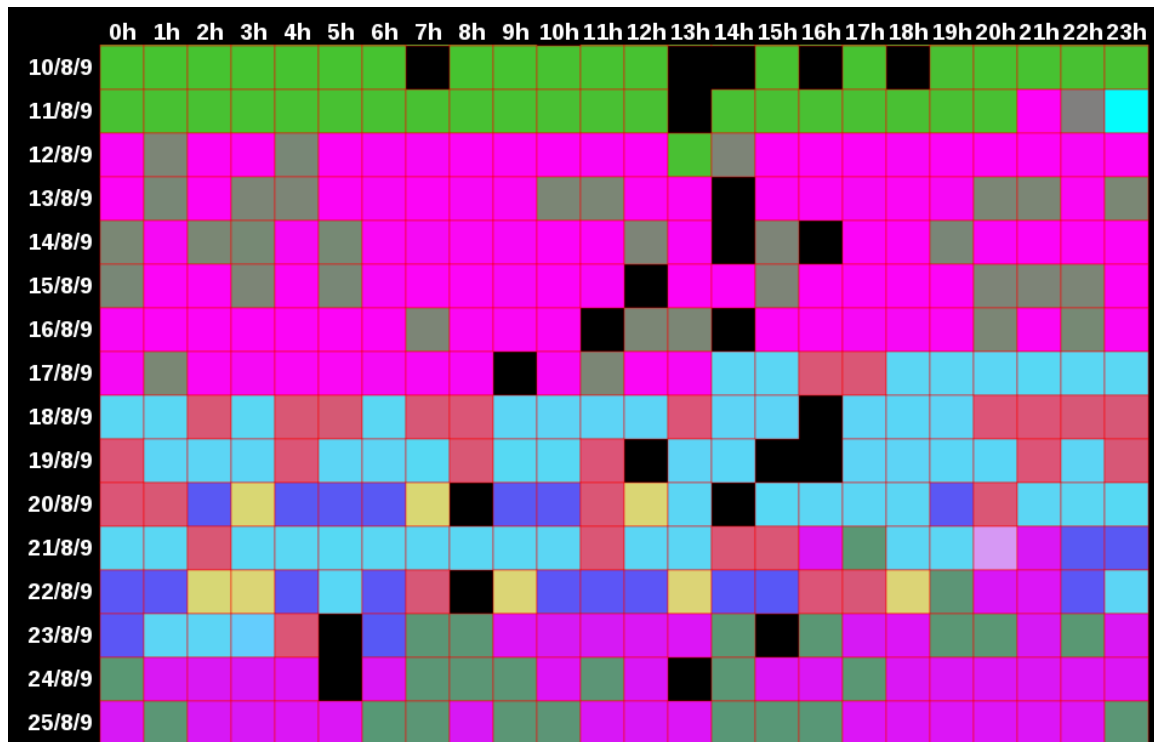
Na representação proposta, as cores são apresentadas em uma matriz na qual cada linha representa um dia presente no eixo temporal e cada coluna representa o intervalo de 1 hora do respectivo dia específico, contendo então, 24 colunas, como pode ser visto na Figura 25.

Essa representação possui a escala temporal mínima de 1 hora. Caso a intenção esteja relacionada a estudar o comportamento de acordo com uma escala de tempo mais abrangente, como diária, semanal ou mensal, a matriz de informações pode ser reelaborada para ressaltar um determinado ciclo de tempo.

Com a representação visual por máscara de bits é possível identificar padrões de cores que definem quais os atributos foram responsáveis pelo acontecimento de um determinado evento. Utilizando o mesmo exemplo do conjunto de geradores da fábrica, caso em um determinado momento, o evento de falha seja identificável pela sequência 0011 e posteriormente se altere para a sequência 0101, esses eventos serão identificados por cores diferentes na matriz de informações, de forma que mesmo se tratando de eventos que apresentam falhas na mesma quantidade de geradores, são sequências que indicam eventos distintos (falha no primeiro e segundo gerador e depois falha no primeiro e terceiro), portanto, serão diferenciados pelas cores das células da matriz.

Esta representação visual é eficiente para a identificação de padrões nos aconteci-

Figura 25 – Exemplo de representação visual por máscara de bits. Na matriz, cada dia é representado por uma linha, cada coluna indica o intervalo de 1 hora e cada cor indica a conversão da sequência de bits para o espaço de cor



mentos de eventos mediante inspeção visual, no entanto, traz algumas limitações impostas pela sua configuração. Uma limitação está relacionada à quantidade máxima de 24 variáveis possíveis de serem representadas na matriz de informações, uma vez que a identificação do evento é realizada pela cor limitando à sequência de 24 bits (RGB) e cada variável define se o evento ocorreu ou não (dois estados) para cada um dos atributos. Outra restrição é a dificuldade de identificação do evento a partir das cores dispostas na matriz, pois é uma abordagem que prioriza a diferenciação entre os eventos que ocorrem ao longo do tempo, sendo que a identificação direta do evento representado só pode ser obtida mediante recursos de interação com as células da matriz que representam os valores das sequências de bits pelas cores.

CAPÍTULO 4

ESPECIFICAÇÕES DO CONJUNTO DE DADOS

Para o desenvolvimento deste trabalho foram utilizados os dados coletados pelo projeto SIMA. Os dados produzidos pelo SIMA representam características ambientais adquiridas por dois tipos de sensores, localizados acima e abaixo da linha da água, que são capazes de coletar um conjunto de variáveis específicas que caracterizam o objeto de análise. Os dados coletados pelas plataformas SIMA são disponibilizados para acesso via Web por meio do geoportal ¹. Neste portal, é possível obter os dados, mediante acesso identificado e autorizado, em estrutura de texto com os dados separados por vírgulas - formato *Comma-Separated Values* (CSV). Detalhes sobre a estrutura dos dados podem ser vistos na Seção 4.1.2.

Todos os dados coletados e armazenados possuem as componentes de espaço (referência geográfica da plataforma SIMA) e tempo (momento de realização da coleta). Foram realizados estudos direcionados para a compreensão da estrutura e formato dos dados armazenados, para permitir a definição de estratégias que garantam a leitura e posterior exploração dos dados. Em razão da existência de atributos espaciais e temporais, o banco de dados PostgreSQL foi utilizado como Sistema Gerenciador de Banco de Dados (SGBD) juntamente com a extensão espacial PostGIS. Tal escolha foi motivada por se tratarem de softwares *open source*, que recebem atualizações constantes pela comunidade desenvolvedora, garantindo um arcabouço robusto de ferramentas para a modelagem de dados, sobretudo de dados com componentes espaciais.

O conjunto de dados utilizados para o desenvolvimento deste trabalho foi aquele

¹ Geoportal SIMA: <dsr.inpe.br/hidrosfera/sima/>

coletado, processado e armazenado pelo projeto SIMA. O entendimento do cenário de coleta de dados, bem como a arquitetura do projeto SIMA é de fundamental importância para compreender a fonte de dados utilizada. As demais Seções detalham as características do projeto SIMA.

4.1 Sistema Integrado de Monitoramento Ambiental - SIMA

O estudo e análise de ambientes aquáticos visando o entendimento de seu funcionamento e acompanhamento de mudanças em seu comportamento exige a observação de diversos atributos relacionados. Metodologias baseadas em medidas de amostras *in situ* e por sensoriamento remoto são amplamente utilizadas para coleta dos dados relacionados. No entanto, tais abordagens podem não ser suficientes para atender análises que demandam alta resolução temporal. Para isto, equipamentos responsáveis pela realização de coletas automáticas e periódicas, como o SIMA, são necessários.

O SIMA é formado por um conjunto de tecnologias aplicadas à coleta de dados e monitoramento da hidrosfera que foi inicialmente desenvolvido em uma parceria entre a Universidade do Vale do Paraíba (UNIVAP) e o Instituto Nacional de Pesquisas Espaciais (INPE). Uma rede de plataformas fundeadas em reservatórios hidrelétricos compõe a arquitetura do projeto. Tais plataformas podem ser concebidas de duas maneiras diferentes no projeto, uma montada em uma boia toroidal construída de fibra de vidro, como pode ser visto na Figura 26, ou em uma plataforma flutuante construída em alumínio naval, como a apresentada na Figura 27.

Apesar de algumas diferenças, existe um conjunto de sensores comum em todas as plataformas, independente da estrutura utilizada para o fundeamento da plataforma SIMA no reservatório. A opção das variáveis medidas pelos sensores da plataforma SIMA levou em consideração fatores relacionados à relevância para definição de um índice ambiental composto por variáveis que descrevem consistentemente às alterações do ambiente aquático, a importância para caracterização dos processos de emissão de gases que colaboram para o efeito estufa, bem como os recursos necessários para a correta aquisição e transmissão dos dados para viabilizar a existência de plataformas autônomas (STECH et al., 2011).

Desde o início do projeto, em meados de 2002, 27 plataformas SIMA foram fundeadas em reservatórios hidrelétricos espalhados pelo território nacional, como pode ser visto na Figura 28. Algumas plataformas podem ser caracterizadas como estações temporárias de coleta que foram movidas para outros reservatórios, situação comum principalmente no início do projeto.

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, tornando as

Figura 26 – Estação SIMA composta por uma boia toroidal de fibra de vidro (INPE, 2014)



Figura 27 – Estação SIMA composta por uma plataforma flutuante em alumínio naval (INPE, 2014)



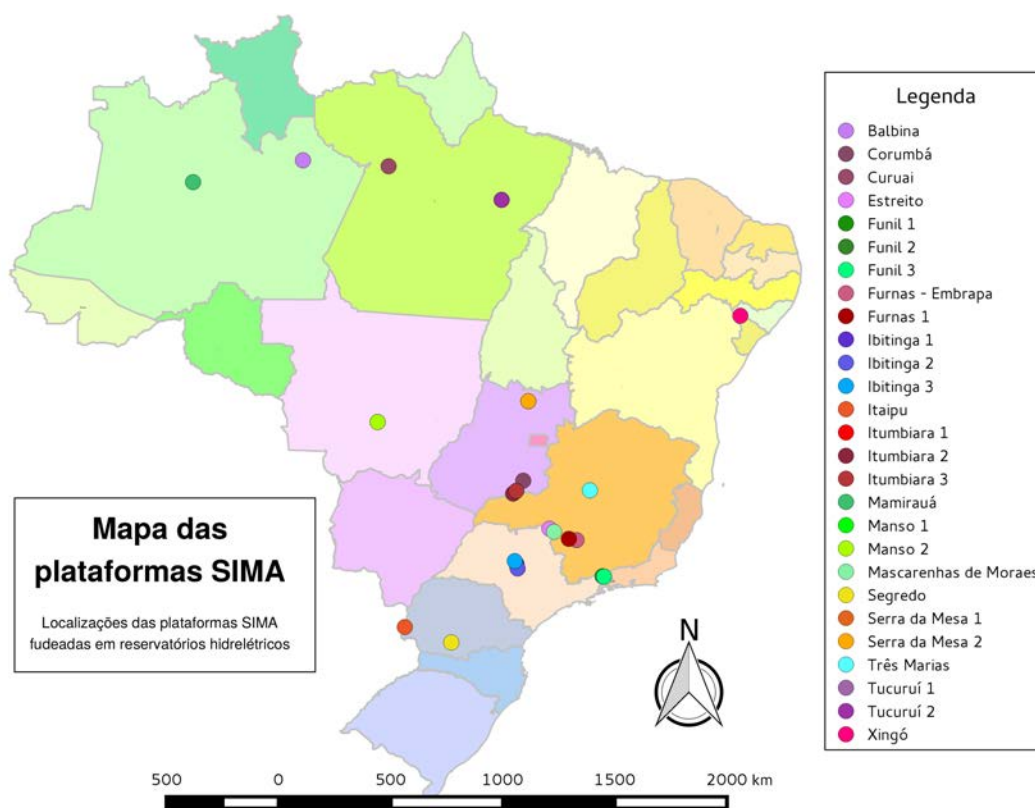
plataformas capazes de realizar coletas de atributos relacionados ao ar e à água, resultando em aproximadamente 20 variáveis coletadas, como pode ser visto na Tabela 2.

Tabela 2 – Variáveis capturadas pelo conjunto de plataformas SIMA

Atributos da água	Atributos do Ar	Outros parâmetros
Clorofila	CO2 High	Precipitação
Condutividade	CO2 Low	Pressão Atmosférica
Inundação	Direção do Vento	Temperatura Interna
NH4+	Intensidade do Vento	
NO3-	Intensidade do Vento 1	
Oxigênio Dissolvido	Temperatura do Ar	
Ph	Umidade Relativa do Ar	
Radiação Incidente	Velocidade Meridional da Corrente	
Radiação Refletida	Velocidade Meridional do Vento	
Temperatura da Água	Velocidade Zonal da Corrente	
Temperatura da Sonda	Velocidade Zonal do Vento	
Turbidez		

As plataformas integrantes do projeto SIMA realizam a leitura dos sinais dos sensores com a periodicidade de uma hora. Após a leitura, os dados coletados são transmitidos via enlace de satélite para servidores intermediários em estações terrestres que são responsáveis por receber os dados transmitidos e realizar a verificação da existência

Figura 28 – Mapa das localizações das 27 plataformas SIMA, nomeadas como: Balbina, Corumbá, Curuai, Estreito, Funil 1, Funil 2, Funil 3, Furnas - Embrapa, Furnas 1, Ibatinga 1, Ibatinga 2, Ibatinga 3, Itaipu, Itumbiara 1, Itumbiara 2, Itumbiara 3, Mamirauá, Manso 1, Manso 2, Mascarenhas de Moraes, Segredo, Serra da Mesa 1, Serra da Mesa 2, Três Marias, Tucuruí 1, Tucuruí 2, Xingó

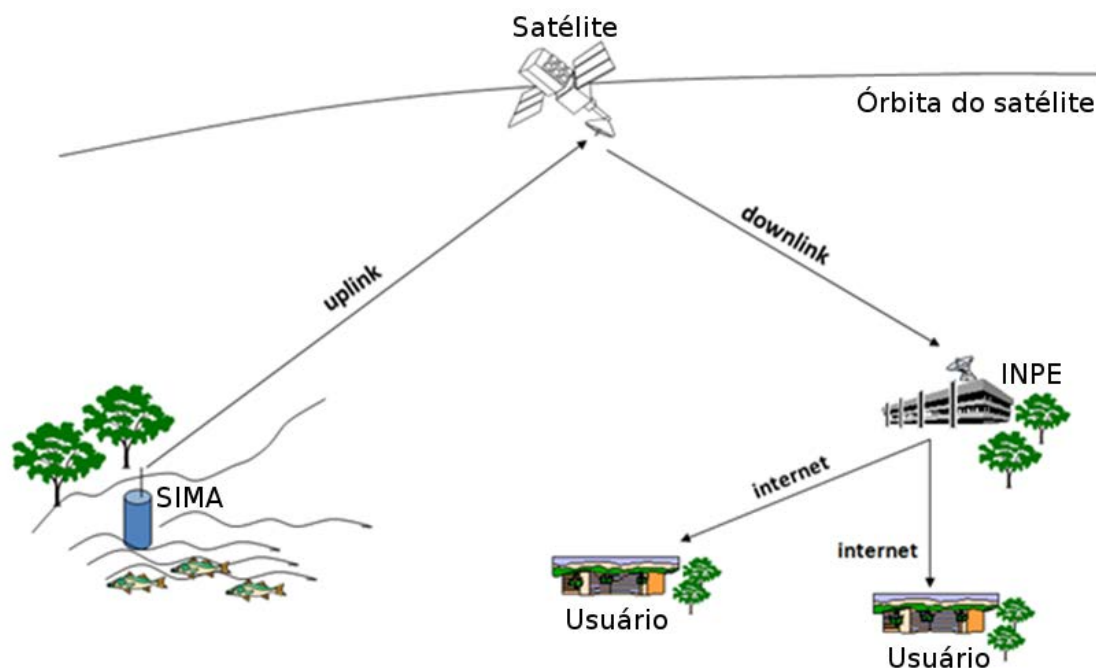


de erros na transmissão dos sinais. Após esta validação dos dados, estes são transmitidos ao servidor no centro de armazenamento, os quais passam pelo processo de decodificação, processamento e armazenamento e ficam disponíveis em um portal da internet (www.dsr.inpe.br/hidrosfera/sima/login.php). A arquitetura de comunicação e fluxo de troca de mensagens pode ser observado na Figura 29.

4.1.1 Características do processo de coleta dos dados

Embora a automatização do processo de coleta de dados ofereça benefícios, relacionados à resolução temporal da realização das coletas e capacidade de coletar múltiplas variáveis, que são difíceis de se obter com outras abordagens e métodos de aquisição de dados, como sensoriamento remoto e medições *in situ*, essa abordagem depende do bom funcionamento de dispositivos frágeis que podem apresentar falhas com o passar do tempo, como discutido por Alcântara et al. (2013). Um dos fatores que prejudicam a qualidade dos dados coletados é a exposição dos sensores às intempéries do ambiente em que se encontra, que influencia diretamente na vida útil dos sensores, como pode ser visto na Figura 30, na qual é apresentada a imagem de uma sonda extraída de uma plataforma

Figura 29 – Arquitetura do projeto SIMA (INPE, 2014)



SIMA durante o processo de manutenção. A degradação constante e contínua dos sensores pode acarretar no prejuízo gradativa da qualidade dos dados coletados.

Figura 30 – Ação do tempo sobre o estado de uma sonda, extraída de uma plataforma SIMA, devido à exposição ao ambiente (INPE, 2014)



Outro fator relevante apontado por Alcântara et al. (2013) é a possibilidade de perdas de dados durante o momento de transmissão para o satélite. Em razão dos recursos de armazenamentos limitados das plataformas, existe um *buffer* de memória pequeno capaz de armazenar temporariamente cerca de oito registros de coletas. A cada ciclo completado no intervalo de uma hora é realizada uma tentativa de envio do registro mais antigo presente na memória para o satélite. Caso a órbita do satélite não esteja visível no alcance da plataforma, o registro é mantido no *buffer*. No entanto, com a necessidade

da realização de novas coletas, caso a memória esteja esgotada, os dados mais antigos são descartados para serem sobrescritos pelos novos dados, podendo resultar na perda de registros inteiros.

Outra possibilidade, porém menos recorrente, é a falha na conversão dos sinais elétricos captados pelos sensores em números discretos. Uma vez identificada essa falha no momento da validação dos dados no servidor intermediário, o dado corrompido é descartado permanentemente. Este processo de validação dos dados, é realizado em um servidor intermediário, que recebe os dados coletados antes de serem enviados para o servidor de armazenamento.

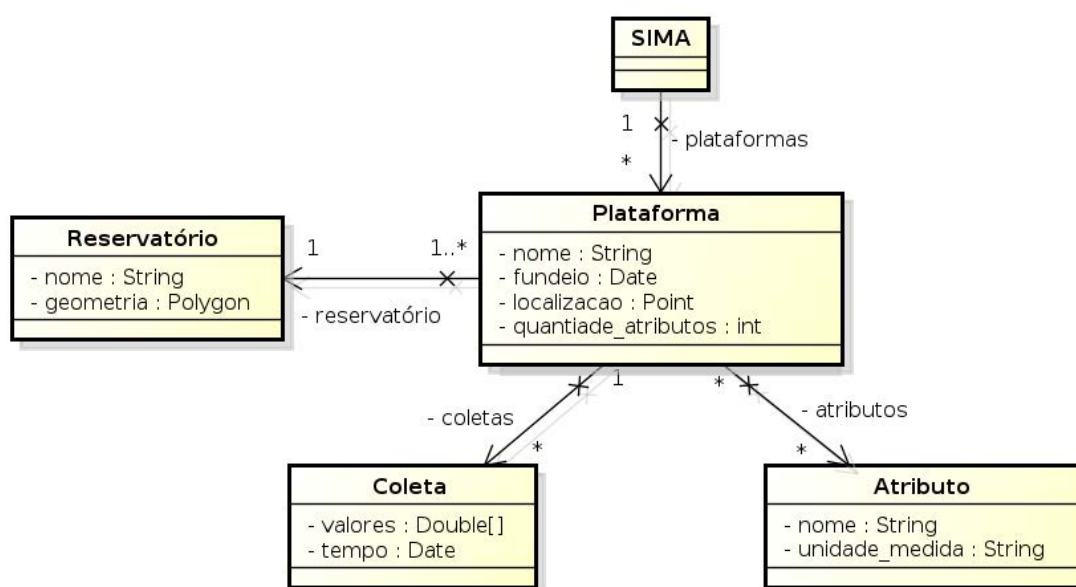
4.1.2 Estrutura dos dados coletados

Os dados coletados pelo sistema SIMA são disponibilizados no portal da internet para obtenção em formato de planilha eletrônica. Primeiramente deve ser aplicado um filtro, correspondente à seleção da plataforma de interesse, período de tempo e atributos desejados. Após este processo, os dados são disponibilizados para aquisição.

No domínio de execução das coletas realizadas pelas plataformas SIMA, existem diversas informações relacionadas com os registros coletados. Um diagrama simplificado das informações relacionadas a este processo pode ser visto na Figura 31, no qual são descritos os relacionamentos entre os elementos que foram deduzidos a partir das estruturas de dados extraídas da estrutura de dados SIMA. Neste diagrama, é possível inferir que no projeto SIMA existe um conjunto formado por várias plataformas, cada uma pertencente a um reservatório que pode conter também outras plataformas. Cada plataforma coleta uma quantidade diferente de atributos expressos em diferentes unidades de medidas. Toda coleta realizada possui a referência temporal e um vetor contendo os valores de cada atributo coletado.

Apesar do portal Web oferecer uma forma para acesso aos dados a partir da internet, aumentando a capacidade de utilização e alcance dos dados, as informações disponibilizadas se restringem aos registros coletados. Não existe qualquer outro dado disponibilizado sobre o processo de coleta, por exemplo, metadados sobre os sensores utilizados, sobre os reservatórios, descrição dos atributos, entre outros.

Figura 31 – Diagrama simplificado deduzido a partir das estruturas de dados que compõem o domínio de informações do SIMA



CAPÍTULO 5

RESULTADOS E DISCUSSÕES

Foram direcionados esforços para a escolha e implementação das técnicas de Visualização de Informação mais eficazes para serem aplicadas sobre o conjunto de interesse. Foi considerada a possível inserção de recursos que permitam a interatividade com o analista, de acordo com o fluxo de VA, discutido no Capítulo 3, de modo que o analista possa interagir diretamente no processo de exploração e análise dos dados, participando ativamente do processo de seleção, filtragem, manipulação e refinamento do processo de análise, de acordo com o *sense-making loop* definido por Keim et al. (2006). Para tanto, os métodos e recursos de análises utilizados neste trabalho foram implementados utilizando bibliotecas voltadas para desenvolvimento Web, como a biblioteca JavaScript D3 (*Data-Driven Documents*)¹ (BOSTOCK; OGIEVETSKY; HEER, 2011), de modo a facilitar a inserção de recursos de interatividade e permitir escalabilidade de acesso a estes recursos, ampliadas pela possibilidade de acesso via navegador de internet em diversas arquiteturas e ambientes computacionais.

A biblioteca D3 é formada por um conjunto de recursos implementados em JavaScript que permite a manipulação direta dos elementos nativos encontrados no *Document Object Model* (DOM)², que compõem as páginas Web escritas com *HyperText Markup Language* (HTML). D3 oferece operadores para realização de vínculos (*bind*) entre os componentes nativos do DOM e documentos de fontes externas, como arquivos CSV, estruturas XML, dados em formato JSON, entre outras formas para manipulação eficiente de documentos baseados em dados (BOSTOCK; OGIEVETSKY; HEER, 2011). Tal capacidade demonstra um potencial significativo para ser integrado com o consumo de

¹ D3 é uma biblioteca JavaScript *open source* disponível em: <<http://d3js.org/>>

² DOM define as estruturas lógicas dos documentos HTML e XML e as maneiras pelas quais estes documentos são acessados e manipulados

dados por meio de *Web Services*, como XML e JSON, uma vez que a própria biblioteca D3 oferece funções para acesso a dados disponíveis em *Web Services*.

A possibilidade de oferecer ferramentas interativas, de representação, exploração e análise de dados, no ambiente Web é um fator importante que pode 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; KELLEHER, 2012). Como o objetivo deste trabalho contempla a exploração das capacidades de utilização de técnicas de representação visual para apoiar o processo de análise de dados, algumas técnicas de Visualização de Informação foram adaptadas, implementadas e organizadas em um protótipo de aplicação Web voltado para a exploração e análise dos dados, nomeado como SimaVis.

Todos os resultados discutidos neste Capítulo foram processados e extraídos a partir de recursos implementados no sistema SimaVis. A implementação do sistema SimaVis em uma arquitetura Web Cliente-Servidor facilitou a exploração dos dados e permitiu a reconstrução dos cenários destacados neste capítulo, bem como sua readequação a outros conjuntos de dados disponíveis no sistema.

As próximas seções tratam dos resultados obtidos com abordagens para exploração dos dados disponibilizados pelo portal do projeto SIMA, alguns aspectos relacionados à elaboração do sistema SimaVis, bem como abordam informações que podem ser obtidas utilizando os recursos implementados no sistema. Tais informações extraídas são importantes para caracterizar o conjunto de dados e também para viabilizar análises posteriores sobre o significado dos fenômenos apontados pelos dados.

Cada seção descreve os resultados de aplicação de cada técnica, compreendendo as informações que podem ser observadas e as perguntas que podem ser respondidas utilizando as representações visuais integradas ao sistema SimaVis.

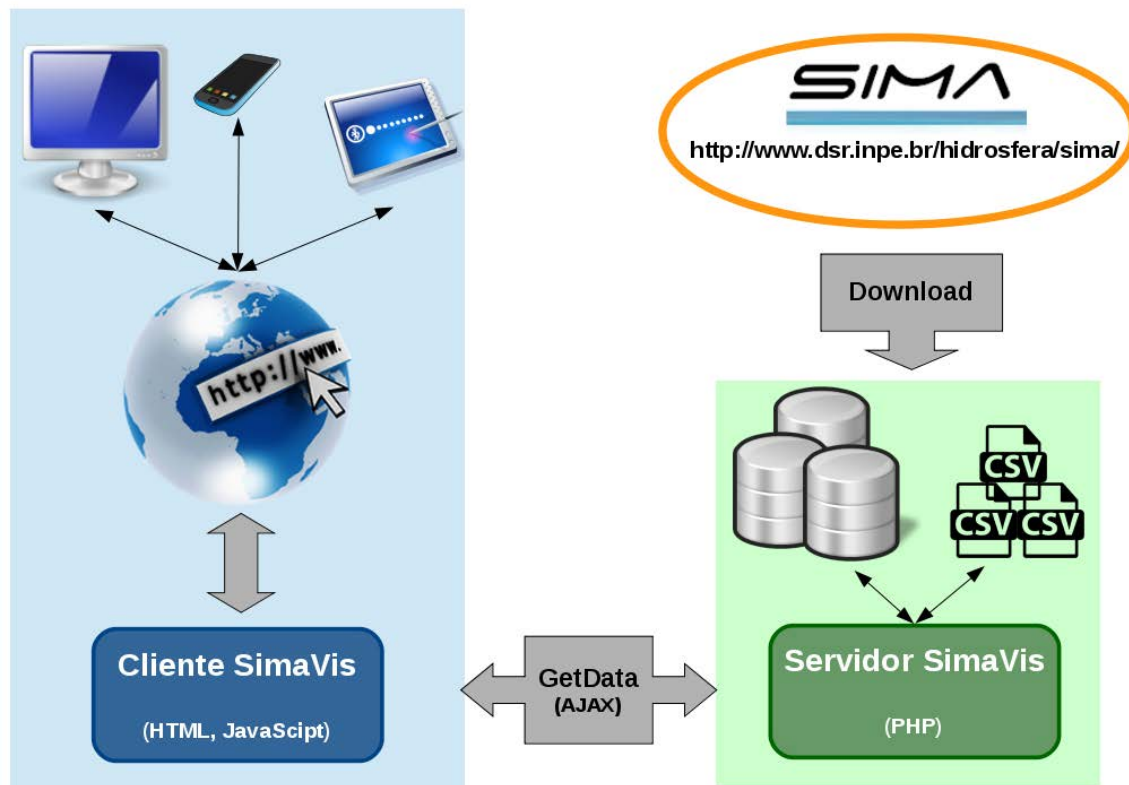
5.1 Arquitetura do SimaVis

A decisão de utilizar a biblioteca D3 para construção das visualizações permite a concepção do sistema SimaVis em uma arquitetura modularizada baseada na infraestrutura Cliente-Servidor³. Todos os recursos de navegação, exploração e interação com as representações visuais podem ser executadas dentro do contexto dos navegadores de internet, em razão da biblioteca D3 ser implementada integralmente sobre a linguagem JavaScript, que é executada em ambiente Cliente e disponível para diferentes categorias de equipamentos: computadores, smartphones e tablets. No lado do servidor, podem ser concentrados apenas os métodos de acesso e obtenção dos dados, como o SGBD e mecanismos importadores de

³ Cliente-Servidor é um modelo de computação que segmenta a arquitetura de aplicações em provedores de serviços (servidores) e consumidores destes serviços (cliente)

novos dados, como pode ser visto na Figura 32.

Figura 32 – Arquitetura modularizada do sistema SimaVis, formada por módulo Cliente, Módulo servidor e Bases de Dados



A arquitetura do sistema SimaVis é composta basicamente por três módulos. Os componentes que integram o módulo cliente são responsáveis por oferecer aos usuários as funcionalidades e recursos de interação com a interface do sistema além da própria apresentação das representações visuais geradas durante a exploração dos dados. Na interface do cliente, que pode ser visto na Figura 33, construída com HTML, JavaScript e CSS (*Cascading Style Sheets*), é possível: selecionar as plataformas que se deseja analisar; escolher os atributos/variáveis de interesse; filtrar os dados; criar diferentes representações visuais e salvar as representações visuais com comentários.

Uma vez que os parâmetros da análise estejam definidos, a consulta para obtenção dos dados (*GetData*) é enviada ao servidor por meio de requisição assíncrona (*Asynchronous Javascript and XML* - AJAX). A consulta é então processada pelo servidor, o qual realiza as operações no conjunto de dados e retorna a resposta ao cliente contendo os dados selecionados. Com o retorno dos dados, a representação visual é construída e apresentada ao usuário utilizando os recursos da biblioteca D3.

A obtenção/atualização do conjunto de dados a partir do portal SIMA está presente na arquitetura, no entanto, é uma operação que pode ser efetuada a qualquer momento fora do fluxo de processamento das consultas, por meio de acesso ao portal e aquisição

Figura 33 – Interface Cliente do sistema SimaVis, na qual é possível: (a) selecionar as plataformas que se deseja analisar; (b) escolher os atributos/variáveis de interesse; (c) limitar o intervalo de tempo para consultar os dados; (d) aplicar filtros de atributos; (e) definir qual a representação visual que se deseja utilizar e (f) exportar relatórios contendo as representações visuais e comentários



autorizada dos novos dados. Com esta abordagem, os sistemas passam a ser totalmente independentes.

5.1.1 Definição e implementação da infraestrutura de manipulação do conjunto de dados

O acesso aos dados a partir da interface cliente utilizando requisições assíncronas aumenta a flexibilidade da construção do ambiente servidor. Na estrutura Cliente-Servidor, a comunicação entre as instâncias cliente e o sistema servidor pode ser garantida desde que a requisição realizada pelo cliente esteja em conformidade com aquela que o servidor é capaz de responder, assim como a resposta do servidor deve possuir a estrutura que o cliente é capaz de interpretar. Tal fato destaca a importância da organização dos dados e dos mecanismos de comunicação entre as instâncias Cliente e Servidor.

Outro fator importante é a garantia da consistência do banco de dados, de modo que a modelagem proposta possa ser capaz de representar todas as informações disponíveis e integradas pela estrutura de dados original, oriunda do projeto SIMA, exibida na Figura 31, detalhada na Seção 4.1.2. Nesse sentido, algumas variações da infraestrutura do servidor e da modelagem do banco de dados são possíveis, como por exemplo: modelagem genérica baseada no modelo Entidade-Atributo-Valor (EAV); modelagem de acordo com o padrão SWE e modelo para múltiplas fontes de dados.

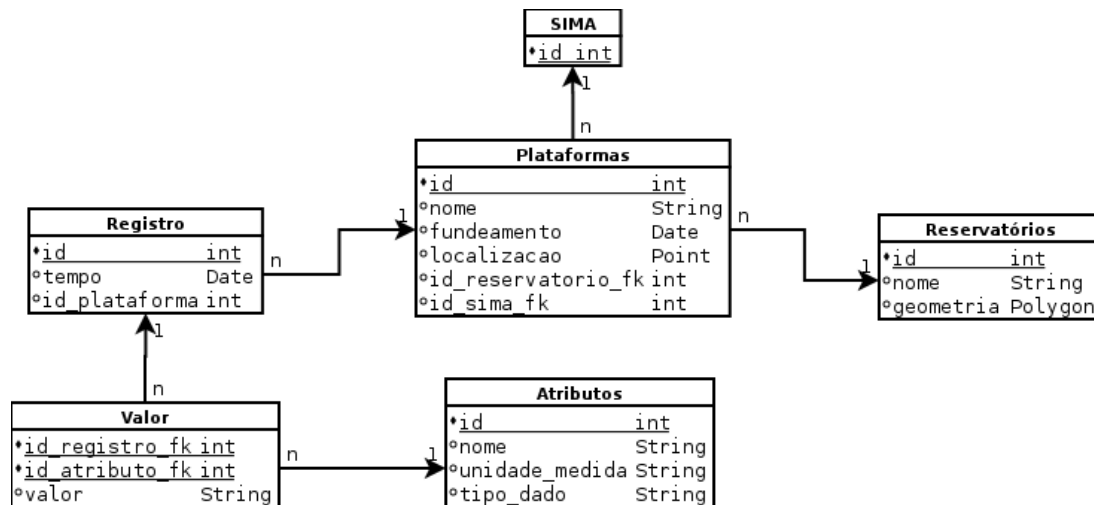
5.1.1.1 Modelo Entidade-Atributo-Valor - EAV

A primeira abordagem aplicada para a definição do modelo de armazenamento dos dados importados do projeto SIMA foi a configuração do banco de dados de acordo com o modelo Entidade-Atributo-Valor (*Entity-Attribute-Value* - EAV) que é útil para a modelagem de dados com estruturas heterogêneas (NADKARNI et al., 1999). Este modelo tem como principal característica o fato de utilizar um modelo dinâmico de banco de dados, muitas vezes nomeado como esquema aberto, ou seja, não é a estrutura física do banco de dados que define como os dados serão registrados, mas sim o esquema lógico definido por meio de metadados, tornando sua modelagem um desafio maior que os modelos convencionais, pois depende diretamente da construção efetiva do conjunto de metadados (DINU; NADKARNI, 2007). Assim como o nome deste modelo indica, os dados são tratados por meio de tuplas que indicam:

- Entidade: a entidade que este registro pertence. Existe uma referência ao conjunto de entidades definidos em outro objeto (tabela) do banco de dados;
- Atributo: Faz referência ao atributo no qual este registro representa o valor. Existe uma referência ao conjunto de atributos definidos em outra entidade do banco de dados;
- Valor: Armazena o valor significativo deste registro, no qual o tipo de dados depende do atributo de referência. Geralmente armazenado em um tipo genérico (como uma cadeia de caracteres), ou então com subconjuntos de entidades específicas para cada tipo de dados.

O comportamento deste modelo é baseado no mesmo conceito de Matrizes Esparsas, existe uma grande quantidade de atributos, mas que nem todos são usados simultaneamente por cada registro, eliminando a existência de elementos nulos (FREITAS et al., 2013). Um modelo EAV preliminar, foi proposto para armazenar os dados exportados do projeto SIMA pode ser visto na Figura 34, na qual o relacionamento entre as tabelas "Registro" e "Plataformas" representa o papel de Entidade, incluindo a componente temporal.

Figura 34 – Modelo EAV para os dados importados do projeto SIMA



A utilização deste modelo seria indicado em razão de existirem dezenas de plataformas de coleta no projeto SIMA, e cada uma delas contém atributos diferenciados entre si, além da alta frequência de lacunas e valores nulos nos dados. No entanto, esse conceito de modelagem incorpora alguns problemas de desempenho associados à recuperação massiva de dados, principalmente quando contém grande parte dos atributos selecionados, pois para responder a esta requisição, são realizadas diversas operações de seleção, união e junção no conjunto de dados.

Apesar da otimização do mecanismo de armazenamento dos dados utilizando o modelo EAV, a recuperação dos dados tem um impacto considerável, pois durante o processo de exploração e análise dos dados, são frequentemente realizadas consultas solicitando o retorno de uma grande quantidade de registros, bem como uma grande quantidade de atributos, tornando inviável a utilização do conceito EAV -em sua concepção original- para realização das análises dos dados do projeto SIMA.

5.1.1.2 Padrão OGC Sensor Observation Service - SOS

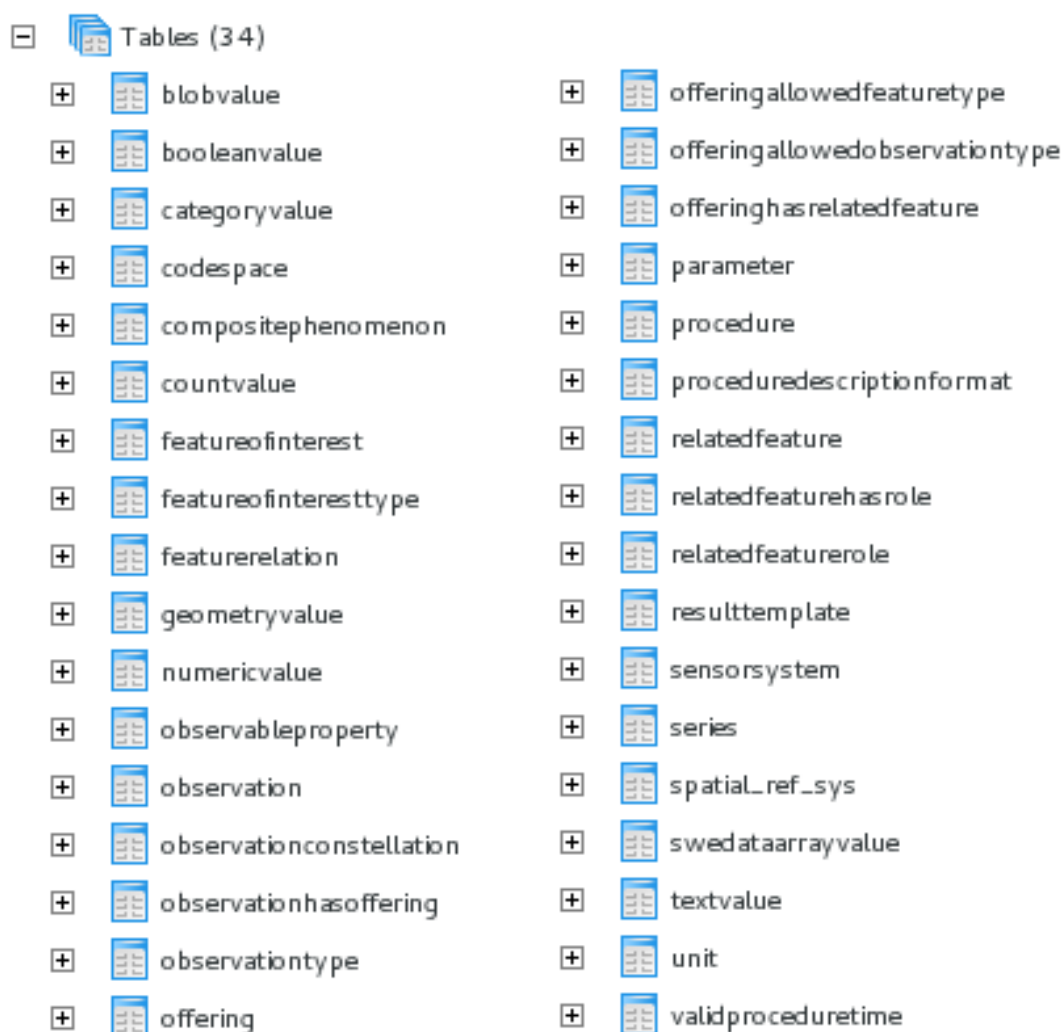
Um dos principais propósitos de trabalhos que lidam com exploração de dados é torná-los acessíveis para que sejam utilizados em outros cenários. No escopo de dados coletados por sensores, existe um esforço para a popularização da utilização de padrões internacionais para a representação e construção da infraestrutura destes dados. O OGC Consortium, como abordado na Seção 2.3, incentiva a adoção e oferece padrões para facilitar a descoberta, reutilização e interoperabilidade entre conjunto de dados. Especificamente sobre a modelagem da infraestrutura de armazenamento e acesso aos dados, as especificações do SOS - um subcomponente do SWE que é descrito na Seção 2.4.1 - trazem as diretrizes que devem ser seguidas para a padronização da modelagem dos dados.

Uma das ferramentas *open source* desenvolvidas e oferecidas gratuitamente é

o servidor 52°North-SOS que implementa todos os recursos em conformidade com as especificações da última versão do padrão SOS (atualmente 2.0), composto de: API SOS para manipulação dos dados; modelo de base de dados; *framework* de persistência no banco de dados PostgreSQL; *Web Service* implementado em Java; interface gráfica de administração e interface de instalação.

Após a realização da instalação e configuração das dependências necessárias – como servidor Apache Tomcat e drivers de acesso ao banco de dados – a instalação do servidor 52°North-SOS foi concluída seguindo as etapas descritas em sua própria interface gráfica de configuração. Posteriormente à definição de acesso ao banco de dados, um novo *schema* é criado automaticamente contendo todas as entidades, atributos e metadados especificados no padrão SOS, como pode ser visto na Figura 35.

Figura 35 – Lista de tabelas que compõem o Schema criado automaticamente pelo servidor 52°North-SOS para receber os dados



De acordo com a especificação do padrão SOS (NA; PRIEST, 2006), existem dois métodos principais de inserção de dados, definidos na extensão transacional do *Web*

Service: InsertObservation e InsertSensor. Estes métodos são implementados para atender a necessidade de inserção de dados em tempo de execução do processo de coleta dos dados, ou seja, o sensor realiza a coleta dos atributos, e esses podem ser inseridos no banco de dados por meio de requisição enviada ao *Web Service*. Esses métodos representam o cenário ideal para atualização do banco de dados por novos registros, no entanto, a transformação de um banco de dados para atender os padrões SOS, contendo previamente uma grande quantidade de dados pode ser onerosa, em razão das inserções serem individuais para cada registro juntamente com seus respectivos metadados.

Para facilitar a inserção direta de conjuntos de dados em uma instância operacional do serviço SOS, *52° North* oferece a ferramenta *52° North-SOS-Importer*. Essa ferramenta é capaz de realizar a leitura de arquivos em formato CSV e auxilia o usuário na definição dos metadados sobre tal conjunto de dados. Ao final do processo, a ferramenta gera um arquivo de configuração da requisição em formato XML, de acordo com o padrão suportado pelo serviço SOS. Um fragmento do arquivo de requisição pode ser visto na Figura 36.

Figura 36 – Fragmento da requisição XML gerada pelo *52°North-SOS-Importer* para importar dados de arquivos CSV no servidor *52°North-SOS*

```

▼<SosImportConfiguration xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://52north.org/sensorweb/sos/importer/0.2/"
  xsi:schemaLocation="http://52north.org/sensorweb/sos/importer/0.2/
    http://svn.52north.org/svn/swe/incubation/SOS-importer/trunk/52n-sos-importer-
    bindings/src/main/xsd/datafile_configuration.xsd">
  ▼<DataFile referenceIsRegularExpression="false">
    ▼<LocalFile>
      ▼<Path>
        /home/alisson/Documentos/ogc_sensors/52north/importer/furnas.csv
      </Path>
    </LocalFile>
  </DataFile>
  ▼<SosMetadata>
    <URL>http://localhost:8080/52n-sos-webapp/sos</URL>
    <Offering generate="true"/>
  </SosMetadata>
  ▼<CsvMetadata>
    ▼<ColumnAssignments>
      ▼<Column>
        <Number>0</Number>
        <Type>DATE_TIME</Type>
        ▼<Metadata>
          <Key>GROUP</Key>
          <Value>1</Value>
        </Metadata>
        ▼<Metadata>
          <Key>PARSE_PATTERN</Key>
          <Value>dd/MM/yy H:m</Value>
        </Metadata>
      </Column>
    </ColumnAssignments>
  </CsvMetadata>
</SosImportConfiguration>

```

A requisição XML é composta por: referência ao arquivo CSV que será importado; localização do serviço SOS que receberá os dados e definição dos metadados do conjunto. No entanto, durante o processo de importação dos dados SIMA utilizando o importador *52° North-SOS-Importer*, ocorreram erros indicados por interrupções causadas pela notificação de exceção "Dados inválidos ou corrompidos". Embora na especificação SOS não conste qualquer restrição sobre a capacidade de lidar com possíveis valores vazios ou nulos, o motivo da falha no processo de importação pode estar relacionado com a capacidade do importador *52°North* em tratar as lacunas de dados frequentemente apresentados sobre o conjunto de dados do projeto SIMA.

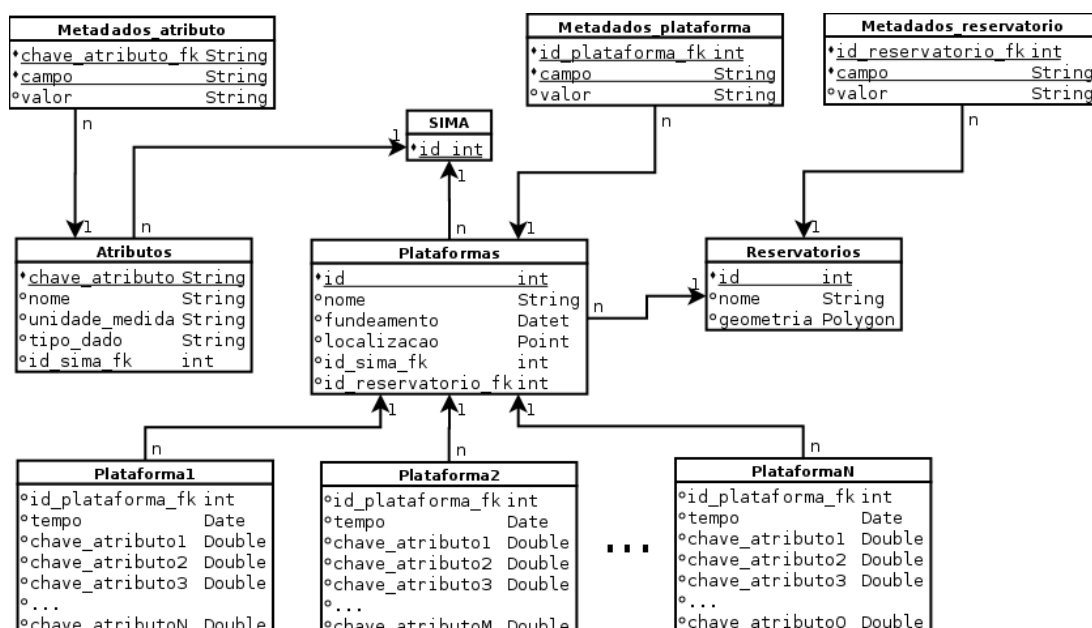
Como não foi possível a inserção efetiva de todos os dados do projeto SIMA sobre a arquitetura do servidor *52° North-SOS*, a análise dos dados utilizando esta modelagem seria prejudicada pela não completitude do conjunto de dados. Por este fato, este *Web Service* foi mantido no trabalho como possibilidade para acesso e interoperabilidade na aquisição dos dados, seguindo o padrão internacional especificado pelo OGC.

5.1.1.3 Modelagem compatível entre conjunto de dados heterogêneos

Cada plataforma SIMA possui uma quantidade diferente de sensores e coleta um conjunto diferente de atributos. Nesse sentido, para manter um equilíbrio entre otimização do armazenamento e desempenho nas consultas e recuperação de dados a serem realizadas, o banco de dados foi modelado considerando cada uma das plataformas como sendo fontes distintas de dados, ou seja, um cenário de múltiplas fontes (Fusão de Dados). Embora as plataformas possam ser consideradas como fontes heterogêneas de dados, a modelagem também precisa manter a compatibilidade entre elas, pois existem atributos em comum entre plataformas diferentes e pode ser necessário integrá-los em algum momento do processo de exploração e análise dos dados.

O Diagrama Entidade-Relacionamento (DER) ⁴ pode ser visto na Figura 37. Neste Diagrama pode ser observado que cada plataforma SIMA é representado em uma Entidade diferente composta pelos atributos específicos de cada plataforma. Todos os atributos, definidos nas colunas das entidades, fazem referência à Entidade de metadados que definem as características destes atributos.

Figura 37 – Diagrama Entidade Relacionamento que define o modelo de dados utilizado. As plataformas SIMA são vistas como múltiplas fontes de dados e podem haver quantas forem necessárias



⁴ DER é um modelo que descreve as informações de um determinado domínio de maneira abstrata

Para manter os dados integrados, a definição dos metadados dos atributos e das plataformas se mostrou como característica importante para assegurar a consistência do armazenamento dos dados e o consumo destes dados pelas técnicas de representação visual, viabilizando a integração dos dados de diferentes plataformas, no momento de recuperação dos dados.

Visando manter o padrão - mesmo que parcialmente - de acesso aos dados e permitir futura integração efetiva com o serviço compatível e implementado sobre os padrões internacionais, foi construído um *Web Service* para acesso aos dados com as mesmas interfaces definidas na extensão do núcleo das especificações do serviço SOS: GetCapabilities; DescribeSensor e GetObservation. A Tabela 3 sintetiza as operações e os parâmetros esperados pelo *Web Service* para cada tipo de operação.

Tabela 3 – Operações implementadas no *Web Service* SimaVis

Requisição	Parâmetro	Valor	Exigido	SOS
GetCapabilities	request	GetCapabilities	Sim	Sim
DescribeSensor	request	DescribeSensor	Sim	Sim
	procedure	identificado do sensor	Não	Sim
	observedProperty	identificadores do atributos	Não	Não
GetObservation	request	GetObservation	Sim	Sim
	observedProperty	identificadores dos atributos	Sim	Sim
	procedure	identificado do sensor	Não	Sim
	eventTime	timestamp inicial / timestamp final	Não	Sim

Entretanto, no contexto do trabalho foi necessário realizar uma limitação quanto à aderência do padrão, restringindo a implementação apenas à interface de acesso aos dados. O tipo de mensagem recebida e devolvida pelo *Web Service* implementado está restrita ao formato JSON e não segue o padrão de codificação Observation&Measurements (O&M) e *Sensor Model Language* (SensorML) para resposta sobre as observações e sobre os sensores, respectivamente.

5.1.2 Recursos de exploração e análise de dados implementados no software Cliente

Uma vez que a infraestrutura de armazenamento e acesso aos dados esteja definida e implementada como um *Web Service*, é possível desenvolver um software cliente capaz de acessar e consumir estes dados, observando a compatibilidade sobre os recursos oferecidos pelo *Web Service*. O software cliente foi construído utilizando tecnologias voltadas para Web para ser integrado em um sistema disponível na Internet. Assim como os recursos utilizados no Servidor, todas as tecnologias e ferramentas empregadas para a construção do software Cliente foram soluções *open source*. Dentre os recursos utilizados para a concepção do software Cliente estão:

- JQuery: Os recursos de interação da biblioteca JavaScript JQuery foram aplicados na construção dinâmica dos elementos gráficos da interface, bem como para acesso e manipulação dos elementos que compõem a interface, por exemplo, definição e recuperação dos atributos dos filtros;
- OpenLayers: A biblioteca JavaScript OpenLayers 3 foi utilizada para permitir a exibição, navegação e interação com o mapa contendo a localização das plataformas SIMA, armazenadas no servidor. Os dados com as informações geográficas das plataformas SIMA são armazenados em um arquivo em codificação GeoJSON, no próprio servidor Web;
- D3: O conjunto de bibliotecas JavaScript foi empregada para a construção e implementação das diferentes técnicas de visualizações, valendo-se dos recursos de interatividade, bem como de acesso aos dados, disponíveis pelo *Web Service*, por meio de requisição assíncrona;

Juntamente com as bibliotecas externas utilizadas, a interface Cliente foi construída a partir das tecnologias nativas suportadas pelos navegadores de internet, como: HTML5, CSS e JavaScript. A integração destas tecnologias permitiu o desenvolvimento de um Cliente Web com interface leve e que pode ser executado em diferentes segmentos computacionais.

5.1.3 Técnicas de Visualização de Informação implementadas

A principal função do sistema Cliente é permitir que o usuário analista consiga explorar os dados oferecidos pelo *Web Service* SimaVis utilizando representações visuais e interativas. Na interface Cliente foram implementadas um total de seis categorias de representações visuais: Mapa de localização; gráficos estatísticos; Visualização baseada em pixel; Visualização por Máscara de Bits; Visualização por *Horizon Charts* e Visualização por Coordenadas Paralelas. Exceto o mapa de localização, que foi construído utilizando a biblioteca OpenLayers, todas as outras representações visuais foram implementadas utilizando a biblioteca D3.

O mapa de localização permite selecionar as plataformas de interesse, a partir das quais se deseja obter os dados e explorá-los por meio de representações visuais. A seleção das plataformas no mapa é sincronizada com o filtro de seleção manual das plataformas, como pode ser visto na Figura 38, na qual as plataformas selecionadas no mapa são automaticamente selecionadas na composição do filtro dos dados.

A partir da escolha das plataformas de interesse, é possível selecionar qual o conjunto de atributos que serão utilizados para a composição da análise. Os filtros pré construídos permitem escolher, todos os atributos das plataformas selecionadas ou apenas os atributos em comum. Também é possível a restrição da consulta dos dados a um

Figura 38 – Mapa de localização das plataformas SIMA. A seleção das plataformas é sincronizada com a composição dos filtros de seleção dos dados. Em destaque estão as plataformas selecionadas, tanto pelo mapa, quanto pelos filtros



determinado intervalo de tempo. Caso nenhum intervalo seja definido, serão utilizados todos os dados existentes no conjunto de dados, de acordo com seleção das plataformas e dos atributos.

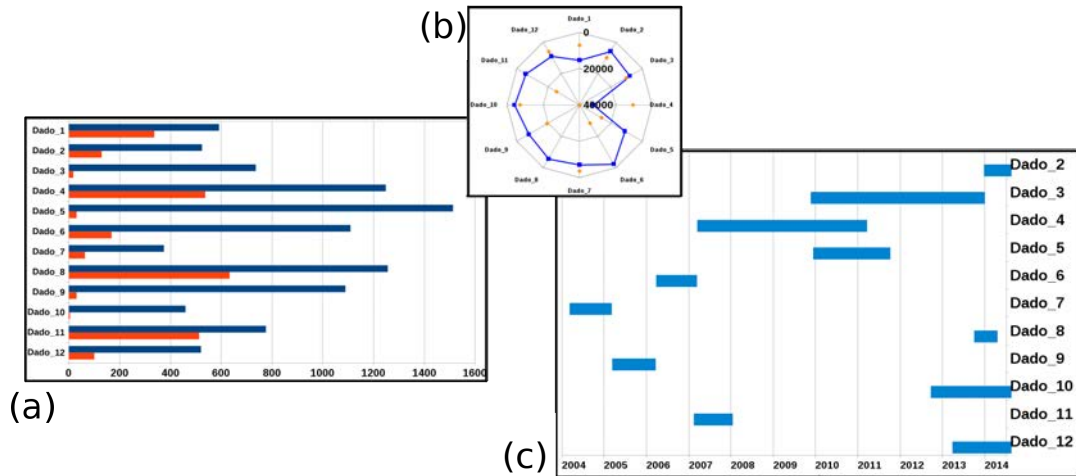
Após a definição do domínio de informação, a requisição é enviada e processada pelo *Web Service* e os dados são devolvidos em formato JSON. Uma vez que os dados tenham sido definidos, diferentes representações visuais podem ser aplicadas sobre estes.

Existem três tipos de gráficos, exibidos na Figura 39 (gerados com dados simulados), que podem ser utilizados para representar inicialmente o comportamento dos dados selecionados e auxiliar a construção das outras visualizações, bem como o refinamento da consulta realizada. O gráfico de intervalo é construído utilizando a data da primeira e da última coleta realizada pelas plataformas selecionadas. Os outros gráficos representam dois atributos quantitativos – relacionados com as falhas nos dados – em formato de barra e radial.

As outras quatro representações visuais foram construídas baseado - ou adaptado - de técnicas conhecidas definidas no escopo de Visualização de Informação que estão detalhadas no Capítulo 3. Todas elas possuem recursos de interatividade com o analista, com os quais ele pode obter detalhes sobre os dados ou registros, por exemplo, utilizando interações com o mouse.

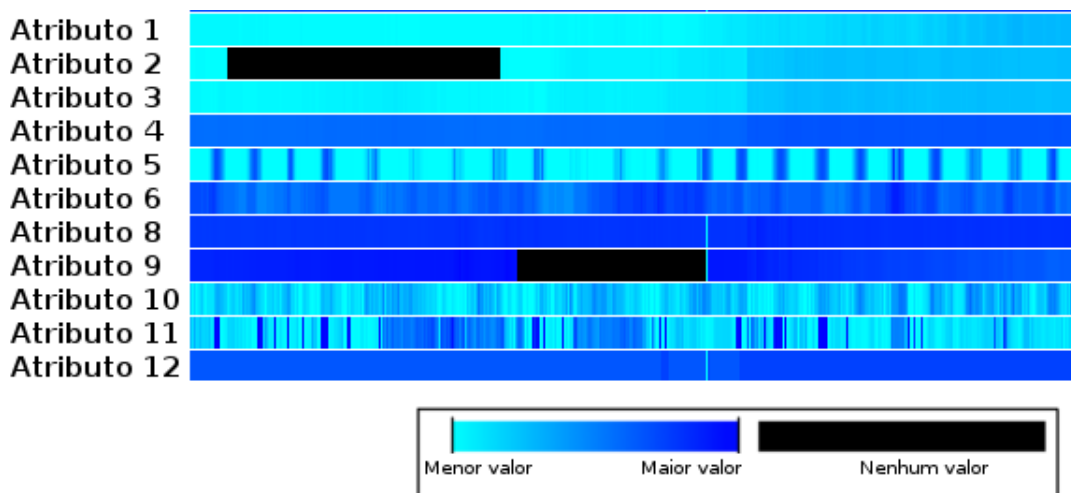
A representação visual baseada em pixel representa a variação de um conjunto de atributos selecionados em uma matriz de dados em que os valores quantitativos dos atributos é exibido de acordo com o mesmo gradiente de cores, mas em escalas específicas para cada atributo, como pode ser visto na Figura 40. Cada linha da matriz identifica um

Figura 39 – Gráficos construídos no SimaVis, gerados a partir de dados simulados: (a) gráfico de barras; (b) gráfico radial e (c) gráficos de intervalos. O gráfico de intervalos considera o primeiro e o último dia de coleta de cada plataforma, enquanto o gráfico radial e de barras representam a proporção, respectivamente, de falhas ocorridas nas coletas e de dias sem coleta de dados



atributo selecionado nos filtros. No eixo das abscissas constam os valores temporais, no qual cada registro de dados selecionados é disposto na linha da matriz, com a dimensão mínima de um pixel, de acordo com o momento de realização da coleta. Nessa representação, o eixo do tempo não é linear, pois os registros são ordenados sequencialmente de acordo com a existência no conjunto de dados, de forma que, se em um determinado intervalo de tempo não existem dados registrados, esta lacuna não será exibida nesta representação.

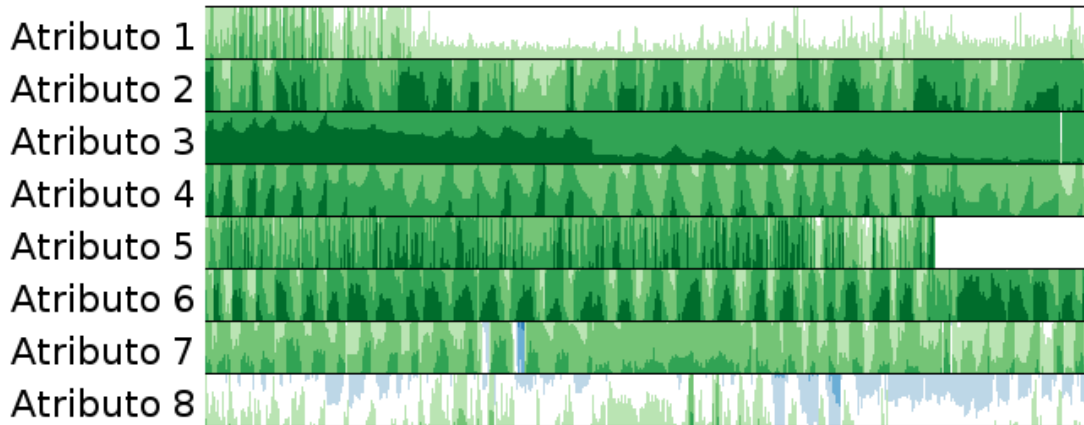
Figura 40 – Visualização baseada em pixel implementada pelo SimaVis, gerada com dados simulados



A técnica de Visualização por *Horizon Charts* foi empregada utilizando o plugin para visualização de séries temporais *Cubism*, desenvolvido com a biblioteca D3. A configuração desta representação visual é semelhante à utilizada na representação baseada em pixel,

como pode ser visto na Figura 41: cada linha representa um atributo identificado pela amplitude da curva e também pelo gradiente de cores e compartilha o mesmo eixo temporal nas abscissas - neste caso de forma linear, ou seja, as lacunas de dados são visíveis.

Figura 41 – Visualização *Horizon Charts* utilizado no SimaVis, gerada com dados simulados



O portal D3 possui um exemplo de implementação da técnica de Coordenadas Paralelas, apresentada e discutida na Seção 3.2, que foi adaptado para compor os recursos do SimaVis. Tal representação permite a reordenação interativa dos eixos paralelos durante a exibição da mesma, de modo que as relações entre os eixos vizinhos possam ser destacadas. Esta representação não é integrada originalmente com o atributo tempo, no entanto, no SimaVis ela foi alterada para representar o componente temporal utilizando um gradiente de cores, como pode ser visto na Figura 42. Desta forma, é possível identificar registros mais novos ou mais antigos por meio da cor, além dos valores ressaltados pelo cruzamento da polilinha (registro) com cada eixo. Essa implementação também permite a seleção de intervalos de atributos em cada eixo, destacando os registros que estão dentro do intervalo selecionado, facilitando o agrupamento e identificação de padrões de variabilidade nos dados.

A representação visual baseada em máscara de bits se originou durante o desenvolvimento deste trabalho, a partir das necessidades observadas para permitir a busca por padrões de falha no conjunto de dados de cada sensor. A implementação foi realizada utilizando a biblioteca D3 para construção e manipulação dos elementos gráficos. Um exemplo desta representação pode ser visto na Figura 43, na qual cada linha representa um dia de dados registrados no banco de dados, cada coluna representa uma hora do dia e cada célula representa pela cor um registro contendo na sequência os sensores que estão funcionando (representados por 1) e os sensores que estão falhando (representados por 0). Mais detalhes sobre a implementação, características e funcionamento desta técnica genérica pode ser obtida na Seção 3.2.1.

Figura 42 – Visualização de Coordenadas Paralelas no SimaVis, gerada com dados simulados. As cores representam o atributo tempo definido de forma relativa no primeiro eixo

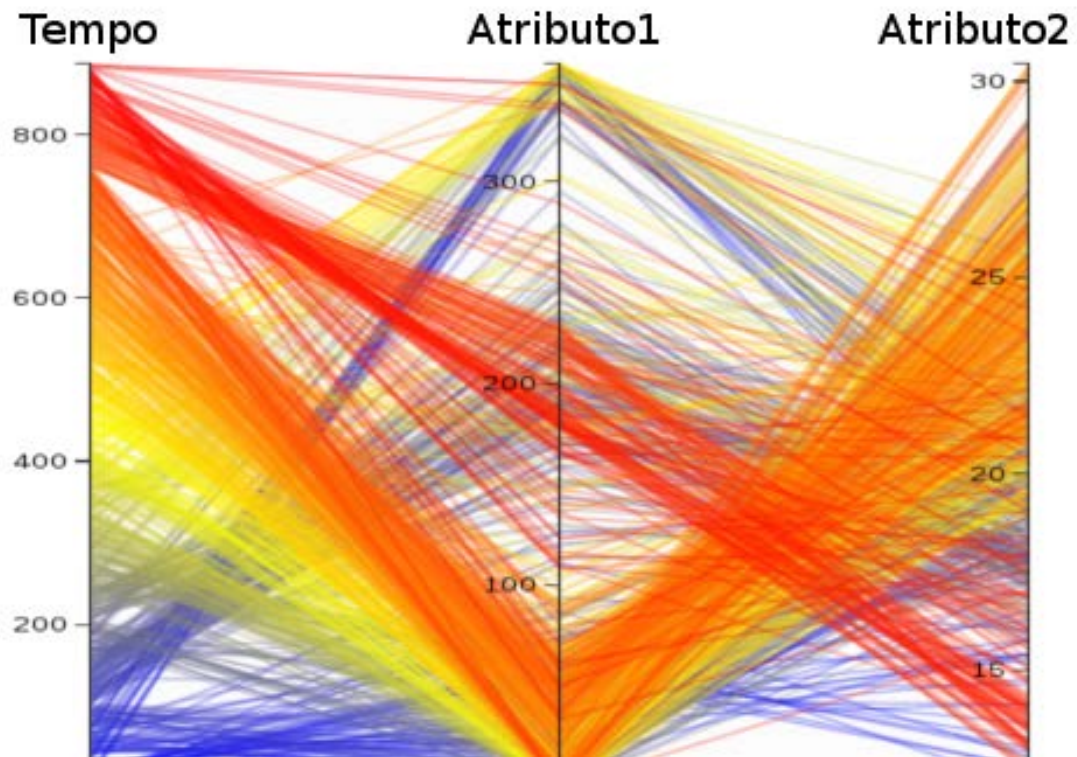
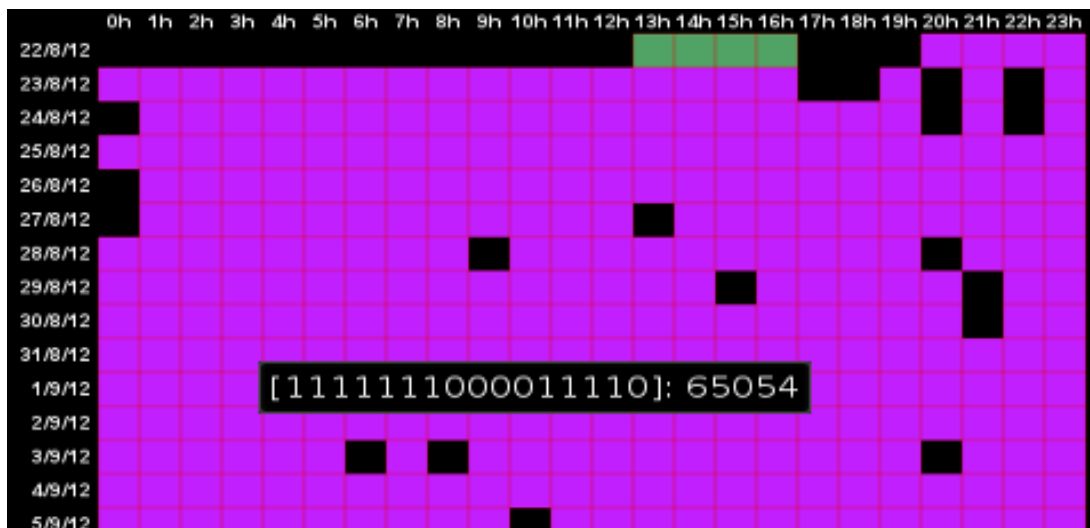


Figura 43 – Visualização por máscara de bit implementada no SimaVis, gerada com dados simulados

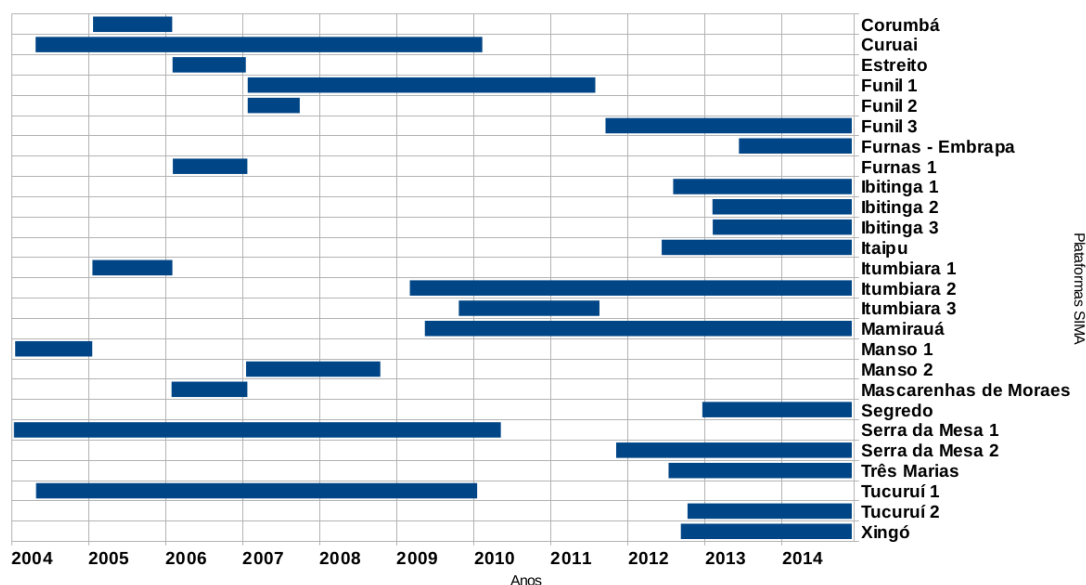


5.2 Tempo de atividade das plataformas SIMA

Um dos primeiros fatores que poderia influenciar a decisão de se utilizar um determinado conjunto de dados está relacionado com a existência de registros em uma determinada localização da plataforma SIMA e em um específico intervalo de tempo. Para extrair este tipo de informação, o gráfico de intervalos pode ser utilizado para exibir o

período de tempo de atividade de cada plataforma, como pode ser visto na Figura 44, que apresenta o intervalo de tempo em que as plataformas estiveram ativas, coletando e transmitindo dados.

Figura 44 – Períodos de atividades das plataformas SIMA



O gráfico apresentado na Figura 44 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 essas plataformas foram estações temporárias de coleta movidas para outros reservatórios, situação provável principalmente no início do projeto.

O gráfico de intervalo permite identificar visualmente as plataformas de acordo com o intervalo de tempo na qual produziram dados. Para extrair essa informação sem a utilização de recursos visuais, a dificuldade de interpretação é acrescida com a quantidade de plataformas analisadas. Na Tabela 4 são apresentadas as mesmas informações presentes no gráfico, divergindo apenas pela estrutura dos dados (Gráfico x Tabela).

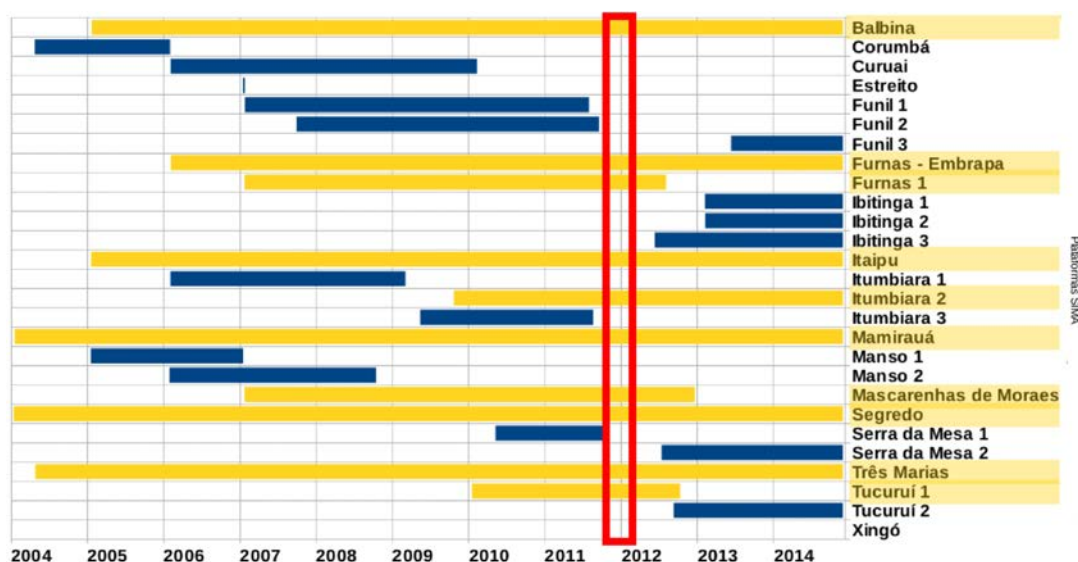
Sem a utilização da representação visual, verificar quais as plataformas possuem dados registrados em uma data específica exige a comparação dos intervalos de cada registro da tabela com a data desejada. Enquanto que, utilizando o gráfico de intervalos, uma inspeção visual sobre o eixo temporal já oferece esta informação de maneira intuitiva (percepção visual), como pode ser visto na Figura 45, na qual uma linha imaginária é traçada em um determinado instante de tempo, destacando as plataformas que estiveram produzindo dados simultaneamente em um determinado instante de tempo.

Vale ressaltar que existe também o custo de processamento necessário para geração da tabela com estes dados, pois, a partir do conjunto total de dados registrados, é preciso extrair e agrupar o primeiro e o último registro para configuração do intervalo, já que

Tabela 4 – Tabela contendo os períodos de atividades de cada plataforma

Nome Plataforma	Primeiro Registro	Último Registro
Balbina	16/08/13	19/01/15
Corumbá	26/01/05	10/02/06
Curuai	25/04/04	11/03/10
Estreito	12/02/06	30/01/07
Funil 1	08/02/07	05/09/11
Funil 2	08/02/07	16/10/07
Funil 3	24/10/11	19/01/15
Furnas - Embrapa	26/07/13	19/01/15
Furnas 1	13/02/06	06/02/07
Ibitinga 1	13/09/12	19/01/15
Ibitinga 2	21/03/13	19/01/15
Ibitinga 3	22/03/13	19/01/15
Itaipu	20/07/12	19/01/15
Itumbiara 1	23/01/05	11/02/06
Itumbiara 2	28/03/09	19/01/15
Itumbiara 3	18/11/09	25/09/11
Mamirauá	08/06/09	19/01/15
Manso 1	18/01/04	22/01/05
Manso 2	31/01/07	06/11/08
Mascarenhas de Moraes	07/02/06	06/02/07
Segredo	31/01/13	19/01/15
Serra da Mesa 1	12/01/04	08/06/10
Serra da Mesa 2	14/12/11	19/01/15
Três Marias	22/08/12	19/01/15
Tucuruí 1	27/04/04	14/02/10
Tucuruí 2	21/11/12	19/01/15
Xingó	20/10/12	19/01/15

Figura 45 – Identificação das plataformas SIMA que estiveram ativas simultaneamente no início de 2012. Em amarelo são destacadas as plataformas que atendem a esta restrição



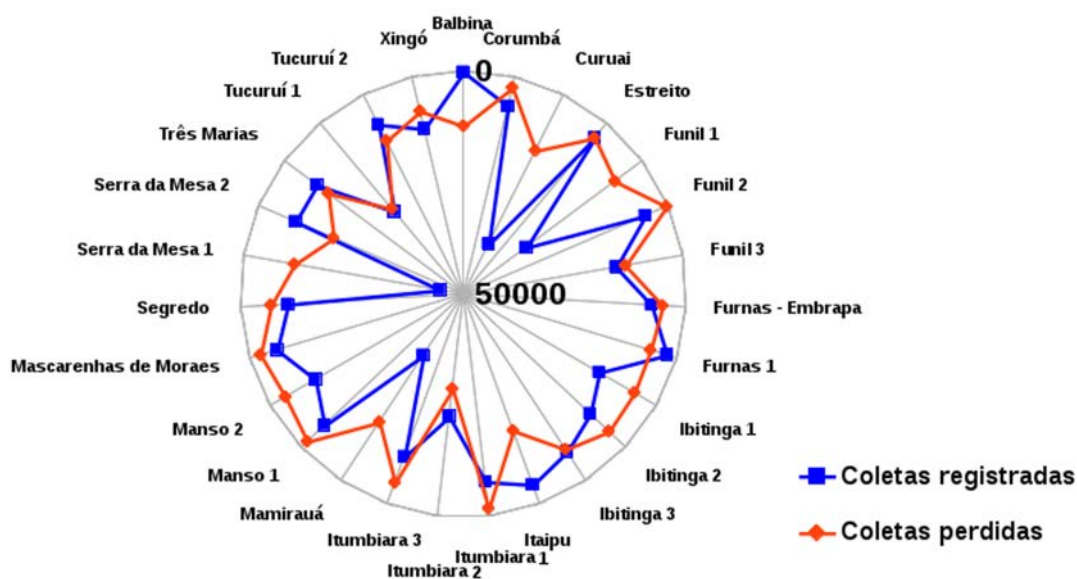
esta informação não está diretamente armazenada no conjunto de dados, ou seja, deve ser obtida a partir dos registros.

5.3 Quantidade de dados faltantes

O gráfico de intervalos considera apenas o primeiro e o último dia de coleta registrado no banco de dados, para a definição do intervalo de atividade das plataformas, mas nessa representação não é possível saber se a plataforma esteve realmente coletando e transmitindo dados plenamente nesse intervalo.

Utilizando um gráfico radial, é possível notar a proporção de dados coletados com sucesso e com falhas, pois são exibidas as quantidades de registros efetivamente armazenados no banco de dados e a quantidade de coletas que não foram registradas, baseada na quantidade de coletas que as plataformas deveriam ter realizado, como pode ser visto na Figura 46, na qual o valor menor (zero) está na borda externa e o valor maior está no centro. A quantidade esperada de coletas que cada plataforma deve realizar pode ser obtida considerando a quantidade de dias em que a plataforma esteve ativa e o fato que todas as plataformas deveriam processar e transmitir 24 coletas diárias – uma coleta a cada hora completa.

Figura 46 – 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.



O percentual de falhas nas coletas está resumido na Tabela 5, na qual a quantidade de coletas com falhas é apresentada em relação à quantidade de coletas que deveriam ter sido realizadas no intervalo de tempo de atividade da plataforma.

Nesta situação, a representação visual utilizando o gráfico radial é um complemento importante para a compreensão dos dados apresentados de forma tabular. Na tabela, é possível identificar a quantidade relativa de falhas existentes a partir do percentual indicado. Já no gráfico, esta relação pode ser observada através da distância entre os

Tabela 5 – Tabela contendo a quantidade relativa de falhas das coletas

Nome Plataforma	%Falha das coletas
Balbina	98,32%
Corumbá	26,43%
Curuai	27,22%
Estreito	50,27%
Funil 1	19,05%
Funil 2	6,88%
Funil 3	46,29%
Furnas - Embrapa	40,38%
Furnas 1	72,18%
Ibitinga 1	27,88%
Ibitinga 2	32,04%
Ibitinga 3	51,83%
Itaipu	79,45%
Itumbiara 1	17,37%
Itumbiara 2	56,14%
Itumbiara 3	30,82%
Mamirauá	31,77%
Manso 1	20,50%
Manso 2	24,87%
Mascarenhas de Moraes	27,79%
Segredo	38,83%
Serra da Mesa 1	20,41%
Serra da Mesa 2	67,10%
Três Marias	57,41%
Tucuruí 1	49,19%
Tucuruí 2	60,73%
Xingó	39,67%

pontos vermelhos (falhas) e os pontos azuis (coletas) de cada eixo, na qual distâncias radiais maiores representam um comportamento com poucas falhas (caso o valor azul seja maior que o vermelho) ou com muitas falhas (caso o valor azul seja menor que o vermelho) e distâncias pequenas (pontos próximos) representam um equilíbrio entre quantidade de falhas e de dados coletados. Como indicado na Figura 47 e confirmado pelo índice percentual exibido na Tabela 5: (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 consideráveis, superiores à 70% do total coletado, como: Furnas 1 (72,18%); Itaipu (79,45%) e Balbina (98,32%). Esse fato foi um indício de que poderiam existir dias completos sem a existência de nenhuma coleta de dados. Para verificar esta possibilidade, foi construído uma representação visual utilizando o gráfico de barras, como pode ser visto na Figura 48.

Dentre as três plataformas que indicaram índices de falhas nas coletas superiores a 70% (Furnas 1, Itaipu e Balbina), apenas Furnas 1 apresenta um baixo índice (menor que 1%) de dias completos sem coletas registradas. Tal fato pode significar que a plataforma SIMA Furnas 1 não está programada para realizar coletas a cada intervalo de 1 hora, mas sim, uma quantidade menor de coletas diárias. Essa justificativa pode ser fortalecida pela 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

Figura 47 – Quantidade de falhas das plataformas SIMA: (a) Serra da Mesa 1 com poucas falhas; (b) Itaipu com muitas falhas; (c) Tucuruí 1 com quantidade similar entre falhas e acertos

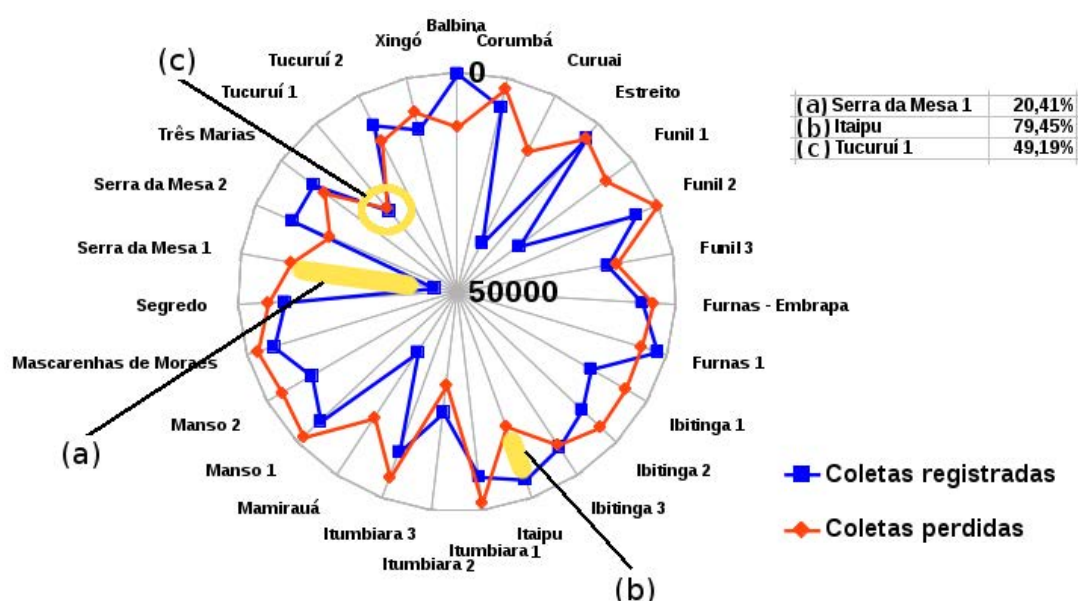
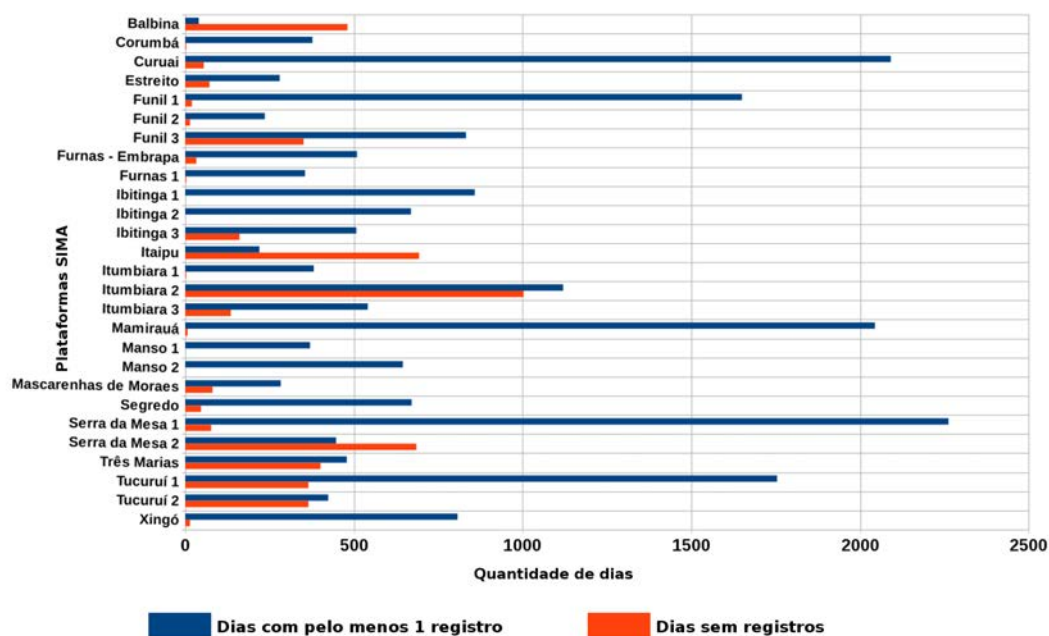


Figura 48 – Coletas realizadas por dia. Em azul a quantidade de dias que existem coletas registradas e em vermelho a quantidade de dias que não possuem nenhum registro.



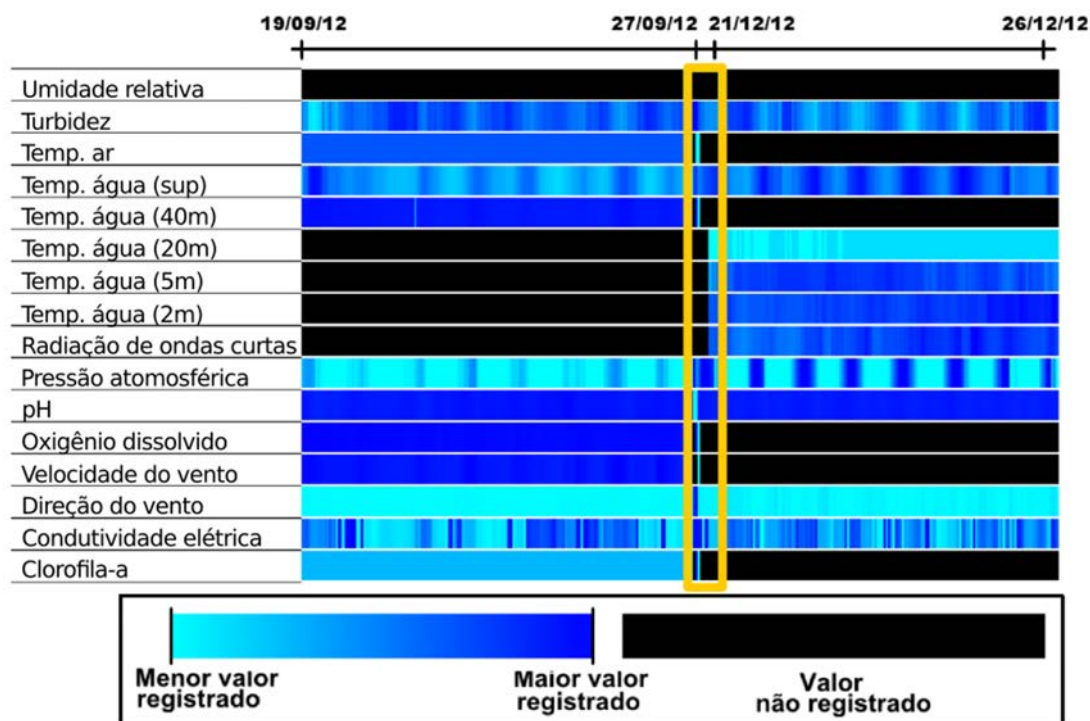
poderia ser parte integrante dos metadados que definem cada plataforma, reforçando a importância da modelagem e organização dos dados.

5.4 Falhas nos atributos das coletas registradas

Os gráficos apresentados nas seções anteriores são úteis para a observação do cenário geral dos dados e extrair informações quantitativas referentes ao sucesso do registro das coletas, sem considerar se os atributos são efetivamente coletados. Para verificar a integridade e o comportamento/variação de cada atributo coletado, outras formas de representação visual podem ser consideradas, de modo que facilitem a percepção e extração de informações. Uma técnica que pode ser aplicada para a representação do comportamento de dados com atributo temporal é a visualização baseada em pixel.

A representação baseada em pixel é útil para observar os valores apresentados por vários atributos simultaneamente de acordo com a escala de tempo. Com esta representação é possível notar a existência de falhas que podem ocorrer simultaneamente em diferentes atributos, como pode ser visto na Figura 49. Para exemplificar a capacidade desta representação visual, foi selecionado um intervalo de dados da plataforma Três Marias, no qual ocorre falha em diferentes grupos de atributos.

Figura 49 – Visualização baseada em pixel de um período de dados da plataforma Três Marias. Em preto encontram-se os valores dos atributos não registrados no banco de dados (valores nulos) e o destaque em amarelo ressalta um intervalo sem a existência de nenhum registro no banco de dados

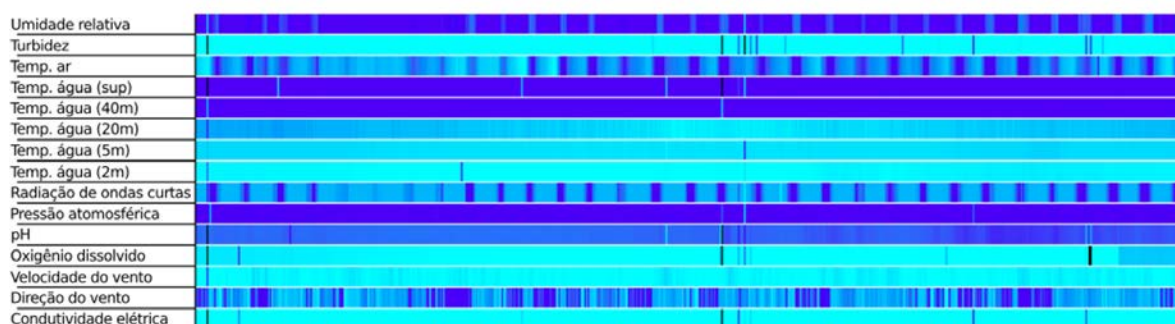


A escala de cores linear e contínua utilizada representa os valores de acordo com sua intensidade. São considerados apenas os registros presentes no conjunto de dados, apresentando-os de maneira ordenada no eixo temporal em uma escala não linear, isto é, não contínua. Na Figura 49, é destacado um intervalo 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.

Esta representação destaca de maneira efetiva a existência de falhas nos atributos, ao longo do período de coleta de dados. Se um atributo não está registrado no banco de dados, possui valor em branco ou nulo, ele é marcado com a cor preta, como pode ser observado na Figura 49. Apesar da frequente existência de falhas nos dados, demonstrada pelas representações visuais exibidas nas Figuras 46 e 48, é possível selecionar intervalos de dados com falhas mínimas. Na Figura 50 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.

Figura 50 – Representação visual baseada em pixel de um intervalo de dados de Funil 1 no qual poucos registros possuem falhas (marcados em preto)

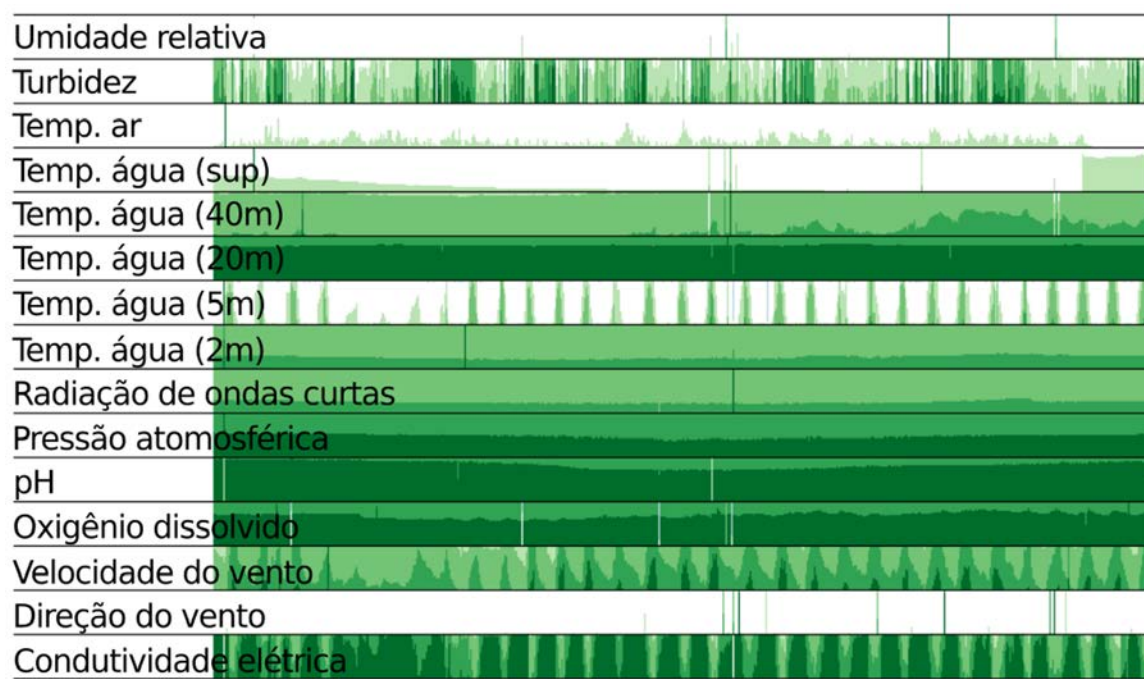


A plataforma Funil 1 possui um dos mais baixos índices de falhas, tanto considerando as falhas em coletas (19,05%) como dias completos sem coleta (1,2%). Além disso, explorando os dados utilizando a representação visual baseada em pixel, é possível notar que também existem poucas falhas nos atributos coletados - pouca existência de lacunas. Essa técnica pode indicar aspectos de qualidade dos dados coletados e armazenados para qualquer uma das plataformas.

Embora a representação baseada em pixel possa auxiliar na detecção isolada de atributos com falhas, ela não permite identificar falhas no registro completo, ou seja, usando esta representação não é possível saber se todas as 24 coletas diárias prevista foram efetivamente registradas ou se houveram períodos em que nenhuma coleta fora realizada, necessitando a integração com outra forma de representação visual, como aquela por gráfico radial ou gráfico de intervalo. Para extrair este tipo de informação a partir de uma única representação visual, a técnica de *Horizon Charts* pode ser utilizada, pois exhibe os valores dos atributos em uma escala de tempo linear. Na Figura 51 é apresentado

o resultado da representação por *Horizon Chart* do mesmo intervalo de dados utilizado na representação baseada em pixel da Figura 50.

Figura 51 – Representação visual, utilizando *Horizon Charts*, para um intervalo de dados da plataforma Funil 1 no qual poucos registros possuem falhas.



Atenção especial deve ser direcionada à interpretação das representações visuais de dados brutos, sem a realização de um processamento prévio. Tomando como exemplo o mesmo intervalo de dados da plataforma de Funil 1, representados nas Figuras 50 e 51, o analista poderia interpretar que existem falhas nos dados de turbidez, radiações de ondas curtas, oxigênio dissolvido, velocidade do vento e condutividade elétrica. Na representação por *Horizon Charts* tal suspeita seria baseada na existência de espaços em branco; e na representação baseada em pixel, pela invariabilidade dos dados. No entanto, tal fato ocorre em razão da existência de valores *outliers* nestes dados, como pode ser visto na Tabela 6 que apresenta algumas estatísticas básicas sobre os dados do intervalo selecionado da plataforma de Funil 1.

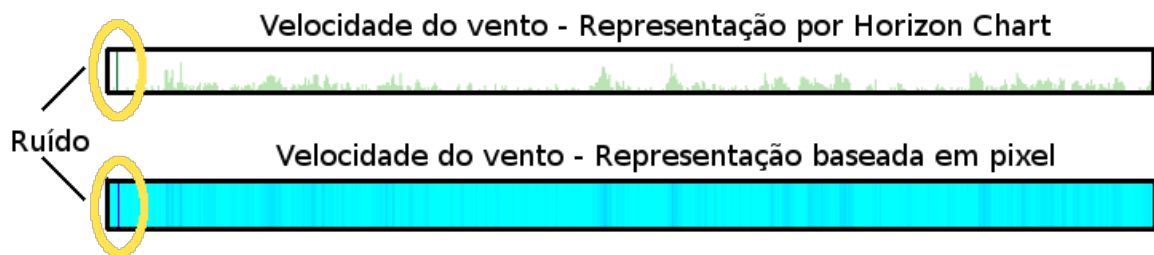
Observando as estatísticas descritivas desses dados, é possível notar uma variação muito alta entre os valores de mínimo e máximo nas variáveis, destacados na Tabela 6, encontrados nos atributos citados que podem indicar a existência de dados corrompidos (ruídos). Estes mesmos dados corrompidos são os que interferem na formação das representações visuais, tanto baseada em pixel como *Horizon Charts*, como pode ser visto na Figura 52 que apresenta os dados do atributo Velocidade do Vento utilizando as duas representações visuais, destacando o valor corrompido que interfere na visualização.

A eliminação do ruído pode ser realizada com a restrição do filtro de valores aplicado sobre a consulta dos dados. Para eliminar o maior valor registrado para o

Tabela 6 – Estatísticas descritivas do conjunto de dados selecionados da plataforma Funil 1. Destacados estão os dados que identificam a presença de outliers (Variação discrepante em relação às medidas de tendência central)

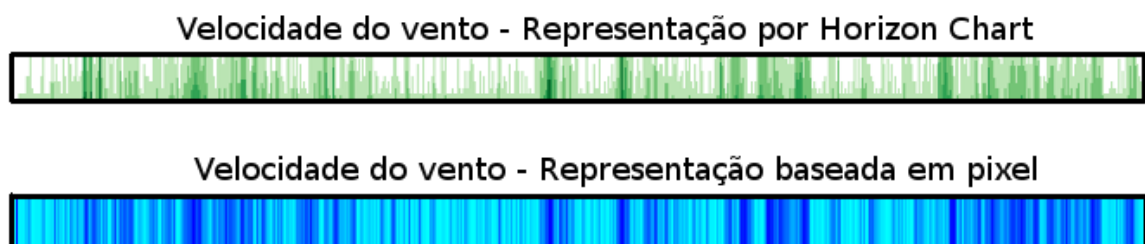
	Mínimo	Máximo	Variação	Média	Mediana	Variância	Desvio Padrão
Condutividade (mS/cm)	1,00	9935,00	9934,00	115,08	65,00	313635,93	678,53
Direção do vento (oNV)	0,00	358,59	358,59	152,91	122,34	8765,75	93,90
Velocidade do vento (m/s)	0,00	44,12	44,12	1,56	1,18	4,38	2,67
Óxigênio dissolvido (mg/l)	0,00	46,09	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 (°C)	20,40	37,09	16,69	21,04	21,05	0,66	1,02
Temp. água a 9.17m (°C)	18,46	36,96	18,50	20,78	20,66	0,51	0,95
Temp. água a 19.17m (°C)	19,75	23,51	3,76	20,48	20,53	0,09	0,32
Temp. água a 39.17m (°C)	5,26	20,53	15,27	19,72	19,75	1,10	1,05
Temp. água a 4.32m (°C)	-5,00	22,94	27,94	21,04	21,22	5,08	2,26
Temp. ar (°C)	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

Figura 52 – Presença de ruído nos valores dos atributos da velocidade do vento, identificáveis nas representações baseada em pixel e *Horizon Charts*



atributo Velocidade do Vento que está interferindo na construção da representação visual, identificado pelas estatísticas descritivas (Tabela 6) como 44,12, é necessário a aplicação de um filtro para restringir a consulta a registros com valores de Velocidade do Vento menores que o limiar desejado. Após o refinamento da consulta, o resultado obtido com a aplicação do filtro para tal atributo pode ser visto na Figura 53.

Figura 53 – Representações visuais baseada em pixel e *Horizon Charts* do atributo Velocidade do Vento da plataforma Funil 1 após a remoção do ruído



O processo de refinamento da consulta pode ser efetuado quantas vezes for necessário. Desta forma, o analista pode reconfigurar a seleção do conjunto de dados e adequar às suas necessidades dentro do processo de exploração e análise dos dados.

A existência de falhas nos atributos coletados e armazenados nos registros do banco de dados permite explorar uma nova possibilidade relacionada com a observação de padrões de falhas de grupos de atributos, como aquele encontrado na Figura 49, representando um intervalo de dados da plataforma Três Marias, no qual ocorre a falha em dois grupos de atributos em momentos distintos.

A visualização por máscara de bits permite explorar a possibilidade de identificação de padrões de comportamento de falhas entre conjuntos de atributos. 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, como detalhado na Seção 3.2.1. Nas Figuras 54 e 55 pode ser visto a representação por meio de máscara de bits do mesmo intervalo de dados utilizado na representação baseada em pixel exibida na Figura 49, referente a um intervalo de dados da plataforma Três Maria, no qual existe falhas em alguns atributos.

Figura 54 – Visualização por máscara de bits do início do intervalo de dados da plataforma Três Marias. O padrão de cores predominante (1111111000011110) indica erro em cinco sensores

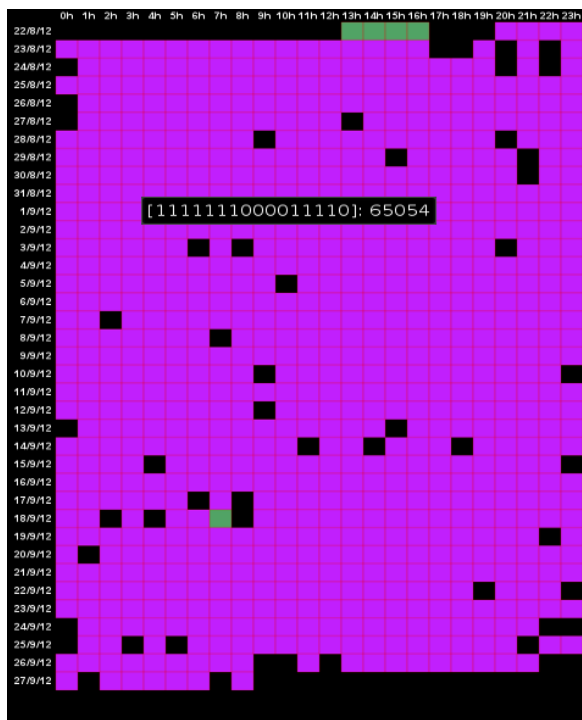
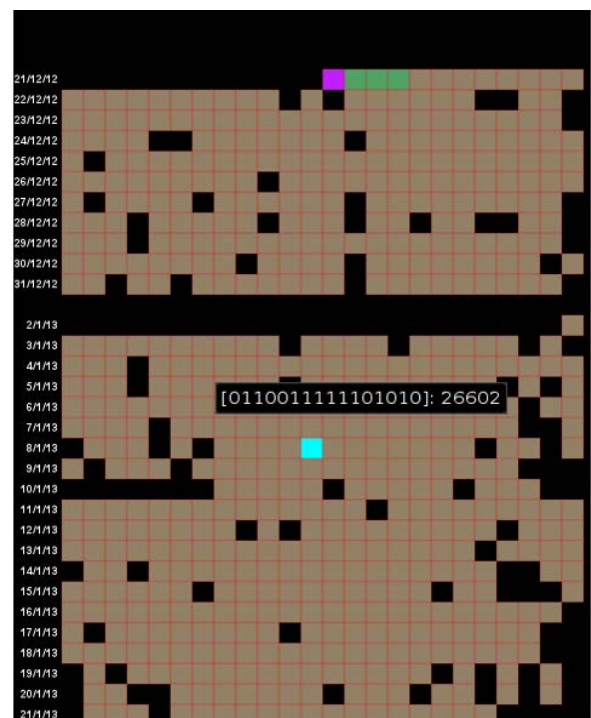
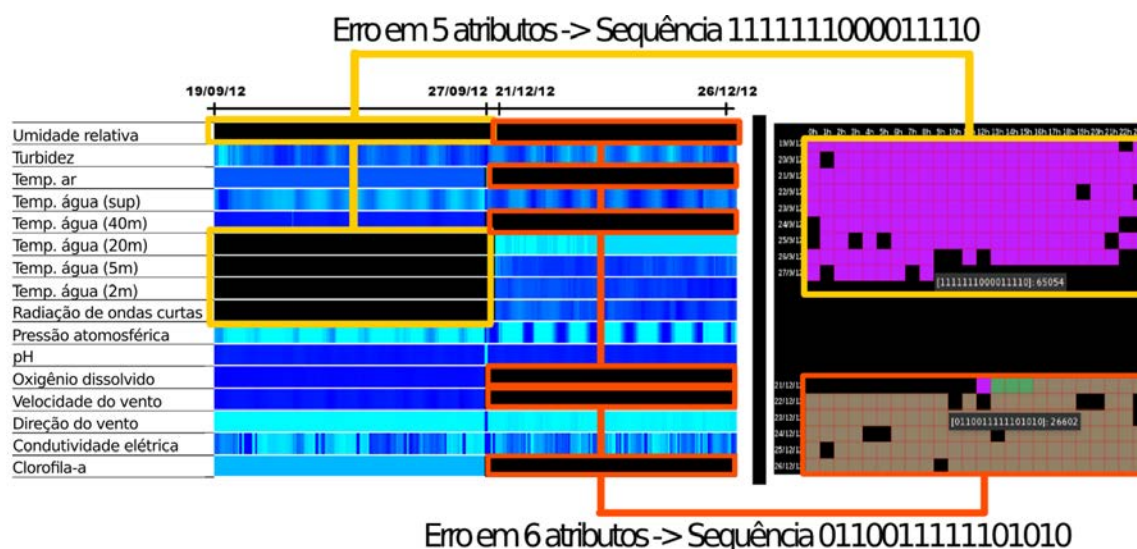


Figura 55 – Visualização por máscara de bits do fim do intervalo de dados da plataforma Três Marias. O padrão de cores predominante (0110011111101010) indica erro em outros seis sensores



Observando ambas as representações visuais simultaneamente, baseada em pixel e por máscara de bits, é possível notar a relação entre a sequência de bits e os atributos que falhando concomitantemente, como pode ser visto na Figura 56. Nesta Figura, estão destacados os cinco atributos com falhas, inicialmente no primeiro intervalo de tempo, e os outros seis atributos que falham no mesmo período de tempo, alterando a sequência de bits e o padrão de cores que identifica o conjunto de atributos com falhas.

Figura 56 – 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 5 sensores e 0110011111101010 falha em outros 6 sensores



Na representação por máscara de bits, a sequência de apresentação dos atributos é importante, pois é esta mesma sequência que será utilizada para conversão para o espaço de cores e definição da cor que representa o conjunto de falha em questão.

5.5 Características de plataformas SIMA fundeadas no mesmo reservatório

Observando o mapa de localização das plataformas e realizando operações de zoom para detalhamento da visão, é possível identificar cenários no qual existem mais de uma plataforma SIMA fundeada no mesmo reservatório. Alguns reservatórios possuem até três plataformas, como é o caso de Funil, Ibitinga e Itumbiara, como pode ser visto na Tabela 7. Existe um total de dezessete plataformas SIMA que compartilham um reservatório com pelo menos mais uma outra plataforma.

A existência de plataformas SIMA espacialmente próximas, oferece a possibilidade de analisar relações entre os comportamentos dos atributos coletados por ambas as plataformas. Dentre os três tipos de erros -falha do sensor, problema na transmissão dos

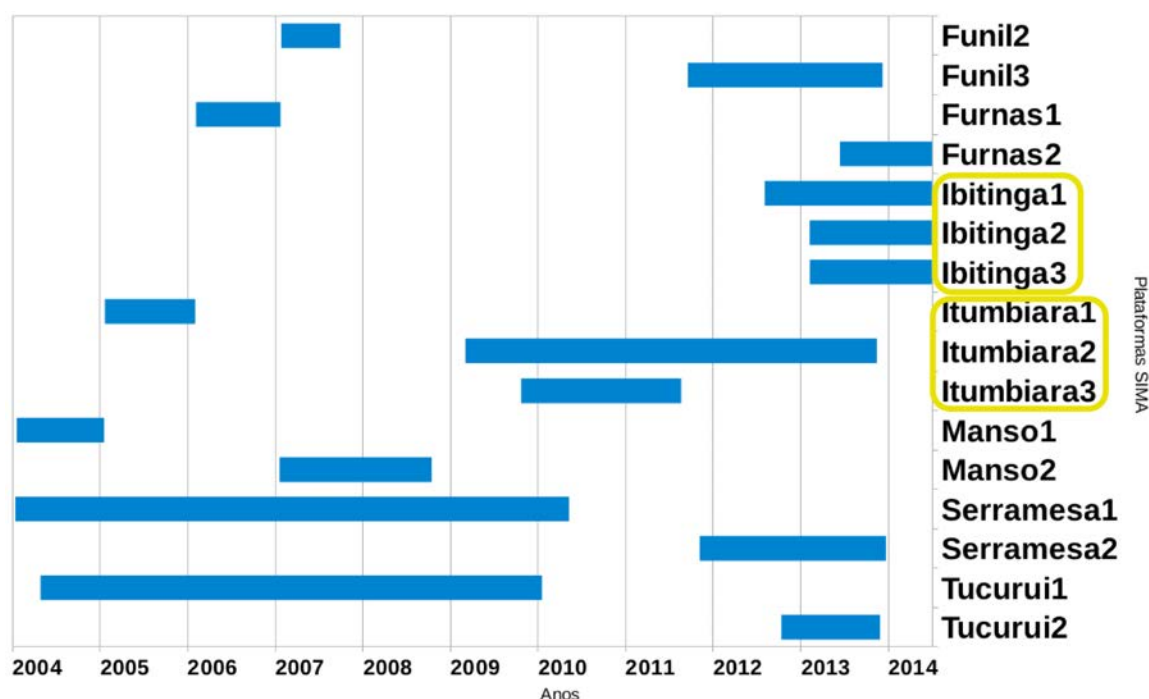
Tabela 7 – Reservatórios com mais de uma plataforma SIMA

Reservatório	Quantidade de Plataformas SIMA
Funil	3
Ibitinga	3
Itumbiara	3
Furnas	2
Manso	2
Serramesa	2
Tucuruí	2
Total	17

dados ou erro na conversão do sinal- que podem afetar o resultado da coleta de dados pelas plataformas SIMA, como discutido na Seção 4.1.1, as falhas ocasionadas por problema na transmissão dos dados coletados, via enlace de satélite, pode ser identificada a partir da observação do comportamento de plataformas próximas (mesma visada do satélite).

Para realizar esse tipo de análise, é necessário primeiramente selecionar apenas as plataformas que compartilham o mesmo reservatório e que possuem tempo de atividade em períodos simultâneos. Esta seleção pode ser efetuada por meio da observação do gráfico de intervalos, exibido na Figura 57.

Figura 57 – Intervalos de tempo ativos das plataformas que compartilham reservatórios. Em destaque estão aquelas plataformas do mesmo reservatório que produzem dados em um mesmo intervalo de tempo



Por meio do gráfico de intervalos apresentado na Figura 57, estão destacados dois

reservatórios possuem plataformas com realizações de coletas em um mesmo intervalo: Itumbiara e Ibitinga. Ibitinga possui três plataformas atualmente ativas e coletando dados no mesmo intervalo de tempo. Itumbiara possui também três plataformas, porém apenas duas com período de coleta de dados com intersecção, sendo que todas elas já foram desativadas.

Outra informação importante é o tipo de atributo que cada plataforma é capaz de coletar, já que se trata de fontes múltiplas de dados. A Tabela 8 sumariza os tipos de atributos coletados por cada plataforma fundeada no mesmo reservatório. Essa tabela foi construída utilizando os metadados definidos na nova modelagem de dados do sistema SimaVis, discutida e apresentada na Seção 5.1.1.3. Cada atributo marcado com X indica que aquela plataforma realiza a coleta de um atributo específico, sendo que a marcação na cor vermelha indica que tal atributo não é coletado por uma das plataformas do mesmo reservatório.

Tabela 8 – Atributos coletados e respectivas plataformas SIMA. Marcados com X indica a presença na plataforma. Aqueles destacados em vermelho indicam a inexistência em uma das plataformas deste reservatório

	Funil1	Funil2	Funil3	Furnas1	Furnas2	Ibitinga1	Ibitinga2	Ibitinga3	Itumbiara1	Itumbiara2	Itumbiara3	Manso1	Manso2	Serramesa1	Serramesa2	Tucuruí1	Tucuruí2	
Chl (ug/l)	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		15
CO2 high (ppm)																	X	1
CO2 low (ppm)																	X	1
Cond. (uS/cm)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Data e horário GMT	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Dir. vento (oNV)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Int. vento (m/s)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
NH4+ (mg/l)	X	X		X					X	X	X	X	X	X		X		10
NO3- (mg/l)	X	X		X					X	X	X	X	X	X		X		10
OD (mg/l)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
pH	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Press. atm. (hPa)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Rad. incid. (W/m2)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Rad. refl. (W/m2)	X			X					X	X	X	X	X	X		X		9
Temp. água (oC)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X		15
Temp. da sonda (oC)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Temp. do ar (oC)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Turb. (NTU)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Umid. rel. ar (%)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Vel. mer. corrente (cm/s)	X			X					X	X	X	X		X				7
Vel. mer. vento (m/s)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Vel. zonal corrente (cm/s)	X			X					X	X	X	X		X				7
Vel. zonal vento (m/s)	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
	21	9	16	21	16	16	16	16	21	21	21	21	19	21	15	19	16	305

Na Tabela 8 é possível notar que apenas dois reservatórios contam com plataformas que coletam exatamente o mesmo conjunto de atributos: Ibitinga e Itumbiara. Ainda é possível observar que Itumbiara se destaca por ser a plataforma que realiza a coleta de um conjunto maior de atributos, fato que pode estar relacionado com o tipo de sensor aplicado sobre estas plataformas.

Também é possível ver na Tabela 8, que apenas a plataforma de Tucuruí 2 é capaz de coletar os atributos CO₂ high e CO₂ low. Essa observação ressalta a necessidade de disponibilização de metadados acerca dos atributos coletados e sensores utilizados, pois tal informação pode contribuir diretamente para a explicação de questões relacionados ao processo de obtenção dos dados.

Apesar da existência de registros de plataformas com identificadores diferentes no mesmo reservatório, não é possível afirmar que seja realmente de outra plataforma – a menos que exista dados simultâneos de ambas. O motivo da existência de um novo registro de plataforma pode ser em razão de:

- Alteração do local de fundeio da plataforma;
- Atualização de componentes eletrônicos da plataforma, como a inserção ou remoção de sensores;
- Desativação para manutenção da plataforma;
- Desativação para recalibração dos sensores.

Tais possibilidades são indicações levantadas a partir da observação do comportamento dos dados. No entanto, a existência de metadados que descrevam tais processos representa uma informação importante, que a priori, é inexistente no banco de dados.

Para tentar encontrar alguma informação sobre a manipulação/manutenção das plataformas SIMA, os gráficos de falhas apresentados nas Figura 58 e 59, bem como o gráfico de intervalo, exibido na Figura 57, poderiam ser utilizados para extrair informação sobre a qualidade dos dados que justificasse ou indicasse alguma ação realizada sobre as plataformas.

Tomando o reservatório de Tucuruí como exemplo, o qual possui dois registros de plataformas SIMA. O gráfico de intervalos (Figura 57) indica que essa plataforma foi desativada no início de 2010 e outra plataforma nomeada Tucurui2 iniciou a coleta de dados no final de 2012. A plataforma Tucurui2 possui algumas diferenças em relação à antiga plataforma Tucurui1 quanto ao conjunto de atributos coletados, como pode ser visto na Tabela 8: dois novos atributos coletados (CO₂ low e CO₂ high) e quatro atributos removidos (NH₄⁺, NO₃⁻, Radiação Refletida e temperatura da água).

Figura 58 – Horas de atividade das plataformas

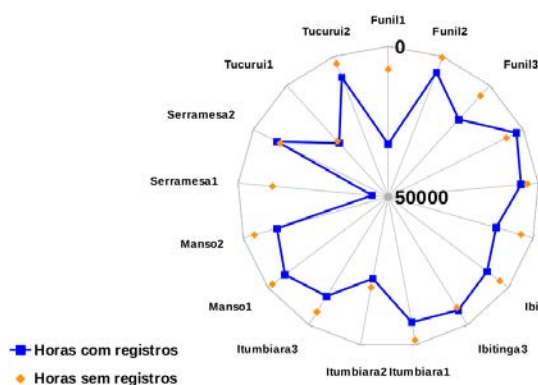
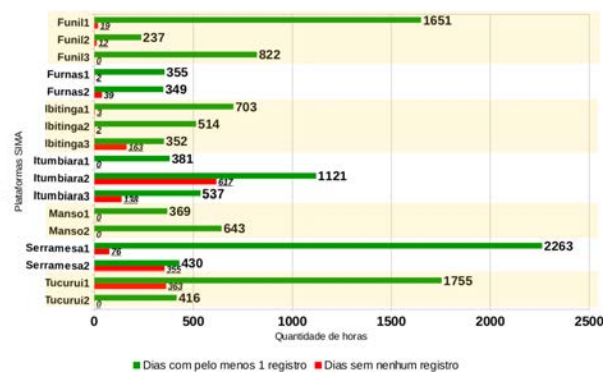


Figura 59 – Dias ativos das plataformas

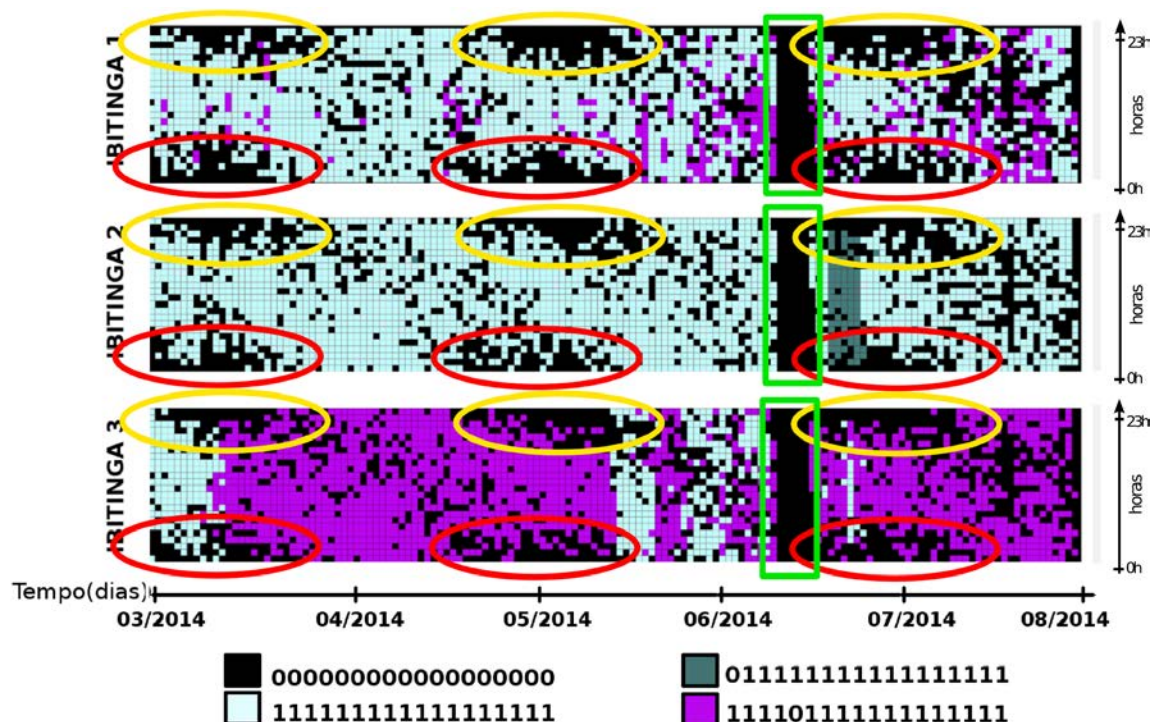


A interrupção da coleta de quatro atributos pode estar relacionado com a degradação dos sensores responsáveis por registrar estas variáveis. Esse indicativo pode ser reforçado observando a grande quantidade de falhas da estação de Tucuruí1 apontadas pelos gráficos apresentados nas Figuras 58 e 59, no qual cerca de 50% das coletas falharam e em razão de haver mais de 20% dos dias (363 falhas de um total de 1755 dias) sem a realização de nenhuma coleta. Já a nova plataforma Tucuruí2, a quantidade de falhas é significativamente menor e com a inexistência de dias sem a realização de coleta de dados.

As três plataformas fundeadas no reservatório de Ibitinga, possuem períodos de tempo com coletas de dados simultâneas, como pode ser visto no gráfico de intervalos da Figura 57, indicando a existência de três plataformas diferentes. As três plataformas possuem a capacidade de coletar o mesmo conjunto composto por 16 atributos diferentes, como indicado na Tabela 8, e ainda temperatura da água em 3 níveis, totalizando 18 variáveis. Exceto Ibitinga 3, todas as outras plataformas possuem um baixo nível de falhas, tanto referentes às coletas quanto acúmulo diário, como mostram os gráficos exibidos nas Figuras 58 e 59.

Observando a representação visual por máscara de bits, que incorpora um intervalo de dados das três plataformas em Ibitinga, apresentada na Figura 60, 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 60, para facilitar a leitura e interpretação, são aplicadas algumas marcações de comportamentos que podem ser destacadas.

Figura 60 – 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 60 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 a 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. Detalhes sobre os Satélites utilizados para a transmissão dos dados das plataformas SIMA podem ser encontrados na Seção 4.1.1.

A existência de períodos com falhas nos dados é comum em outros intervalos do conjunto de dados, como apontado por Valerio et al. (2011), 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 mal funcionamento.

Ainda na Figura 60, 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.

A similaridade entre as plataformas fundeadas no mesmo reservatório de Ibitinga se estende também aos valores coletados pelas mesmas. Na Tabela 9 são apresentadas algumas estatísticas descritivas que auxiliam a observação de variações próximas entre os valores dos dados coletados.

Tabela 9 – Estatísticas descritivas para o mesmo intervalo de dados das plataformas SIMA do reservatório de Ibitinga

	Ibitinga 1				Ibitinga 2				Ibitinga 3			
Variável	Média	Variância	Mínimo	Máximo	Média	Variância	Mínimo	Máximo	Média	Variância	Mínimo	Máximo
Chl(ug/l)	24,048	261,571	1,17	120,55	28,389	1366,399	0,78	370,25	24,869	591,733	2,74	228,57
Cond.(uS/cm)	185,34	2012,01	19	275	256,47	253,19	216	299	225,4	177,75	192	260
Dir.vento(oNV)	176,81	953,49	0	355,78	182,25	6940,75	0	358,59	154,82	5187,97	0	357,19
Int.vento(m/s)	19,615	26,886	0	61,96	1,5733	1,8749	0	10,2	2,481	2,9527	0	9,02
pH	7,6969	0,0956	7,1	9,27	7,6799	0,1575	7,19	9,45	7,8856	0,3696	7,08	9,71
Press.atm.(hPa)	970,13	9,88	962,12	987,29	970,9	9,86	962,12	979,45	969,65	9,31	962,12	978,43
Rad.incid.(W/m2)	193,52	69407,59	0	1029,41	194,88	73330,01	0	1194,12	202,11	78933,9	0	1164,71
Temp.dasonda(oC)	23,718	7,39	3,02	30,32	23,257	6,708	18,87	29,59	24,063	7,436	5,23	30,91
Temp.doar(oC)	21,219	21,963	5,8	34,97	20,323	22,698	10	33,47	21,466	19,786	10,25	32,46
Turb.(NTU)	707,03	29045,03	159,49	1000	36,089	2074,555	0,98	650,68	345,52	71519,07	2,94	999,02
Umid.rel.ar(%)	69,919	283,91	19,61	96,08	38,562	426,895	14,51	97,65	69,817	260,803	19,22	99,22
Vel.mer.vento(m/s)	17,917	41,732	-41,83	29,62	0,795	2,2402	-5,08	6,92	1,3199	4,6315	-6,59	7,82
Vel.zonalvento(m/s)	-0,125	48,898	-61,79	34,78	0,0847	1,4696	-5,25	10,12	-0,7179	2,2178	-5,63	7,53

Não é foco deste trabalho apresentar discussões sobre o mérito do significado dos valores coletados para cada atributo. No entanto, as estatísticas descritivas permitem observar um comportamento muito próximo dos atributos coletados, demonstrando que as plataformas possuem correspondências na variabilidade espacial dos dados além de semelhança que pode ser notada sobre o comportamento de falhas dos sensores.

5.6 Métodos estatísticos complementados por representações visuais

As falhas nos dados podem ocorrer frequentemente em diferentes intervalos da série temporal dos dados das plataformas SIMA. Os dados corrompidos podem fornecer informações sobre a característica e comportamento de cada plataforma, no que diz respeito à realização das coletas. Os resultados obtidos nas seções anteriores demonstram que é possível ressaltar relações entre características de coletas de dados por diferentes plataformas, como aquelas fundeadas no mesmo reservatório. As representações visuais podem ser utilizadas também para realçar possíveis relações entre os atributos coletados por uma mesma plataforma. Para tanto, técnicas estatísticas podem ser aplicadas juntamente com técnicas de Visualização de Informação para facilitar a interpretação dos resultados.

Com o objetivo de destacar possíveis relações entre os atributos coletados e as possibilidades de complementação das análises estatísticas com representações visuais, é necessário que o conjunto de dados utilizado não possua falhas ou dados corrompidos, pois estes teriam influência direta sobre a variabilidade dos dados. Para isto, foi realizada uma exploração prévia sobre o conjunto de dados, visando selecionar um subconjunto no qual não existissem dados espúrios.

Após análise exploratória da base de dados da plataforma SIMA do reservatório de Itaipu, foi possível selecionar um período em que todas as variáveis foram coletadas, compreendido entre 20 de julho de 2012 à 30 de agosto de 2012, contendo 887 registros.

Inicialmente, para verificar o comportamento geral das diferentes variáveis, algumas métricas de estatísticas descritivas foram calculadas sobre o conjunto de dados. A Tabela 10 exhibe o conjunto de dados descrito pelas métricas estatísticas.

Tabela 10 – Métricas da estatística descritiva sobre o conjunto de dados da plataforma Itaipu

Variable	N	Mean	StDev	Variance	Minimum	Median	Maximum	Range
condutividade	887	0,043337	0,002369	0,000006	0,040000	0,045000	0,050000	0,010000
O2_dissolvido	887	7,2191	2,7411	7,5137	1,0800	8,8100	10,3200	9,2400
pH	887	7,3497	0,1646	0,0271	7,0000	7,3600	8,1800	1,1800
agua_temp_0m	887	21,231	1,057	1,118	19,220	21,660	23,520	4,300
agua_temp_3.5m	887	21,031	1,056	1,116	19,240	21,440	22,470	3,230
agua_temp_5m	887	20,992	1,065	1,134	19,110	21,310	22,340	3,230
agua_temp_10m	887	20,864	1,100	1,211	19,110	21,180	22,210	3,100
agua_temp_30m	887	20,221	1,163	1,354	18,720	19,880	22,080	3,360
turbidez	887	8,4992	2,3959	5,7405	5,8700	7,8300	17,6100	11,7400
vento_dir	887	107,89	103,07	10623,59	0,00	57,66	358,59	358,59
vento_vel	887	3,4463	2,1852	4,7753	0,0000	3,1400	9,8000	9,8000
pressao_atm	887	990,24	3,05	9,29	983,53	989,65	998,82	15,29
radiacao_oc	887	165,50	242,03	58577,95	0,00	5,88	852,94	852,94
ar_temp	887	21,074	3,824	14,623	12,010	21,170	30,450	18,440
humidade_relativa	887	70,860	16,329	266,644	25,100	70,590	99,220	74,120

As métricas de estatísticas descritivas demonstram, conforme esperado, uma grande similaridade entre os dados que representam o mesmo fenômeno relacionado à tempe-

Figura 61 – Série temporal dos dados de temperatura da água. A proximidade das curvas indica possível alta correlação entre os dados

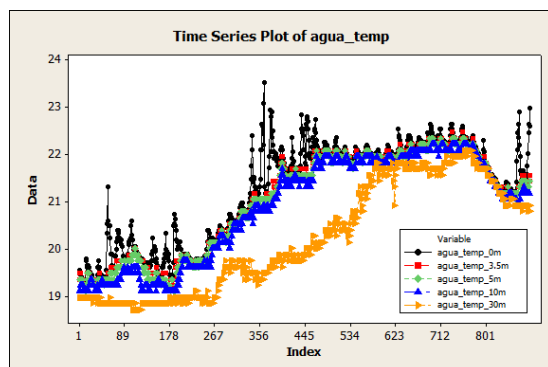
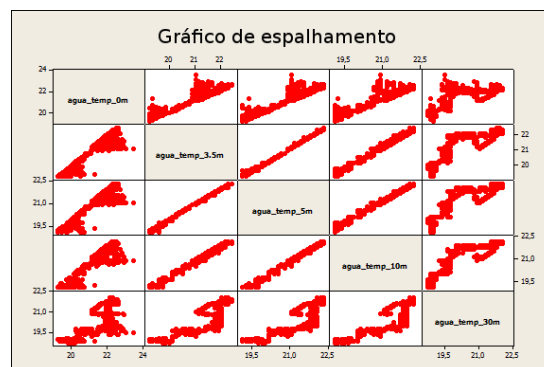


Figura 62 – Scatter Plot dos dados de temperatura da água. Linearidade do espalhamento indica possível alta correlação entre os dados



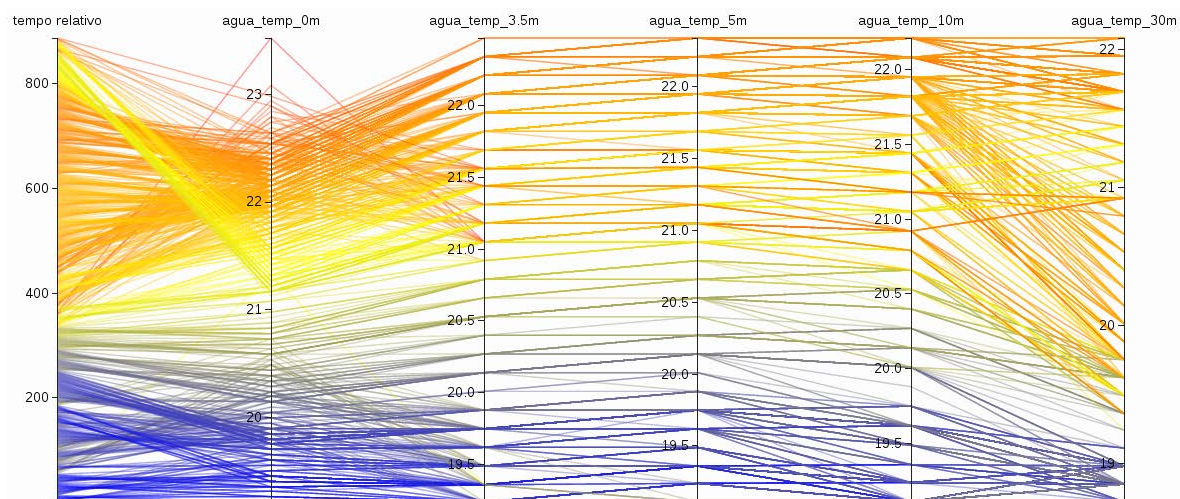
ratura da água em diferentes profundidades, referentes aos atributos: `agua_temp_0m`, `agua_temp_3.5m`, `agua_temp_5m`, `agua_temp_10m`, `agua_temp_30m`. Tal evidência pode ser observada também por meio de algumas representações visuais clássicas, como o gráfico de série temporal, Figura 61, e a matriz de Scatter Plot, Figura 62, relacionados a estes dados.

Além dos métodos gráficos convencionais de representação, exibidos nas Figuras 61 e 62, outras técnicas de Visualização de Informação podem facilitar a observação de relação entre a variabilidade dos dados, como a técnica de Coordenadas Paralelas. A representação visual por Coordenadas Paralelas permite exibir simultaneamente várias dimensões do conjunto de dados, ou seja, vários atributos. A relação direta entre os dados apresentados segundo esta técnica, pode ser ressaltada por meio de linhas concentradas paralelamente que conectam cada eixo, como pode ser visto na Figura 63, para os dados de temperatura da água em diferentes profundidades.

A representação visual por Coordenadas Paralelas na Figura 63 complementa as informações de existência de relação entre os dados de temperatura da água em diferentes níveis de profundidade, indicados também nos gráficos das Figuras 61 e 62 e as estatísticas descritivas apresentadas na Figura 10. Esta situação apresenta um cenário evidente e esperado de alta correlação entre os dados, pois contempla a observação do mesmo fenômeno em diferentes níveis de profundidade. No entanto, foi utilizado para exemplificar e apresentar a capacidade da representação por Coordenadas Paralelas em destacar tais correlações.

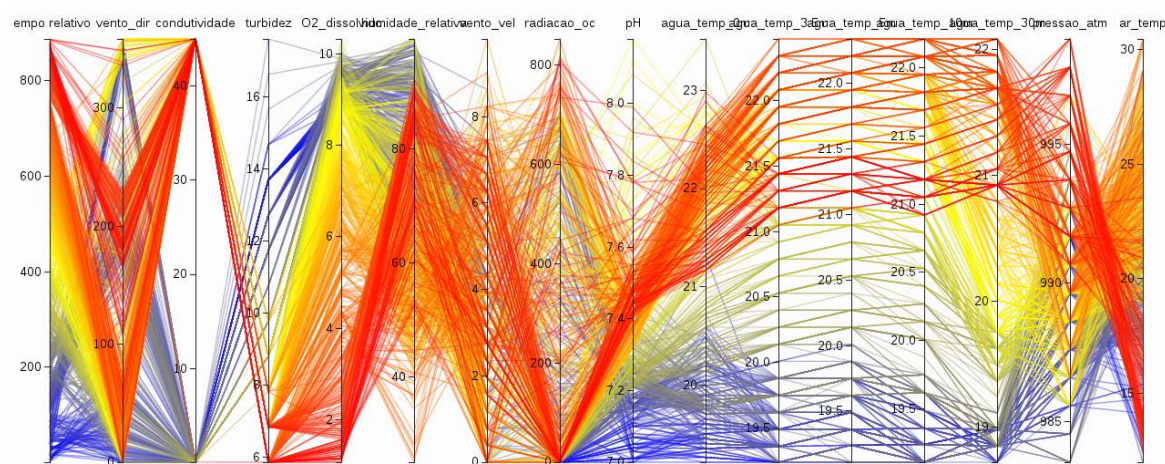
Um dos principais fatores que influenciam diretamente o resultado da utilização da representação por Coordenadas Paralelas está relacionado com a quantidade de variáveis e a ordenação de apresentação dos eixos. Utilizando este mesmo conjunto de dados de

Figura 63 – Representação visual por Coordenadas Paralelas. A simetria entre as arestas que conectam os eixos confirma a alta correlação entre a variabilidade dos dados. A cor está relacionada com o atributo temporal, exibido no primeiro eixo



Itaipu, contendo 887 registros e 15 variáveis (e mais o tempo da coleta), a representação por Coordenadas Paralelas por se tornar confusa, como pode ser visto na Figura 64

Figura 64 – Representação visual por Coordenadas Paralelas exibindo todos os atributos coletados pela plataforma Itaipu sem ordenação dos eixos, dificultando a interpretação



Na Figura 64, a identificação de padrões de comportamento não é uma tarefa intuitiva, em razão da existência de uma grande quantidade de atributos apresentados sem ordenação. Algumas abordagens podem ser realizadas para facilitar a interpretação da representação gerada. Uma delas é a segmentação do conjunto de atributos de acordo com a correlação entre os dados.

Nas análises anteriores, a alta correlação entre os atributos relacionados à água já foi demonstrada e exemplificada, portanto, para diminuir a quantidade de atributos utilizados para a análise, pode ser criado um único índice que represente a variabilidade

dos cinco atributos que representam a temperatura observável.

Uma abordagem possível para a sintetização desses atributos em uma única nova variável é a utilização da técnica de Análise de Componentes Principais (PCA - *Principal Component Analysis*). De forma geral, segundo Johnson (1998), PCA utiliza algumas combinações lineares entre variáveis originais para a criação de um novo conjunto de variáveis capazes de explicar a estrutura de variância e covariância do conjunto, com o objetivo de reduzir a dimensão dos dados e facilitar a interpretação. Essa técnica é comum na literatura em diversos contextos, como avaliação da qualidade da água (GUEDES et al., 2012), classificação de elementos químicos (LYRA et al., 2010), entre outros.

A técnica PCA foi aplicada sobre os dados de temperatura da água, resumizando em um único fator capaz de explicar mais de 94% da variabilidade dos dados de temperatura da água, como pode ser observado na Figura 65 que exhibe os autovalores e autovetores relevantes.

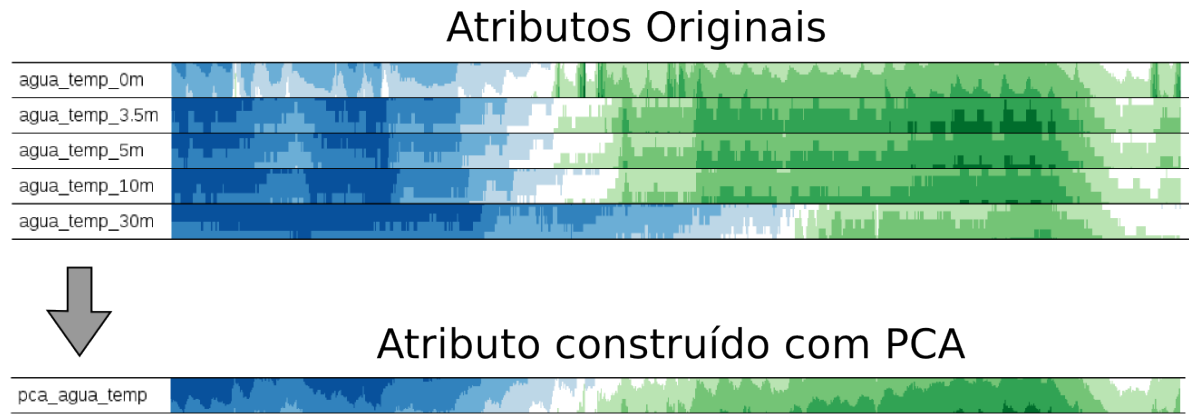
Figura 65 – Componentes Principais dos dados da temperatura da água de um conjunto de dados extraído de Itaipu

Autovalor	4,7302	0,2052	0,0571	0,0058	0,0017
Explicação	0,946	0,041	0,011	0,001	0,000
Acumulado	0,946	0,987	0,998	1,000	1,000

Componentes Principais				
Variáveis	PC1	PC2	PC3	PC4
agua_temp_0m	0,442	-0,452	-0,769	-0,094
agua_temp_3.5m	0,457	-0,159	0,267	0,535
agua_temp_5m	0,457	-0,133	0,331	0,306
agua_temp_10m	0,457	-0,070	0,394	-0,781
agua_temp_30m	0,422	0,865	-0,270	0,033

Considerando as contribuições homogêneas do primeiro Componente Principal com peso distribuído entre todas as cinco variáveis, foi criada uma nova variável para representar a temperatura da água obtida com PCA. O comportamento deste novo atributo pode ser visto com sua representação por *Horizon Charts*, na Figura 66, indicando a transformação das cinco séries temporais referentes aos atributos da temperatura da água em um única nova série temporal derivada.

Figura 66 – Comportamento do novo fator temperatura da água, obtido com a aplicação da técnica PCA, de um conjunto de dados extraído de Itaipu



Após a sintetização dos atributos da temperatura da água em uma única nova variável, o conjunto antes formado por 15 atributos, e mais a componente temporal, foi

reduzido para 11 atributos. Esta abordagem foi utilizada para ilustrar uma possibilidade para redução da dimensionalidade do conjunto, ação que pode ser aplicada em diferentes contextos de análise.

Com o conjunto de atributos reduzidos, pode ser simplificada a tarefa de ordenação dos eixos das Coordenadas Paralelas para ressaltar as relações entre os mesmos. Existem algumas abordagens que podem ser aplicadas para viabilizar a ordenação dos eixos. Peng, Ward e Rundensteiner (2004) discutem sobre alguns algoritmos que podem ser utilizados para redução da poluição causada pelas representações visuais - dentre elas Coordenadas Paralelas - baseados na reordenação dos das dimensões dos dados.

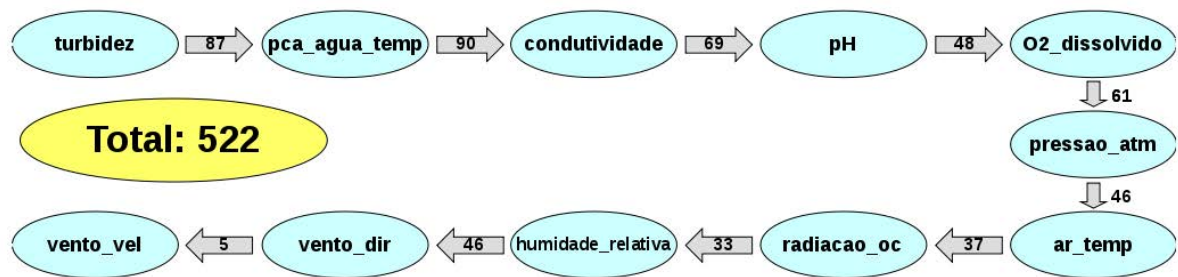
A reordenação ótima dos atributos, ou seja, aquela que melhor apresenta o conjunto de dados, diminui a poluição visual e facilita a identificação de padrões e pode ser obtida a partir de índices que representam a correlação entre cada dimensão dos dados. Uma das formas de se obter valores quantitativos associados a estas relações pode ser a própria matriz de correlação dos dados, com a vantagem dos valores presentes na matriz de correlação serem adimensionais – sem unidade de medida definida. A matriz de correlação calculada para esse conjunto de dados pode ser vista na Tabela 11.

Tabela 11 – Matriz de correlação dos dados de um conjunto extraído de Itaipu

	condutividade	O2_dissolvido	pH	pca_agua_temp	turbidez	vento_dir	vento_vel	pressao_atm	radiacao_oc	ar_temp	umidade_relativa
condutividade	1	-0,56247	0,69572	0,90233	-0,83436	-0,08491	0,17378	0,34151	0,19308	0,29465	-0,45497
O2_dissolvido	-0,56247	1	-0,48641	-0,64459	0,64616	-0,09208	-0,28885	-0,61517	-0,062	0,18717	0,1965
pH	0,69572	-0,48641	1	0,6855	-0,70744	0,09111	-0,06792	0,29791	0,2179	0,20438	-0,21609
pca_agua_temp	0,90233	-0,64459	0,6855	1	-0,87945	-0,13875	0,26052	0,32453	0,14822	0,36824	-0,55314
turbidez	-0,83436	0,64616	-0,70744	-0,87945	1	0,05325	-0,20911	-0,45484	-0,12543	-0,22195	0,38728
vento_dir	-0,08491	-0,09208	0,09111	-0,13875	0,05325	1	0,05571	0,25167	0,07301	-0,37793	0,46629
vento_vel	0,17378	-0,28885	-0,06792	0,26052	-0,20911	0,05571	1	0,21971	0,34284	0,11972	-0,26849
pressao_atm	0,34151	-0,61517	0,29791	0,32453	-0,45484	0,25167	0,21971	1	0,0931	-0,4629	0,16845
radiacao_oc	0,19308	-0,062	0,2179	0,14822	-0,12543	0,07301	0,34284	0,0931	1	0,37883	-0,33711
ar_temp	0,29465	0,18717	0,20438	0,36824	-0,22195	-0,37793	0,11972	-0,4629	0,37883	1	-0,7744
umidade_relativa	-0,45497	0,1965	-0,21609	-0,55314	0,38728	0,46629	-0,26849	0,16845	-0,33711	-0,7744	1

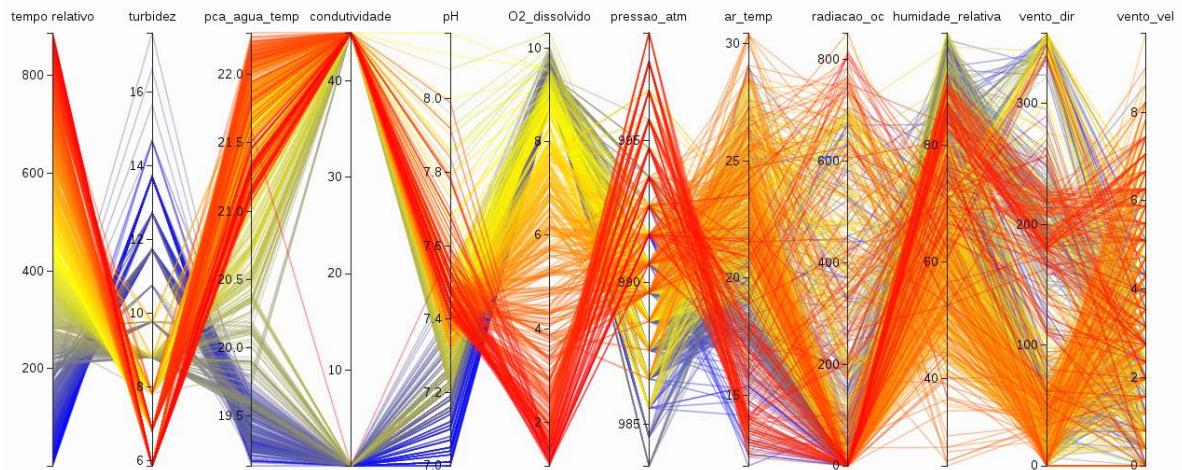
A reordenação dos eixos da representação visual por Coordenadas Paralelas utilizando os índices da matriz de correlação dos dados pode ser realizada a partir da definição da sequência de atributos que contempla os maiores valores de correlação. A escolha da combinação de atributos pode ser visto como um problema de percurso de um dígrafo conexo -grafo bidirecional com todas as conexões possíveis de arestas (Teoria dos Grafos): A matriz de correlação pode ser vista como um dígrafo, no qual cada atributo/variável representa um vértice, e cada valor absoluto de correlação representa o peso da aresta que conecta ambos os atributos, ou seja, o valor de correlação entre os atributos A e B representa o peso da aresta que conecta os vértices A e B. Com esta abordagem, a ordenação dos atributos se resume a encontrar o percurso elementar (caminho aberto sem retornar ao vértice de origem) de custo máximo (considerando todos os cursos como positivos) em que todos os vértices sejam visitados. Um caminho que atende a estas premissas pode ser visto no Figura 67, na qual, para ilustrar, constam apenas os valores inteiros relacionados à cada peso das arestas (valores de correlação).

Figura 67 – Exemplo de percurso sobre o dígrafo representado pela matriz de correlação para definir a ordem dos atributos na representação por Coordenadas Paralelas



O caminho apresentado na Figura 67 pode não ser o caminho ótimo que maximiza o custo do percurso, isto é, essa sequência de atributos pode não ser a que destaca o maior grau de correlação entre os atributos. Porém, mesmo não sendo uma solução ótima, a reordenação dos eixos seguindo esta sugestão traz uma melhoria para a representação visual por Coordenadas Paralelas, como pode ser visto na Figura 68, na qual a distinção e agrupamento de registros que apresentam comportamentos similares é facilitada pela nova ordenação dos eixos.

Figura 68 – Representação em Coordenadas Paralelas dos dados selecionados, após sintetização dos atributos de temperatura da água e reordenação dos eixos de acordo com a matriz de correlação



A representação por Coordenadas Paralelas apresentada na Figura 68 ressalta algumas relações indicadas pela matriz de correlação, como a relação inversa entre turbidez e temperatura da água (87,9%), relação direta entre temperatura da água e condutividade (90,2%), relação direta entre condutividade e pH (69,5%), entre outras.

Após a reordenação dos eixos, os atributos presentes nas últimas dimensões da representação continuam com um aspecto de variação confuso, o qual não permite a identificação de nenhum padrão nos registros, como é o caso de: temperatura do ar; radiação de ondas curtas; umidade relativa; direção do vento e velocidade do vento. Observando a matriz de correlação, é possível notar que estes atributos não possuem índices de correlações

significativos com nenhum outro do conjunto. Com isto, a variabilidade destes dados não apresenta similaridade com nenhuma outra variável.

5.7 Integração dos resultados das técnicas de representações visuais

O *Web Service* implementado no protótipo nomeado como SimaVis facilita a reusabilidade, favorece o acesso aos recursos desenvolvidos e integra todos os materiais e ferramentas utilizados. Além do *Web Service* implementado para fornecer os dados para o Cliente do sistema SimaVis, a arquitetura de serviço e modelagem oferecida pelo *framework* da iniciativa 52°North, no qual atende efetivamente à especificação SOS 2.0 do OGC, foi configurada para receber os dados originais do projeto SIMA. Esta possibilidade oferece um recurso essencial para garantir a interoperabilidade entre diferentes fontes de dados e de serviços que estejam em conformidade com tais padrões. Em virtude de que nem todos os dados coletados do SIMA puderam ser disponibilizados no *framework* 52°North-SOS - como discutido na Seção 5.1.1.2 - optou-se por implementar o *Web Service* personalizado no sistema SimaVis - abordado na Seção 5.1.1.3.

Aproveitando a possibilidade de escalabilidade e flexibilidade oferecida pela construção do sistema SimaVis em um ambiente Web, alguns recursos que contemplam técnicas de Visualização de Informação foram integrados. Tais recursos derivaram uma ferramenta capaz de ser adaptada para aplicação sobre diferentes conjuntos de dados, para facilitar o processo de exploração e análise dos dados.

A representação da localização das plataformas de coleta apresentadas no mapa interativo permitiu que a realização da seleção das fontes de dados, compostas pelas plataformas SIMA, pudessem ser efetuadas considerando o componente espacial dos dados. Com este recurso, integrado com a composição dos filtros, foi possível selecionar plataformas de acordo com a distância entre elas ou aquelas responsáveis por coletar dados em uma determinada região, observando, por exemplo, a existência de plataformas presentes no mesmo reservatório.

Os gráficos interativos implementados 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 intervalo de tempo de operação das plataformas SIMA pôde ser facilmente identificado utilizando o gráfico de intervalo, facilitando a descoberta de outros tipos de informações, como elencar as plataformas que se encontram atualmente ativas coletando dados, identificar as plataformas que possuem registros de dados no mesmo intervalo de tempo, entre outras possibilidades.

O índice de falhas associado às coletas é um fator que pode ser utilizado para

representar a qualidade dos dados coletados ou a eficiência da realização de coletas pelas plataformas. 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, respectivamente, relacionado à identificação de atributos com falhas no registro e sobre a variação dos dados para identificação de registros espúrios (ruídos). Estas representações permitem extrair rapidamente informações sobre qual o conjunto de atributos efetivamente coletados em um determinado instante de tempo e se a variabilidade destes dados apresenta algum comportamento anormal ou inesperado.

A representação visual por máscara de bit permite a exploração dos dados em busca de padrões de comportamentos na ocorrência de erros entre um determinado conjunto de sensores. A concepção da ideia da representação visual por máscara de bit surgiu a partir da necessidade de verificar se o conjunto de atributos com falhas simultâneas poderia apresentar algum padrão sistemático que pudesse 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 por exemplo: falha e sucesso) de dados multivariados.

A abordagem com a técnica de Coordenadas Paralelas, se mostrou um recurso útil para permitir a compreensão das relações dos dados por usuários que não dominam outros métodos de análise, podendo ser um recurso importante e significativo para entendimento de outras abordagens analíticas, como análises estatísticas. As capacidades desta representação em ser integrada com outras ferramentas de análise ressaltam a eficiência da percepção e cognição visual para otimização da interpretação, formulação de hipóteses e descobertas de relações entre os comportamentos dos atributos.

A integração entre diversas técnicas de análise, inclusive representação visuais e abordagens estatísticas são essenciais para amplificar as capacidades e refinar os resultados do processo de exploração e análise de dados. Cada técnica possui características específicas que justificam sua aplicação e motivam a complementação com outras abordagens. Na Tabela 12 são sintetizadas algumas características apresentadas por cada uma das técnicas de representação visual empregada neste trabalho, destacando as principais informações extraídas de cada uma delas em sua utilização neste trabalho.

Na Tabela 12, a coluna indicada por "Múltiplas fontes" se refere a capacidade da representação visual de incorporar dados oriundos de mais de um conjunto de dados, neste trabalho utilizado para mais de uma plataforma simultaneamente. As possibilidades de

Tabela 12 – Síntese das características das técnicas de representações visuais utilizadas

Técnica	Múltiplos atributos	Múltiplas fontes	Temporal	Interatividade	Informação extraída
Gráfico de intervalos	Não	Sim	Sim	Detalhamento	- Intervalo de atividade das plataformas SIMA - Plataformas ativas simultaneamente
Gráfico radial	Não	Sim	Sim	Detalhamento	- Proporção de falhas e sucessos nas coletas
Baseada em pixel	Sim	Não	Sim	Seleção	- Variabilidade dos atributos de acordo com a cor - Atributos com falhas em cor destacada
Horizon Charts	Sim	Não	Sim	Seleção	- Variabilidade dos atributos de acordo com a cor - Variabilidade dos atributos de acordo com a linha - Atributos com falhas de acordo com a linha - Coletas não registradas
Máscara de bits	Sim	Não	Sim	Detalhamento	- Coletas não registradas em cor destacada - Conjunto de Atributos com falha em cada registro - Padrões de falhas entre os conjuntos de atributos
Coordenadas Paralelas	Sim	Não	Não	Reposicionamento	- Correlação entre a variabilidade dos atributos - Padrões no comportamento do conjunto

interatividade foram agrupadas em recursos destinados: ao Detalhamento, para indicar interações com o mouse para a exibição de mais informações sobre o registro; à Seleção, para indicar a capacidade de refinamentos e manipulação de subconjuntos de atributos e o Reposicionamento para caracterizar a possibilidade de reconfigurar posições dos elementos gráficos da representação.

De modo geral, as representações visuais demonstraram possibilidades para potencializar o processo de análise, sobretudo em relação a facilidade de interpretação em razão da capacidade de percepção visual dos comportamentos dos dados e refinamento com recursos de interação. Este fato reforça a potencialidade da integração de técnicas de Visualização de Informação para ampliar o entendimento de outras abordagens analíticas, assim como apontado por Jeong et al. (2009a), que apresentam um estudo comparativo entre os resultados obtidos com a aplicação de análises estatísticas e sua integração com técnicas de Visualização de Informação dispostas em um ambiente de múltiplas visões coordenadas na plataforma interativa IPCA (*Interactive Principal Component Analysis*). As 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., 2009b). Baseados na acurácia, velocidade e qualidade das respostas dos usuários/analistas que foram submetidos aos testes, Jeong et al. (2009a) concluem que a integração com as representações visuais interativas ajudam os usuários a entenderem melhor as relações entre os dados e as características estatísticas extraídas.

CAPÍTULO 6

CONCLUSÕES

O monitoramento ambiental por meio de dispositivos que realizam a coleta automática e periódica de dados tem demonstrado ser uma abordagem que agrega ganhos significativos à esta tarefa. No entanto, a análise do conjunto de dados coletados pelo processo automático de leitura realizada por sensores pode ser prejudicada em razão da inconsistência e falhas dos dados. As falhas nos dados podem ocorrer pois o processo de coleta envolve diversos fatores. No contexto do projeto SIMA, os erros dos dados podem ser inseridos por três principais motivos: falha do equipamento no momento da leitura ou armazenamento; erro na transmissão dos dados via enlace de satélite ou ainda na conversão do sinal elétrico obtido pelo sensor para uma grandeza física. Por esta razão, principalmente no momento de exploração, filtragem e seleção do conjunto de dados a ser utilizado, técnicas para apoiar a análise podem ser úteis para a definição de características do conjunto de dados, como: a localização geográfica das realizações das coletas; o intervalo de tempo com dados disponíveis; a confiabilidade dos dados e a identificação de padrões de comportamento no conjunto de dados.

O desenvolvimento deste trabalho permitiu um estudo sobre as possibilidades e recursos que podem ser utilizados para potencializar o processo de exploração e análise dos dados. A integração entre diferentes abordagens analíticas, como métodos estatísticos e técnicas conhecidas de representação visual dos dados, como gráficos interativos, Coordenadas Paralelas, *Horizon Charts* e outras variações de representações baseadas em pixel, se mostraram eficazes enquanto abordagens para facilitar a extração e interpretação de informações, baseadas na capacidade de percepção e cognição associada ao sistema visual humano.

As representações visuais utilizadas neste trabalho facilitaram a observação de

várias informações e características presentes no conjunto de dados. Informações sobre o período de coleta em que cada plataforma esteve ativa puderam ser intuitivamente extraídas usando o gráfico de intervalos. Os gráficos interativos permitem a extração de termos quantitativos e proporcionais da quantidade de dados coletados e não coletados (falhas), enquanto as representações visuais baseadas em pixel e *Horizon Charts* beneficiaram a percepção de falhas nos atributos coletados e em quais intervalos de tempo elas ocorrem. Padrões no comportamento dos dados, relacionados às falhas sistemáticas e variabilidade dos dados podem ser destacados, respectivamente, com as representações por máscara de bits e por Coordenadas Paralelas. Diante deste cenário, pode ser confirmada a hipótese de que "a exploração visual e interativa potencializa a análise dos dados coletados pelo Sistema Integrado de Monitoramento Ambiental e facilita a observação de características do conjunto de dados, tais como: disponibilidade suficiente de dados; funcionamento dos sensores; evidências de falhas sistemáticas ou não, as quais influenciam a qualidade da análise".

Considerando os resultados discutidos no Capítulo 5, pode-se concluir que foi alcançado o objetivo geral de aplicar técnicas de análise exploratória visual interativa de dados para facilitar a caracterização do conjunto de dados. Este objetivo é primordial e deve ser considerado pelo analista ambiental para a realização das tarefas analíticas para estudos dos fenômenos e significado dos dados, pois a caracterização inicial do conjunto de dados, principalmente associada à existência de falhas e dados espúrios, é um fator que pode influenciar diretamente a qualidade dos resultados e define como o especialista irá abordar e minimizar tais falhas.

As principais contribuições deste trabalho podem ser caracterizadas em dois aspectos, distintos mas diretamente relacionados: as possibilidades de interpretação e extração de informação complementadas por técnicas de Visualização de Informação e a ainda sobre a interoperabilidade e flexibilidade na utilização dos dados por meio de *Web Service* com padrão OGC. Tais contribuições beneficiam: a ação e intervenção de analistas e especialistas do domínio de informação, pois permitem obter de forma rápida e intuitiva, informações preliminares sobre a caracterização do conjunto de dados; projetos de novas plataformas que podem se basear no padrão de falhas encontradas nas plataformas existentes para realizar alterações em sua concepção; e gestores do projeto que podem obter informações sobre o desempenho e qualidade relacionada aos dados coletados pelas plataformas.

A área de Geociências, assim como diversos outros contextos, utiliza mecanismos de coleta de dados baseados em sensores. No entanto, a descoberta e acesso a estes dados é um desafio atual, tendo em vista a grande quantidade de dados coletados, observando diferentes atributos, seguindo diferentes métodos, utilizando ferramentas diversificadas e armazenadas em múltiplas fontes de dados (diferentes iniciativas de instituições). Diante deste cenário, metodologias para padronização e definição da infraestrutura de dados,

como aquelas definidas pelo OGC, são necessárias para favorecer a interoperabilidade dos dados, principalmente no que diz respeito à disponibilização de metadados.

Outro aspecto relevante é a possibilidade de aplicar técnicas interativas de Visualização de Informação para complementar a metodologia de análise e exploração dos dados. Neste trabalho, foram propostas e aplicadas diferentes tipos de representações visuais que foram úteis para auxiliar na caracterização dos dados, permitindo identificar, de maneira rápida, falhas, dados espúrios e padrões de comportamento.

Como possibilidades futuras para expansão dos assuntos abordados neste trabalho pode-se citar recomendações sobre:

- Extensão do escopo da investigação para além da caracterização dos dados, utilizando técnicas de Visualização de Informação também para a análise dos fenômenos associados aos significados dos dados;
- Abordagens para viabilizar a integração de múltiplas fontes de dados, utilizando padrões e outros conceitos como Fusão de Dados, LOD e *Web of Things*;
- Aprofundamento de estudos sobre as possibilidades de associação do padrão SWE como fonte de acesso aos dados utilizados no processo de exploração e análise visual e interativa;
- Aplicação de técnicas de coordenação entre as representações visuais para que os recursos de interatividade sejam ampliados durante o processo de exploração e análise de dados.

Por fim, espera-se que os resultados obtidos e demonstrados 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 sensores (como o SIMA), bem como de estimular a utilização de técnicas de Visualização de Informação como complemento e potencializador de outros procedimentos analíticos no processo de exploração e análise de dados.

REFERÊNCIAS

- 52NORTH. *52N Sensor Web Community - Sensor Observation Service*. 2015. <<http://52north.org/communities/sensorweb/sos/>>. Acesso em Janeiro de 2015. Citado na página 43.
- 52NORTH. *History - 52^o North Initiative for Geospatial Open Source Software GmbH*. 2015. <<http://52north.org/about/52north/history>>. Acesso em Janeiro de 2015. Citado na página 42.
- AGOSTINHO, A. A.; JR, H. F. J.; BORGHETTI, J. R. Considerações sobre os impactos dos represamentos na ictiofauna e medidas para sua atenuação. um estudo de caso: reservatório de itaipu. *Revista Unimar*, v. 14, n. ssupl, 1992. Citado na página 27.
- AHLBERG, C.; SHNEIDERMAN, B. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1994. (CHI '94), p. 313–317. ISBN 0-89791-650-6. Disponível em: <<http://doi.acm.org/10.1145/191666.191775>>. Citado na página 46.
- AIGNER, W. et al. *Visualization of Time-Oriented Data*. Springer, 2011. (Human–Computer Interaction Series). ISBN 9780857290793. Disponível em: <<http://books.google.com.br/books?id=YnDivwba2nkC>>. Citado 2 vezes nas páginas 11 e 54.
- ALCÂNTARA, E. et al. A system for environmental monitoring of hydroelectric reservoirs in brazil. *Ambiente e Água - An Interdisciplinary Journal of Applied Science*, v. 8, n. 1, 2013. ISSN 1980-993X. Disponível em: <<http://www.ambi-agua.net/seer/index.php/ambi-agua/article/view/1088>>. Citado 3 vezes nas páginas 65, 66 e 100.
- ALONSO, E. M. B. Plataforma de coleta de dados. In: *II Congresso Brasileiro de Meteorologia*. [S.l.: s.n.], 1982. p. 394–402. Citado na página 30.
- ANKERST, M. *Visual data mining*. [S.l.]: Dissertation. de, 2001. Citado na página 48.
- ARAYA, P. R.; AGOSTINHO, A. A.; BECHARA, J. A. The influence of dam construction on a population of leporinus obtusidens (valenciennes, 1847)(pisces, anostomidae) in the

yacyretá reservoir (argentina). *Fisheries Research*, Elsevier, v. 74, n. 1, p. 198–209, 2005. Citado na página 27.

BERNERS-LEE, T. Linked data-design issues. URL <http://www.w3.org/DesignIssues/LinkedData.html>, 2006. Citado na página 32.

BERNERS-LEE, T.; CONNOLLY, D. Notation3 (n3): A readable rdf syntax. *W3C Team Submission (January 2008)* <http://www.w3.org/TeamSubmission/3>, 1998. Citado na página 32.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1–22, 2009. Disponível em: <<http://eprints.soton.ac.uk/271285/>>. Citado na página 32.

BORGO, R. et al. Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *Visualization and Computer Graphics, IEEE Transactions on*, IEEE, v. 16, n. 6, p. 963–972, 2010. Citado 2 vezes nas páginas 11 e 52.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3; data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, v. 17, n. 12, p. 2301–2309, Dec 2011. ISSN 1077-2626. Citado na página 69.

BRÖRING, A. et al. New generation sensor web enablement. *Sensors*, v. 11, n. 3, p. 2652–2699, 2011. ISSN 1424-8220. Disponível em: <<http://www.mdpi.com/1424-8220/11/3/2652>>. Citado 2 vezes nas páginas 36 e 42.

BRÖRING, A. et al. Development of sensor web applications with open source software. In: *First Open Source GIS UK Conference (OSGIS 2009)*. [S.l.: s.n.], 2009. v. 22. Citado 2 vezes nas páginas 10 e 37.

BRÖRING, A.; STASCH, C.; ECHTERHOFF, J. Ogc sensor observation service interface standard. *Open Geospatial Consortium, Inc*, 2012. Citado 3 vezes nas páginas 10, 39 e 41.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. *Readings in information visualization: using vision to think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN 1558605339. Citado 2 vezes nas páginas 47 e 49.

CRESSIE, N.; WIKLE, C. K. *Statistics for spatio-temporal data*. [S.l.]: John Wiley & Sons, 2011. Citado na página 51.

DEEGREE. *About deegree*. 2015. <<http://www.deegree.org/About>>. Acesso em Janeiro de 2015. Citado na página 43.

DEEGREE. *deegree3/SensorObservationService - deegree Wiki*. 2015. <<http://wiki.deegree.org/deegreeWiki/deegree3/SensorObservationService>>. Acesso em Janeiro de 2015. Citado na página 43.

DEGU, A. M. et al. The influence of large dams on surrounding climate and precipitation patterns. *Geophysical Research Letters*, Wiley Online Library, v. 38, n. 4, 2011. Citado na página 27.

DEKKER, A.; VOS, R.; PETERS, S. Comparison of remote sensing data, model results and in situ data for total suspended matter (tsm) in the southern frisian lakes. *Science of the Total Environment*, Elsevier, v. 268, n. 1, p. 197–214, 2001. Citado na página 29.

- DEVILLERS, R.; JEANSOULIN, R. et al. *Fundamentals of spatial data quality*. [S.l.]: ISTE London, 2006. Citado na página 28.
- DI, L. Standards, critical evaluation of remote sensing. In: SHEKHAR, S.; XIONG, H. (Ed.). *Encyclopedia of GIS*. [S.l.]: Springer, 2008. p. 1128–1135. ISBN 978-0-387-30858-6. Citado na página 31.
- DINU, V.; NADKARNI, P. Guidelines for the effective use of entity–attribute–value modeling for biomedical databases. *International journal of medical informatics*, Elsevier, v. 76, n. 11, p. 769–779, 2007. Citado na página 73.
- ESLING, P.; AGON, C. Time-series data mining. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 45, n. 1, p. 12:1–12:34, dez. 2012. ISSN 0360-0300. Disponível em: <<http://doi.acm.org/10.1145/2379776.2379788>>. Citado na página 30.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado na página 47.
- FEARNSIDE, P. M. Greenhouse gas emissions from a hydroelectric reservoir (brazil's tucuruí dam) and the energy policy implications. *Water, Air, and Soil Pollution*, Springer, v. 133, n. 1-4, p. 69–96, 2002. Citado na página 27.
- FREITAS, A. et al. Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data & Knowledge Engineering*, Elsevier, v. 88, p. 126–141, 2013. Citado na página 73.
- GIANNECCHINI, S. et al. Supporting interoperable geospatial data fusion by adopting ogc and iso tc 211 standards. In: *Information Fusion, 2006 9th International Conference on*. [S.l.: s.n.], 2006. p. 1–8. Citado na página 30.
- GLENN, E. P. et al. Integrating remote sensing and ground methods to estimate evapotranspiration. *Critical Reviews in Plant Sciences*, Taylor & Francis, v. 26, n. 3, p. 139–168, 2007. Citado na página 29.
- GUEDES, H. A. S. et al. Aplicação da análise estatística multivariada no estudo da qualidade da água do rio pomba, mg. *Revista Brasileira de Engenharia Agrícola e Ambiental*, scielo, v. 16, p. 558 – 563, 05 2012. ISSN 1415-4366. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-43662012000500012&nrm=iso>. Citado na página 105.
- HARTMANN, C. et al. Analise comparativa de custos entre métodos convencionais e de satélite na região da laguna dos patos, rs. *Simpósio Brasileiro de Sensoriamento Remoto, 6 (SBSR)*, City São José dos Campos, p. 587–595, 1990. Citado na página 29.
- HAUSER, H.; LEDERMANN, F.; DOLEISCH, H. Angular brushing of extended parallel coordinates. In: *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*. [S.l.: s.n.], 2002. p. 127–130. ISSN 1522-404X. Citado 2 vezes nas páginas 11 e 57.
- HEALEY, C. G.; BOOTH, K. S.; ENNS, J. T. Visualizing real-time multivariate data using preattentive processing. *ACM Trans. Model. Comput. Simul.*, ACM, New York, NY, USA, v. 5, n. 3, p. 190–221, jul. 1995. ISSN 1049-3301. Disponível em: <<http://doi.acm.org/10.1145/217853.217855>>. Citado na página 58.

- HEER, J.; AGRAWALA, M. Design considerations for collaborative visual analytics. In: *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*. Washington, DC, USA: IEEE Computer Society, 2007. p. 171–178. ISBN 978-1-4244-1659-2. Citado 3 vezes nas páginas 10, 47 e 49.
- HEER, J.; KONG, N.; AGRAWALA, M. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2009. (CHI '09), p. 1303–1312. ISBN 978-1-60558-246-7. Disponível em: <<http://doi.acm.org/10.1145/1518701.1518897>>. Citado 2 vezes nas páginas 11 e 53.
- HEINRICH, J.; WEISKOPF, D. State of the art of parallel coordinates. In: THE EUROGRAPHICS ASSOCIATION. *Eurographics 2013-State of the Art Reports*. [S.l.], 2012. p. 95–116. Citado 3 vezes nas páginas 11, 55 e 56.
- HINTON, J. Gis and remote sensing integration for environmental applications. *International Journal of Geographical Information Systems*, Taylor & Francis, v. 10, n. 7, p. 877–890, 1996. Citado na página 29.
- ICKE, I. *Visual Analytics: A Multifaceted Overview*. 2009. Citado na página 49.
- INPE, H. *SIMA: Sistema Integrado de Monitoramento Ambiental*. 2014. <<http://www.dsr.inpe.br/hidrosfera/sima/>>. Acesso em Março de 2014. Citado 4 vezes nas páginas 11, 22, 64 e 66.
- INSELBERG, A.; DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*. [S.l.: s.n.], 1990. p. 361–378. Citado 2 vezes nas páginas 11 e 55.
- JENSEN, J. R. et al. *Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres*. [S.l.]: Parêntese Editora, 2009. Citado na página 29.
- JEONG, D. H. et al. ipca: An interactive system for pca-based visual analytics. In: *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2009. (EuroVis'09), p. 767–774. Disponível em: <<http://dx.doi.org/10.1111/j.1467-8659.2009.01475.x>>. Citado na página 110.
- JEONG, D. H. et al. Understanding principal component analysis using a visual analytics tool. *2009 US-Korea Conference on Science, Technology and Entrepreneurship*, 2009. Citado na página 110.
- JOHNSON, R. *Applied multivariate statistical analysis*. 4. ed. ed. Upper Saddle River, NJ [u.a.]: Prentice Hall, 1998. ISBN 013834194X. Disponível em: <http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+237428075&sourceid=fbw_bibsonomy>. Citado na página 105.
- KEIM, D. et al. *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics Association, 2010. ISBN 9783905673777. Disponível em: <<http://books.google.com.br/books?id=vdv5wZM8ioIC>>. Citado 2 vezes nas páginas 10 e 48.

- KEIM, D. A. Designing pixel-oriented visualization techniques: Theory and applications. *Visualization and Computer Graphics, IEEE Transactions on*, IEEE, v. 6, n. 1, p. 59–78, 2000. Citado 2 vezes nas páginas 11 e 52.
- KEIM, D. A. et al. Challenges in visual data analysis. In: IEEE. *Information Visualization, 2006. IV 2006. Tenth International Conference on*. [S.l.], 2006. p. 9–16. Citado 5 vezes nas páginas 10, 25, 49, 50 e 69.
- KESSLER, C.; JANOWICZ, K. Linking sensor data-why, to what, and how? In: *SSN*. [S.l.: s.n.], 2010. Citado na página 34.
- KNOOP, H. International standardization and management of gis-activities. *INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY AND REMOTE SENSING*, INTERNATIONAL SOCIETY FOR PHOTOGRAMMETRY & REMOTE, v. 33, n. B6; PART 6, p. 164–171, 2000. Citado na página 28.
- KOJIMA, I. et al. Implementation of the fukushima radiation lod framework. *W3C OGC Linking Geospatial Data*, Scientific Research Publishing, 2014. Citado na página 32.
- KRESSE, W.; FADAIE, K. Iso standards for geographic information. In: _____. [S.l.]: Springer, 2004. cap. Liaison members of ISO/TC211, p. 185–212. Citado na página 31.
- LAMMARSCH, T. et al. Hierarchical temporal patterns and interactive aggregated views for pixel-based visualizations. In: IEEE. *Information Visualisation, 2009 13th International Conference*. [S.l.], 2009. p. 44–50. Citado 2 vezes nas páginas 11 e 52.
- LEVKOWITZ, H.; KELLEHER, C. Cloud and mobile web-based graphics and visualization. In: *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2012 25th SIBGRAPI Conference on*. [S.l.: s.n.], 2012. p. 21–35. Citado na página 70.
- LIN, G.; WANG, Y.; HU, Z. Hydrodynamic pressure on arch dam and gravity dam including absorption effect of reservoir sediments. In: IOP PUBLISHING. *IOP Conference Series: Materials Science and Engineering*. [S.l.], 2010. v. 10, n. 1, p. 012234. Citado na página 27.
- LONGHORN, R. A. Geospatial standards, interoperability, metadata semantics and spatial data infrastructure. In: *Proceedings NIEeS Workshop on Activating Metadata, July 6-7 2005, Cambridge, UK*. [S.l.: s.n.], 2005. p. 1–23. Citado na página 35.
- LYRA, W. da S. et al. Classificação periódica: um exemplo didático para ensinar análise de componentes principais. *Química Nova*, scielo, v. 33, p. 1594 – 1597, 00 2010. ISSN 0100-4042. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422010000700030&nrm=iso>. Citado na página 105.
- MACIEJEWSKI, R. et al. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 2, p. 205–220, mar. 2010. ISSN 1077-2626. Disponível em: <<http://dx.doi.org/10.1109/TVCG.2009.100>>. Citado na página 51.
- MAPSERVER. *SOS Server - MapServer 6.4.1 documentation*. 2015. <http://mapserver.org/ogc/sos_server.html>. Acesso em Janeiro de 2015. Citado na página 43.

- MAPSERVER. *Welcome to MapServer - MapServer 6.4.1 documentation*. 2015. <<http://mapserver.org/>>. Acesso em Janeiro de 2015. Citado na página 43.
- MASINTER, L.; BERNERS-LEE, T.; FIELDING, R. T. Uniform resource identifier (uri): Generic syntax. 2005. Citado na página 32.
- MCFERREN, G. et al. Evaluating sensor observation service implementations. In: *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*. [S.l.: s.n.], 2009. v. 5, p. V-363-V-366. Citado 2 vezes nas páginas 44 e 45.
- MORADKHANI, H. Hydrologic remote sensing and land surface data assimilation. *Sensors, Molecular Diversity Preservation International*, v. 8, n. 5, p. 2986-3004, 2008. Citado na página 29.
- MOUMEN, A. et al. A sensor web for real-time groundwater data monitoring in morocco. *Journal of Geographic Information System*, Scientific Research Publishing, v. 6, n. 06, p. 613, 2014. Citado na página 31.
- MULLER, W.; SCHUMANN, H. Visualization methods for time-dependent data - an overview. In: *Simulation Conference, 2003. Proceedings of the 2003 Winter*. [S.l.: s.n.], 2003. v. 1, p. 737-745 Vol.1. Citado na página 58.
- NA, A.; PRIEST, M. Opendgis sensor observation service implementation specification. *Open Geospatial Consortium Implementation Specification*, v. 91, 2006. Citado na página 75.
- NADKARNI, P. M. et al. Organization of heterogeneous scientific data using the eav/cr representation. *Journal of the American Medical Informatics Association*, BMJ Publishing Group Ltd, v. 6, n. 6, p. 478-493, 1999. Citado na página 73.
- NETWORK, O. *Sensor Observation Service (SOS) Introduction*. 2014. <http://www.ogcnetwork.net/sos_intro>. Citado 2 vezes nas páginas 10 e 40.
- PELLEC-DAIRON, M. L. *From chains to platforms: valuing remote sensing data for environmental management*. 2012. Citado na página 29.
- PENG, W.; WARD, M.; RUNDENSTEINER, E. Clutter reduction in multi-dimensional data visualization using dimension reordering. In: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*. [S.l.: s.n.], 2004. p. 89-96. ISSN 1522-404X. Citado na página 106.
- PENG, Z.-R.; TSOU, M.-H. Internet gis: distributed geographic information services for the internet and wireless networks. In: _____. [S.l.]: John Wiley & Sons, 2003. cap. Standards for Distributed GIServices, p. 257-297. Citado na página 31.
- PERCIVALL, G. *Geospatial Temporal Open Standards for Big Data from Space*. 2014. Slides de palestra na 2014 Conference on Big Data from Space, BIDS 14, 12-14 de Novembro 2014, Frascati, Itália. Disponível em: <<http://pt.slideshare.net/GeorgePercivall/geospatial-temporal-open-standards-for-big-data-from-space-bids2014>>. Citado 2 vezes nas páginas 10 e 35.

- POORAZIZI, M. E.; LIANG, S. H.; HUNTER, A. J. Testing of sensor observation services: a performance evaluation. In: ACM. *Proceedings of the First ACM SIGSPATIAL Workshop on Sensor Web Enablement*. [S.l.], 2012. p. 32–38. Citado 3 vezes nas páginas 10, 43 e 44.
- REED, C. et al. Ogc sensor web enablement:overview and high level achhitecture. In: *Autotestcon, 2007 IEEE*. [S.l.: s.n.], 2007. p. 372–380. ISSN 1088-7725. Citado 5 vezes nas páginas 10, 36, 37, 38 e 39.
- RITCHIE, J. C.; ZIMBA, P. V.; EVERITT, J. H. Remote sensing techniques to assess water quality. *Photogrammetric Engineering and Remote Sensing*, v. 69, n. 6, p. 695–704, 2003. Disponível em: <http://www.asprs.org/a/publications/pers/2003journal/june/2003_jun_695-704.pdf>. Citado 3 vezes nas páginas 23, 27 e 29.
- SCHMACHTENBERG, M. et al. Linking open data cloud diagram 2014. *LOD Community* (<http://lod-cloud.net/>), 2014. Citado 3 vezes nas páginas 10, 33 e 34.
- SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the linked data best practices in different topical domains. In: *The Semantic Web–ISWC 2014*. [S.l.]: Springer, 2014. p. 245–260. Citado na página 33.
- SHETH, A.; HENSON, C.; SAHOO, S. Semantic sensor web. *Internet Computing, IEEE*, v. 12, n. 4, p. 78–83, July 2008. ISSN 1089-7801. Citado na página 34.
- SHNEIDERMAN, B. *Designing the user interface: strategies for effective human-computer interaction*. [S.l.]: Addison-Wesley Reading, MA, 1992. Citado na página 46.
- SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE. *Visual Languages, 1996. Proceedings., IEEE Symposium on*. [S.l.], 1996. p. 336–343. Citado na página 48.
- SILVA, S. F.; CATARCI, T. Visualization of linear time-oriented data: a survey. In: *Web Information Systems Engineering, 2000. Proceedings of the First International Conference on*. [S.l.: s.n.], 2000. v. 1, p. 310–319 vol.1. Citado na página 57.
- STASKO, J.; GÖRG, C.; LIU, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, Palgrave Macmillan, Sacramento, v. 7, p. 118–132, April 2008. ISSN 1473-8716. Citado na página 49.
- STECH, J. L. et al. Uso de tecnologia espacial para coleta automática de dados limnológicos e meteorológicos: Aplicações nos reservatórios hidrelétricos de manso e corumbá. In: ALCÂNTARA, E. H. d.; NOVO, E. M. L. d. M.; STECH, J. L. (Ed.). *Novas tecnologias para o monitoramento e estudo de reservatórios hidrelétricos e grandes lagos*. [S.l.]: OParêntese, 2011. Citado na página 63.
- STECH., J. L. et al. Uso de tecnologia espacial para coleta automática de dados limnológicos e meteorológicos: Aplicações nos reservatórios hidrelétricos de manso e corumbá. In: *Novas tecnologias para o monitoramento e estudo de reservatórios hidrelétricos e grandes lagos*. [S.l.]: Parêntese, 2011. cap. 4, p. 119–162. Citado na página 22.
- THOMAS, J. J.; COOK, K. A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. [S.l.]: National Visualization and Analytics Ctr, 2005. Paperback. ISBN 0769523234. Citado 3 vezes nas páginas 46, 49 e 51.

- TUNDISI, J. G. Reservatórios como sistemas complexos: Teoria, aplicações e perspectivas de usos múltiplos. *HENRY, R. (Ed.): Ecologia de reservatórios: estrutura, função e aspectos Sociais*, p. 29–38, 1999. Citado na página 28.
- VALERIO, A. d. M. et al. Variabilidade dos dados da bóia sima analisados pelo operador de fragmentação assimétrico. *XV Simpósio Brasileiro de Sensoriamento Remoto (SBSR-2011)*, 2011. Citado na página 100.
- WANG, Z. et al. Single aircraft integration of remote sensing and in situ sampling for the study of cloud microphysics and dynamics. *Bulletin of the American Meteorological Society*, American Meteorological Society, v. 93, n. 5, p. 653–668, 2012. Citado na página 29.
- WARD, M.; GRINSTEIN, G.; KEIM, D. *Interactive Data Visualization: Foundations, Techniques, and Applications*. [S.l.]: A K Peters, Ltd, 2010. ISBN 9781568814735. Citado 2 vezes nas páginas 25 e 50.
- WIATKOWSKI, M. Influence of slup dam reservoir on flow and quality of water in the nysa szalona river. 2011. Citado na página 27.
- WIJK, J. J. van. The Value of Visualization. In: . [S.l.: s.n.], 2005. p. 79–86. Citado na página 49.
- WIJK, J. van; SELOW, E. V. Cluster and calendar based visualization of time series data. In: *Information Visualization, 1999. (Info Vis '99) Proceedings. 1999 IEEE Symposium on*. [S.l.: s.n.], 1999. p. 4–9, 140. ISSN 1522-404X. Citado na página 60.
- YUAN, X. et al. Scattering points in parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, v. 15, n. 6, p. 1001–1008, Nov 2009. ISSN 1077-2626. Citado 2 vezes nas páginas 11 e 56.