UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

FELIPE MARX BENGHI

VISUAL ANALYTICS E OUTLYING ASPECT MINING: CONTEXTUALIZAÇÃO DE ANOMALIAS CONSIDERANDO QUESTÕES TEMPORAIS E MULTIDIMENSIONAIS

DISSERTAÇÃO DE MESTRADO

CURITIBA
2020

FELIPE MARX BENGHI

VISUAL ANALYTICS E OUTLYING ASPECT MINING: CONTEXTUALIZAÇÃO DE ANOMALIAS CONSIDERANDO QUESTÕES TEMPORAIS E MULTIDIMENSIONAIS

Visual Analytics and Outlying Aspect Mining: contextualization of anomalies considering temporal and multidimensional issues

Dissertação de Mestrado apresentada como requisito parcial à obtenção do título de Mestre em Computação Aplicada, do Programa de Pós-Graduação em Computação Aplicada , da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Luiz Celso Gomes Jr.

CURITIBA

2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Curitiba

FELIPE MARX BENGHI

VISUAL ANALYTICS E OUTLYING ASPECT MINING: CONTEXTUALIZAÇÃO DE ANOMALIAS CONSIDERANDO QUESTÕES TEMPORAIS E MULTIDIMENSIONAIS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Sistemas Computacionais.

Data de aprovação: 17 de Novembro de 2020

Prof Luiz Celso Gomes Junior, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Andre Santanche, Doutorado - Universidade Estadual de Campinas (Unicamp)

Prof.a Myriam Regattieri De Biase Da Silva Delgado, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 17/11/2020.

AGRADECIMENTOS

A minha família pelo suporte e dedicação desde sempre.

A meu orientador pela compreensão e incentivo, mesmo nas mudanças de tema e nos momentos em que não pude me dedicar à pesquisa.

“Things always become obvious after the fact”

(Nassim Nicholas Taleb).

RESUMO

BENGHI, Felipe Marx. **Visual Analytics e Outlying Aspect Mining: contextualização de anomalias considerando questões temporais e multidimensionais**. 2020. 53 f. Dissertação de Mestrado (Mestrado Profissional em Computação Aplicada) – Universidade Tecnológica Federal do Paraná. Curitiba, 2020.

Outlying Aspect Mining (OAM) é um novo método para o tratamento de anomalias que, em vez de focar somente na detecção, também fornece uma explicação para o estado anormal. Para tanto, é apresentado um subespaço com os atributos considerados mais relevantes para a compreensão dos aspectos excepcionais da amostra. Existem muitos desafios na aplicação de OAM, como a explosão combinatória do espaço de busca e a habilidade de se comparar métricas calculadas para subespaços com diferentes dimensionalidades. Ainda assim, listar um grupo de atributos não é o bastante para um especialista humano compreender a situação e tomar as medidas necessárias. Uma abordagem visual e de alto nível pode melhorar o processo e fornecer melhores indícios cognitivos para especialistas. Neste trabalho, descrevemos a aplicação de uma técnica de OAM em um problema de detecção de falhas em locomotivas. A partir da experiência adquirida neste caso de uso, propusemos e desenvolvemos uma plataforma de Análise Visual para processamento e apresentação de dados de forma amigável a humanos. Uma novidade disponível nesta plataforma são gráficos de coordenadas paralelas que exibem dados temporais multidimensionais. Esta representação busca contornar as limitações do sistema visual humano e ajuda na investigação de anomalias. Para explorar e validar a usabilidade da ferramenta desenvolvida, o caso de uso de operação de locomotivas é novamente empregado.

Palavras-chave: Outlying Aspect Mining. Explicação de Anomalias. Detecção de Outliers. Análise Visual. Operação de Locomotivas.

ABSTRACT

BENGHI, Felipe Marx. **Visual Analytics and Outlying Aspect Mining: contextualization of anomalies considering temporal and multidimensional issues**. 2020. 53 p. Dissertation (Master's Degree in Course Name) – Universidade Tecnológica Federal do Paraná. Curitiba, 2020.

Outlying Aspect Mining (OAM) is a new way of handling outliers that, instead of focusing solely on the detection, also provides an explanation for the abnormal status. For this purpose, a subspace of attributes considered as the most relevant for understanding the sample outlying aspects is presented. There are many challenges associated with the application of OAM, such as combinatorial explosion of the search space and ability to compare metrics calculated for subspaces with different dimensionalities. Even so, listing a group of attributes is not sufficient for a human specialist to comprehend the situation and take the necessary actions. A higher-level, visual approach can improve the process by providing better cognitive clues to experts. Here we describe the application of an OAM technique in a fault detection problem for locomotives. Based on the experience obtained in this use case, we proposed and developed a Visual Analytics platform for the processing and representation of data in a user-friendly interface. A novelty available on this platform are parallel coordinates plots that also display temporal multidimensional data. Such representation tries to circumvent human visual system limitations and helps the outlier investigation. To explore and validate the applicability of the developed tool, the locomotive operation use case is employed again.

Keywords: Outlying Aspect Mining. Anomaly Explanation. Outlier Detection. Visual Analytics. Locomotive Operation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

ABREVIATURAS

GPSVEL	Velocidade de deslocamento
iForest	Isolation Forest
IGA	Corrente Gerador Auxiliar
IGP	Corrente Gerador Principal
iPath	Isolation Path
LOF	Local Outlier Factor
MT1	Corrente Motor de tração
OAM	Outlying Aspect Mining
PCA	Principal Component Analysis
PCOMB	Pressão combustível
PCP	Parallel Coordinate Plot
PLUB	Pressão lubrificante
POT	Potência
PTO	Ponto de operação
ROT	Rotação
SVM	Support Vector Machine
TH2O	Temperatura da água
TOA	Temperatura óleo antes do arrefecimento
TOD	Temperatura óleo depois do arrefecimento
VA	Visual Analytics
VGA	Corrente Gerador Principal
VGP	Tensão Gerador Principal
WRST	Wilcoxon rank sum test

SUMÁRIO

1 INTRODUÇÃO

A crescente integração entre sistemas e a maior disponibilidade e variabilidade de sensores faz com que conjuntos de dados de *alta dimensionalidade e variantes com o tempo* sejam cada vez mais comuns. Por exemplo, uma única locomotiva moderna em operação pode gerar dados para centenas de variáveis, com atributos discretos para a supervisão do acionamento de diferentes equipamentos e variáveis contínuas para a inspeção de grandezas físicas como a temperatura dos motores, tensão de geradores etc. Neste contexto de abundância de dados e variáveis, o segmento de detecção de anomalias ou *outliers* vem despertando bastante interesse (CHANDOLA *et al.*, 2009).

Estudos dos elementos que destoam (*outliers*) das demais amostras em um conjunto (*inliers*) são aplicados hoje para detecções de fraudes financeiras, identificação de invasores e até mesmo diagnósticos de câncer (CHANDOLA *et al.*, 2009). Por sua vez, uma abordagem mais recente, Outlying Aspect Mining (OAM), não se limita a somente encontrar *outliers*, mas também busca apresentar uma explicação ou interpretação para a anormalidade das amostras. Dentre as técnicas de OAM, algumas são capazes de identificar de outliers, no entanto, a ênfase destes algoritmos está em encontrar os atributos que contextualizem de forma mais representativa as diferenças do ponto sob análise em relação aos outros elementos. Neste sentido, nos casos em que a técnica de OAM não é capaz de identificar anomalias, ambas as estratégias (detecção e OAM) podem ser usadas de forma complementar.

Dentre as técnicas de OAM, destaca-se o *isolation path* (VINH *et al.*, 2016). Este algoritmo utiliza partições sucessivas e aleatórias do conjunto de dados para determinar os atributos mais relevantes para a interpretação da anormalidade de uma amostra. Assim, dentre as vantagens da aplicação desta técnica destacam-se a ausência de cálculos complexos e a possibilidade de realização do cálculo somente para o ponto de interesse em vez de todo o conjunto dados, propriedades relevantes no âmbito de OAM dado que limitações computacionais são recorrentes neste ramo.

Apesar dos avanços trazidos pelas técnicas citadas acima, a análise de especialistas continua imprescindível em um ambiente de bancos de dados multidimensionais, porém mais complexa. Como a representação gráfica de dimensionalidades mais altas em superfícies e monitores de vídeo implica em perdas ou simplificações, soluções alternativas precisam ser consideradas para tornar os dados palpáveis a humanos.

Neste sentido, as técnicas de Visual Analytics (VA) ou Análise Visual frequentemente vêm ao auxílio dos especialistas, pois não se limitam à geração de imagens, mas também buscam determinar *o que* e *como* os elementos devem ser exibidos a partir da vasta quantidade e variedade de referências disponíveis. Para tanto, VA engloba pesquisas sobre a interação homem-máquina, gerenciamento/análise de dados e estatísticas (KEIM *et al.*, 2010). Com isso, VA tenta ampliar a capacidade humana de compreender e raciocinar sobre o mundo através de representações visuais (THOMAS; COOK, 2005).

A preocupação com a análise e identificação de *outliers* também está presente em diversos estudos de VA (NOVOTNY; HAUSER, 2006; WEBGA; LU, 2015). No entanto, existe um déficit de pesquisas que busquem avaliar a interpretação de anomalias do ponto de vista de OAM. Tal solução é relevante porque a simples identificação de falhas não é suficiente: também é necessário que a explicação fornecida pelas técnicas de OAM seja *apresentada* de forma adequada (função que normalmente cabe às técnicas de VA), para que especialistas possam tomar as atitudes necessárias e, se possível, prevenir a reincidência.

Uma plataforma de VA adequada ao contexto de OAM possui algumas especificidades em relação a uma ferramenta tradicional:

- Seleção e priorização de atributos: em OAM, a explicação para uma anomalia se dá através da seleção dos atributos mais relevantes para a explicação da anormalidade da amostra.
- Necessidade de exibição do componente temporal: a plataforma deve guiar o especialista em questões como "Quando o problema começou?"ou "Existiu algum evento que desencadeou o processo?"
- Preferência por técnicas que mantenham os atributos originais: a aplicação de técnicas como Principal Component Analysis (PCA) que constroem um novo espaço vetorial dificultariam a explicação das falhas por não representarem grandezas reais.

O objetivo desta pesquisa é oferecer uma ferramenta com as características acima.

Além disso, devido ao carácter incipiente dos estudos de OAM, a utilização de dados oriundos de aplicações reais ainda é rara. Por isso, este trabalho inova ao propor que, a partir de falhas conhecidas, técnicas de OAM sejam aplicadas para a explicação do *status* anômalo de amostras reais no contexto de locomotivas. Esta é uma aplicação interessante devido à complexidade dos sistemas, que possuem grande quantidade de atributos monitoráveis, e aos

diferentes fatores a que os equipamentos estão sujeitos (e.g. intempéries climáticas, variação de carga, operador).

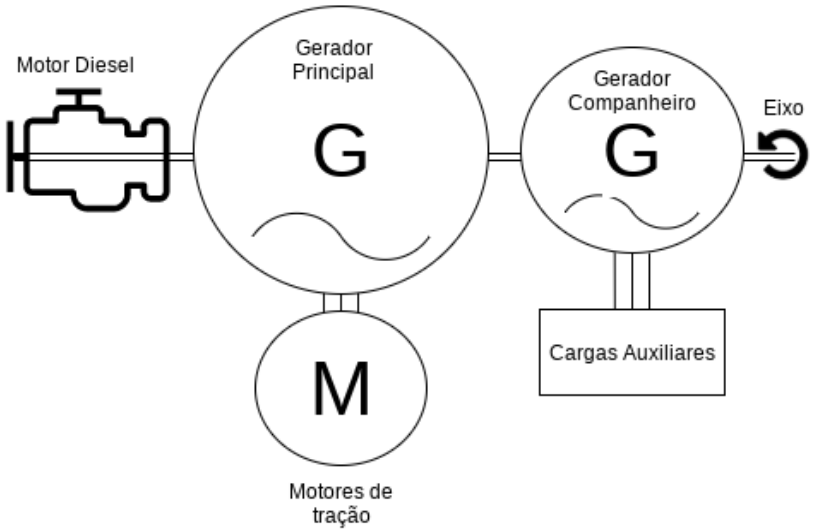
Este trabalho é organizado de forma a inicialmente apresentar uma breve descrição do funcionamento de locomotivas (Seção 2.1) e conceitos fundamentais como OAM (Seção 2.3) e *Visual Analytics* (Seção 2.4). Posteriormente, apresentam-se as experiências obtidas a partir da aplicação de OAM a dados operacionais reais (Capítulo 3) e a implementação de uma plataforma de VA adequada a OAM (Capítulo 4). Por fim, discutem-se os resultados e são propostas melhorias (Capítulo 5).

2 REVISÃO DA LITERATURA

2.1 LOCOMOTIVAS

Dentre os diferentes formatos de veículos que operam sobre trilhos (e.g. locomotivas, trens de alta velocidade, veículos leves sobre trilhos e trem-unidade elétrico), considera-se que as locomotivas demandam maior esforço de seus componentes tanto para aceleração quanto manutenção de altas velocidades (NATEGH S., 2018). Tal desempenho foi possibilitado pelos atuais modelos diesel-elétricos (HAPEMAN *et al.*, 1986), em que geradores movimentados por um motor diesel fornecem energia elétrica para tração e sistemas auxiliares (HAPEMAN *et al.*, 1986).

Figura 1 – Modelo de locomotiva diesel-elétrica.



Fonte: Hapeman *et al.* (1986)

A Figura 1 apresenta um diagrama de funcionamento de uma locomotiva diesel-elétrica. Nestes veículos, a fonte primária de energia é o diesel que abastece um motor com eixo acoplado a geradores. Os geradores, alternador principal e alternador companheiro, convertem a energia mecânica em energia elétrica AC trifásica. O gerador principal é maior, mais potente e alimenta os motores de tração. Por sua vez, o alternador companheiro, fornece energia para sistemas auxiliares como o pneumático (General Motors, 1978; General Motors, 1997).

As locomotivas modernas possuem microcontroladores embarcados que monitoram e controlam os diversos dispositivos que compõem o veículo (HAPEMAN *et al.*, 1986). São monitoráveis/controláveis tanto variáveis contínuas (e.g. rotação dos motores, excitação dos

campos dos geradores) quanto variáveis discretas como ativação/desativação de equipamentos por meio de contatores.

2.1.1 Falhas em locomotivas

Conforme expresso em Chandola *et al.* (2009), situações de falha em unidades industriais precisam ser identificadas o mais rápido possível e costumam ser consequência do desgaste causado pelo uso.

Para tanto, uma prática comum na indústria de locomotivas é a criação de regras manualmente por especialistas. Estas regras buscam detectar padrões relacionados a problemas nos equipamentos ou condições que possam diminuir a vida útil de componentes. No seu modo mais simples, esta análise é feita offline. A Tabela 1 exibe alguns exemplos destas regras que podem, por exemplo, monitorar se os geradores estão trabalhando dentro de intervalos adequados ou se o controle de temperatura e lubrificação do motor diesel está funcionando.

A especificação de regras apresenta diversas desvantagens, como custo de construção e manutenção. Além disso, rigidez e dificuldade na adaptação para novos contextos ou casos de uso são problemas frequentemente apontados para a utilização destas estruturas. Tais limitações levaram ao crescente interesse e aplicação das técnicas automáticas de detecção de *outliers* discutidas na próxima seção.

Tabela 1 – Exemplos de regras manuais para a detecção de falhas.

Regra
Temp. do óleo antes do arrefecimento $> X^oC$ AND Pressão do lubrificante $< X^oC$ AND Ponto == 8
Pressão da água $< X$ psi AND Ponto == 8
Corrente do Motor de tração 1 $> X\%$ da média dos outros motores
Motor diesel == Ligado AND (Tensão Gerador Companheiro $< X$ V OR Tensão Gerador Companheiro $> X$ V)

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

A descentralização (i.e. divisão em subsistemas) também tem se mostrado eficiente na detecção de elementos que demandem manutenção. Esta estratégia diminui o uso de redes de comunicação, melhora a relação custo-benefício e aumenta a conveniência para expansão (GAO *et al.*, 2015). Assim, a análise de trabalhos correlatos relacionados a falhas nos subsistemas de uma locomotiva mostra-se pertinente.

Devido à complexidade de suas estruturas e condições ambientais adversas de uso,

o diagnóstico de falhas em motores a diesel é desafiador, porém de grande interesse para indústria (ZHAO *et al.*, 2019). Isermann (2005) descreve um motor a diesel como a junção de três subsistemas: (i) sistema de admissão, (ii) injeção, (iii) combustão, virabrequim e sistema de exaustão de gás. O autor então propõe que a identificação de falhas seja feita com base no processamento de sinais de cada uma destas subpartes e na comparação com modelos de referência. Geng *et al.* (2003) sugerem um modelo analítico que, a partir da análise nos domínios de tempo e frequência das vibrações do motor, permita identificação de padrões de falhas. Zhao *et al.* (2019), por sua vez, obtiveram bons resultados com análise sonora, principalmente na identificação de defeitos relacionados a folgas nas válvulas dos motores.

Apesar de serem considerados robustos e confiáveis (SINGH; KAZZAZ, 2003), motores elétricos também estão sujeitos a falhas. Estas podem ser divididas de acordo com sua origem: elétricas, mecânicas ou ambientais. As primeiras referem-se a situações como transientes de tensão/corrente, falhas em dielétricos e tensão desbalanceada. Do ponto de vista mecânico, problemas podem ter como origem sobrecarga, montagem errada, falhas em rolamentos etc. Por sua vez, umidade, temperatura e limpeza são fatores ambientais que afetam a vida útil dos componentes.

Dentre as falhas mecânicas, problemas em rolamentos atraem bastante atenção da academia. Singh e Kazzaz (2003) citam diversos estudos feitos a partir da análise das vibrações para detectar tais falhas, principalmente com a análise de harmônicas. Este tipo de abordagem tem como desvantagem a necessidade de sensores adicionais, uma vez que medições de vibração normalmente não são usadas para o controle de motores elétricos. Neste sentido, Frosini e Bassi (2010) sugerem a utilização das medições de corrente do estator para a determinação de diferentes tipos de falhas. Os resultados obtidos mostraram que somente esta variável não é suficiente para gerar diagnósticos precisos em todos os casos analisados. Por exemplo, uma falha relacionada à deformação do selo do rolamento não seria identificada pelo método proposto pois o espectro da corrente não seria afetado. Para superar esta limitação, os autores então sugerem que o estudo seja complementado com a adição de dados de rotação do motor e torque. A primeira variável já está normalmente disponível nas máquinas, enquanto a segunda necessitaria de sensores adicionais.

Além dos tipos de falhas mencionadas em Singh e Kazzaz (2003), Najafabadi *et al.* (2011) defendem que a falha de sensores (principalmente de corrente) é comum em motores elétricos. Para tanto, Najafabadi *et al.* (2011) propõem um observador adaptativo capaz de

identificar se um (e não mais) dos sensores de tensão de link DC, correntes de fase e velocidade do motor está defeituoso. O observador cria um modelo dos fluxos do rotor, das correntes de fase e da resistência do rotor para estimar valores para os sensores. Quando às diferenças entre valores estimados e medidos forem substanciais, diagnostica-se uma falha.

Alguns tipos de aeronaves também possuem geradores trifásicos sem escovas (semelhantes aos usados na ferrovia). Assim, a análise realizada em Batzel e Swanson (2009) se mostra relevante a este trabalho. No estudo, os autores identificaram curto-circuitos no enrolamento de campo e falhas em diodos retificadores como os defeitos mais comuns nos geradores utilizados na aviação. A partir da função de transferência, estes tipos de falhas puderam ser previstos e um filtro de Kalman foi utilizado para determinar o tempo para a substituição dos componentes.

A revisão bibliográfica de trabalhos relacionados à detecção de falhas nos diversos dispositivos que compõem uma locomotiva evidenciou uma preferência da academia por abordagens baseadas em modelos matemáticos. No entanto, conforme apontado em Mao *et al.* (2017), a crescente complexidade dos sistemas, a dificuldade para se representar situações incomuns como ruídos/distúrbios e as diferentes condições de operações tornam este tipo de abordagem bastante complexa. Por exemplo, a descrição da dinâmica das forças que atuam sobre um vagão convencional envolveria uma equação diferencial de 84^a ordem (GOODALL; KORTüM, 2002). Neste sentido, estudos a partir de dados históricos podem ser empregados como forma a simplificar a abordagem. Contudo, como este tipo de estratégia normalmente demanda muitos recursos computacionais, não se costuma aplicá-las em sistemas embarcados na ferrovia (GARRAMIOLA *et al.*, 2018) mas sim de forma remota (XUE *et al.*, 2006).

Luwei *et al.* (2018), por exemplo, utilizam dados no domínio da frequência obtidos de diferentes sensores como entrada para redes neurais. Uma primeira rede neural é responsável por indicar a existência de alguma falha e cabe a uma segunda rede a identificação do tipo de anomalia.

Chen *et al.* (2018) defendem que a distância entre as funções de densidade de probabilidade obtidas a partir do cálculo da distância de Hellinger seja aplicada para detectar falhas nos sensores do sistema de tração de trens de alta velocidade. Diferente de outros trabalhos, o algoritmo proposto é capaz de diagnosticar falhas em tempo real fazendo uso de inferências de Bayes.

Conforme indicado em Garramiola *et al.* (2018), publicações com dados reais do sistema ferroviário são raras. Na maior parte dos casos, os estudos são feitos com base em simulações ou

em laboratórios equipados com alguns dos dispositivos usados na vida real, como em Chen *et al.* (2018). Neste sentido, Xue *et al.* (2006) apresentam um trabalho diferenciado uma vez que empregam dados reais obtidos de locomotivas da *General Electric*.

Xue *et al.* (2006) empregam Wilcoxon rank sum test (WRST) para determinar se uma amostra desviou da condição adequada à operação. WRST é um teste estatístico não paramétrico para execução de testes de hipótese entre duas amostras, uma reconhecidamente normal (*inlier*) e outra indicando o estado atual da locomotiva. Os resultados fornecidos pelo WRST são usados como entrada para uma rede neural de regressão generalizada, que fornece o resultado final: amostra é uma anomalia ou não. Para validação do método proposto, focou-se no sistema de turbo do motor a diesel. Foram empregadas nove variáveis, no entanto devido a questões proprietárias, nomes e descrições destes atributos foram omitidos. A rede neural conseguiu identificar corretamente todas as 579 condições de falhas analisadas.

Conforme esta seção demonstrou, pesquisas destinadas à detecção de falhas costumam ser baseadas em modelos específicos para os diversos componentes que formam o sistema de uma locomotiva. Mesmo nos poucos casos em que algoritmos de aprendizado de máquina são aplicados, os dados raramente são reais. Não obstante o trabalho de Xue *et al.* (2006) ser uma exceção por utilizar dados verídicos, os autores consideraram somente nove atributos para sua análise de anomalias no sistema turbo do motor diesel. Neste sentido, o presente trabalho supre lacunas na bibliografia uma vez que propõe-se uma análise da locomotiva como um todo, ou seja, não se limitando a alguns componentes e utilizando-se de dados reais. Devido à complexidade, tal abordagem seria inviável se fosse baseada em modelos, . Assim, a análise a partir de dados históricos faz-se necessária.

2.2 OUTLIERS

Outliers ou anomalias são definidos como amostras que destoam do esperado, i.e. estão tão distantes das demais observações em um conjunto de dados (*inliers*) a ponto de ser razoável supor que foram originadas por um processo distinto ou que seguem uma lógica própria (HAWKINS, 1980; CIELEN *et al.*, 2016).

Dependendo do problema considerado, *outliers* podem representar um obstáculo ou o objetivo final do estudo. No primeiro caso, as anomalias estão fora da zona de interesse do analista e devem ser removidas da base de dados, pois seus valores extremos podem distorcer medidas (CHANDOLA *et al.*, 2009; CIELEN *et al.*, 2016). No segundo caso, os diferentes

padrões de *outliers* fornecem informações críticas a diversas áreas como detecção de fraudes, identificação de invasores ou diagnóstico de tumores (CHANDOLA *et al.*, 2009). Para este trabalho considera-se o segundo caso, pois espera-se que diagnósticos relevantes sobre falhas em locomotivas sejam extraídos a partir da identificação de *outliers*.

O estudo de técnicas para a detecção de *outliers* é um campo bastante amplo e difundido por diversos domínios e os resultados fornecidos tipicamente consistem em uma classificação binária (*outlier* ou *inlier*). Em outros casos, tem-se uma pontuação (quanto maior/menor o valor desta medida, mais distante/próximo é uma amostra das demais), este é o caso por exemplo das técnicas *Local Outlier Factor*(BREUNIG *et al.*, 2000) e *Isolation Forest* (LIU *et al.*, 2008)

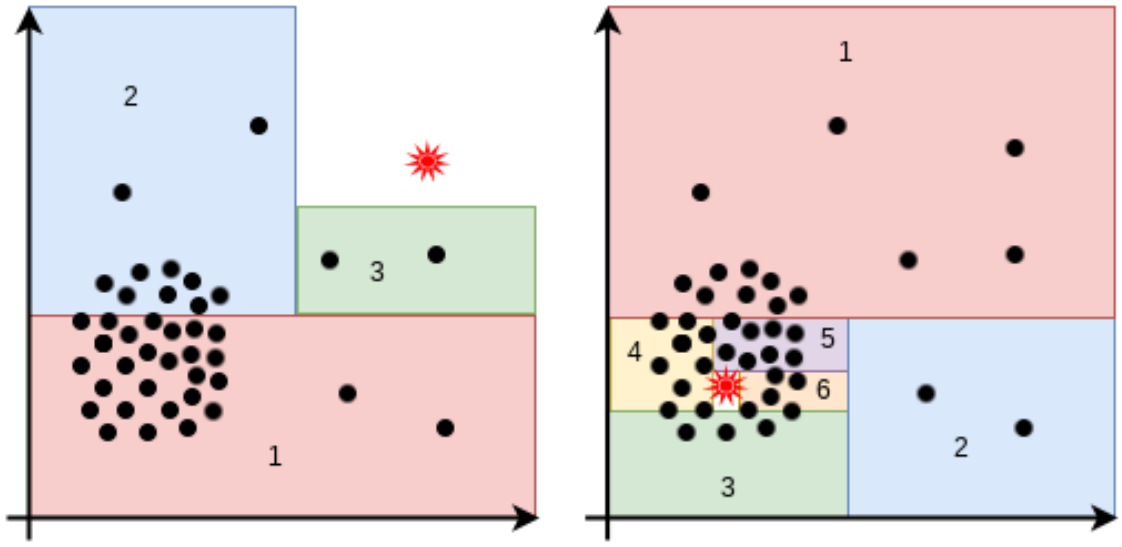
Local Outlier Factor (LOF) (BREUNIG *et al.*, 2000) adota o conceito de densidade local para determinar anomalias: pontos com densidade muito menor que seus vizinhos são considerados anômalos. Isto garante que elementos dentro ou na parte externa de *clusters* obtenham pontuações próximas a 1 (*inliers*) independentemente da dimensionalidade.

Para Isolation Forest (iForest) (LIU *et al.*, 2008), constroem-se árvores binárias dividindo-se aleatoriamente o conjunto de dados até que cada uma das amostras seja isolada. O número de divisões necessário para isolar cada elemento serve de base para o cálculo do fator de anormalidade. O particionamento é realizado sorteando-se um valor dentre o intervalo possível para um atributo: objetos com valores menores são direcionados para um ramo da árvore e os demais são direcionados para outro. Assume-se que anomalias estão presentes em menor quantidade e que apresentam características diferentes de elementos normais, portanto, podem ser isoladas com menos divisões. Quando diversas árvores são construídas (no estudo utilizam-se 100 árvores), o tamanho do caminho ou número de partições para que um mesmo objeto seja isolado converge para um valor específico. Liu *et al.* (2008) também propõem a subamostragem dos dados, no caso, cada árvore utilizava somente 256 elementos, independentemente do tamanho da população considerada.

A Figura 2 demonstra como o isolamento de *outliers* (a) tende a ser mais curto se comparado a pontos normais (b), justificando a heurística empregada no método *isolation path*. Na imagem, a amostra sob análise é representada em vermelho e diferentes cores são utilizadas para evidenciar as regiões descartadas a cada nova partição. Assim, para se isolar o elemento *outlier* em (a) seriam necessárias três partições e, sete interações para segregar o *inlier* em (b).

Desde a publicação em Liu *et al.* (2008) outros estudos buscaram aplicar o método proposto. Em Puggini e McLoone (2018) *isolation forest* foi usado na detecção de anomalias em

Figura 2 – Número de ramificações para isolar (a) *outlier* e (b) *inlier*.



Fonte: Liu *et al.* (2008)

dados de Espectroscopia de Emissão Óptica obtidos a partir da fabricação de semicondutores. Susto *et al.* (2017) aplicam o algoritmo para a identificação de anomalias em dados de corrosão por plasma. Em aplicações industriais, como na calibração de equipamentos, muitas vezes é importante encontrar uma amostragem relevante da população, i.e. sem *outliers*, por isso, Chen *et al.* (2016) utilizam iForest com o objetivo de a remover anomalias. Scherman e Bülow (2018) conseguiram identificar com sucesso 76% dos casos de ataques internos a uma rede empresarial (realizados por funcionários da própria empresa), com somente 7% de alarmes falsos. Em um *benchmark* realizado com diversos métodos e dados reais, *isolation forest* obteve o melhor desempenho junto com o método Ensemble Gaussian Misture Model (GLODEK *et al.*, 2013), superando SCiForest (LIU *et al.*, 2010), LOF (BREUNIG *et al.*, 2000) e Support Vetor Data (TAX; DUIN, 2004).

Em pesquisas subsequentes foi identificado que o desempenho do iForest não é adequado à localização de *outliers* locais (VINH *et al.*, 2016) e que seu desempenho é afetado por fatores como distribuição dos dados (GUHA *et al.*, 2016; HARIRI; KIND, 2018), alta correlação entre variáveis (PUGGINI; MCLOONE, 2018) e alta densidade de *clusters* anômalos (LIU *et al.*, 2008). Estas situações podem resultar em alta instabilidade nos resultados (CHEN *et al.*, 2013) e muitos alarmes falsos (GUHA *et al.*, 2016).

2.3 OUTLYING ASPECT MINING

A *interpretação* de *outliers* é formalmente estabelecida no contexto de OAM em Liu *et al.* (2018): "a partir de um conjunto de dados $\mathcal{X}$ e os *outliers* $\mathcal{O}$ identificados, a interpretação de cada *outlier* $\mathbf{o}_i \in \mathcal{O}$ é definida pelo conjunto $\mathcal{E}_i = \{ \mathcal{A}_i, \lceil(\mathbf{o}_i), \mathcal{C}_i = \{\mathcal{C}_{i,l} | l \in [1,L]\}\}$. Em que $\mathcal{C}_i$ é o contexto de $\mathbf{o}_i$, $\mathcal{A}_i$ inclui os atributos anormais de $\mathbf{o}_i$ em relação a $\mathcal{C}_i$ e $\lceil(\mathbf{o}_i)$ é a pontuação para a anormalidade do *outlier*". O contexto $\mathcal{C}_i$ do outlier $\mathbf{o}_i$ por sua vez, é definido pelas $\mathbf{k}$ instâncias normais mais próximas de $\mathbf{o}_i$ e $\mathcal{C}_i$ pode ser composto por grupos menores (*clusters*) $\mathcal{C}_{i,1}, \mathcal{C}_{i,2}, ..., \mathcal{C}_{i,L}$.

Assim, considera-se em OAM que uma explicação para a anormalidade de uma amostra é fornecida pelo subespaço, i.e. conjunto de atributos, que tenha obtido o primeiro lugar para uma dada métrica de anormalidade dentre todos os subespaços possíveis. Isto impede que diversas técnicas tradicionais de mineração de dados sejam aplicáveis ao contexto de OAM, seja por limitações conceituais ou mesmo a capacidade computacional disponível.

Do ponto de vista conceitual, diversas técnicas não são adequadas por não manterem as mesmas propriedades matemáticas em subespaços com diferentes dimensionalidades. Por sua vez, limitações computacionais existem porque alguns métodos precisam calcular um fator de anormalidade (*outlierness score*) para todas as combinações de subespaços possíveis. Conforme indicado em Duan *et al.* (2015), um conjunto de dados com 100 dimensões resultaria em uma combinação de $1{,}27 \times 10^{30}$ subespaços, quantidade proibitiva.

Quanto aos grupos $\mathcal{C}_{i,l}$ que compõem o contexto $\mathcal{C}_i$ para o *outlier* $\mathbf{o}_i$, justifica-se esta distinção em *clusters* porque a interpretação de um evento raro não deveria somente avaliar os pontos mais próximos dos objetos de interesse. Caberia também a técnica de OAM comparar o *outlier* ao grupo de dados a que ele pertence em detrimento de uma análise global, levando em consideração as características específicas deste *cluster*. Apesar desta definição formal, a maior parte dos trabalhos em OAM não leva em conta esta divisão de *clusters* na análise. Da mesma forma, no presente trabalho também foi considera que os dados pertenciam a um único grupo.

Dentre as técnicas de OAM, existem dois grupos principais: seleção de atributos e pontuação-e-procura (MICENKOVÁ *et al.*, 2013; VINH *et al.*, 2016). Enquanto a primeira estratégia baseia-se em métodos para a seleção não-supervisionada de atributos conforme já aplicado em outros contextos (como identificação de *outliers*), a segunda busca definir métricas para medição da anormalidade dos objetos e, em seguida, aplicá-las nas diversas combinações de

subespaços. Os subespaços melhor classificados fornecem então a interpretação da anormalidade dos objetos.

Considerado por Vinh *et al.* (2016) como o primeiro método de seleção de atributos aplicado a OAM, Micenková *et al.* (2013) utilizam as técnicas Support Vector Machine (SVM) (CORTES; VAPNIK, 1995) e lasso (TIBSHIRANI, 1996) para incorporar classes consideradas relevantes para a compreensão de anomalias a um subespaço. Os autores salientam que diferentes algoritmos poderiam ter sido utilizados, desde que a estratégia para a inclusão de atributos fosse mantida. Esta estratégia busca enfatizar a vizinhança de uma anomalia no espaço completo e realiza a subamostragem nas demais dimensionalidades.

Dentre as principais vantagens apontadas para esta abordagem baseada na seleção de atributos destaca-se a rapidez, uma vez que os algoritmos propostos não costumam considerar todas as combinações de subespaços possíveis (VINH *et al.*, 2016). Quando às desvantagens, salienta-se o modo *caixa-preta* de algumas técnicas e a ausência de uma lista com os subespaços melhor avaliados (afinal nem todas as dimensionalidades são consideradas) (VINH *et al.*, 2016).

Conforme definido em Micenková *et al.* (2013), algoritmos de pontuação-e-procura buscam determinar uma função que atribua a um objeto uma pontuação que sirva de métrica para o grau de anormalidade deste ponto em relação aos demais elementos. Pontos que divirjam mais devem receber valores mais altos, enquanto *inliers* recebem valores mais baixos.

Apesar de ter sido idealizado para a detecção de *outliers*, LOF (BREUNIG *et al.*, 2000) é uma métrica de pontuação bastante usada no contexto de OAM, inclusive como algoritmo base para comparação na proposição de novos métodos (VINH *et al.*, 2016; MICENKOVÁ *et al.*, 2013; DANG *et al.*, 2014; ANGIULLI *et al.*, 2017). Segundo Vinh *et al.* (2016), LOF possui diversas características importantes no âmbito do OAM, como dimensionalidade pouco influente nos resultados e a possibilidade de se calcular a pontuação para somente o objeto de interesse.

Dang *et al.* (2014) propõem a construção de um grafo para a interpretação das características determinantes de *outliers*. Neste grafo, vizinhanças são modeladas com vértices representando objetos e a ligação entre elementos se dá quando estes forem vizinhos mais próximos. Um peso para a conexão é atribuída com base na semelhança entre os objetos.

Após a construção do grafo inicial, noções de grafo Laplaciano são aplicadas para diminuir a dimensionalidade. No novo espaço gerado, as características do local *outliers* são mantidas, i.e. cada item é avaliado em relação a uma subpopulação de vizinhos e um fator de anormalidade do elemento pode ser calculado.

Duan *et al.* (2015) propõem uma estratégia a partir da estimativa da densidade do kernel (SCOTT, 2015). Neste algoritmo, OAM representa a busca pelo *subespaço mínimo do outlier*, i.e. a procura pelo subespaço com o menor combinação de atributos em que o objeto de interesse teve a melhor posição no ranking. Formalmente, dado um elemento q e um espaço multidimensional $\mathcal{X}$, um subespaço $\mathcal{S}_1 \subseteq \mathcal{X}$ é chamado de subespaço mínimo do *outlier* se:

1. Não existir um outro subespaço $\mathcal{S}_2 \subseteq \mathcal{X}(\mathcal{S}_2 \neq \varnothing)$, tal que $rank_{\mathcal{S}_2}(q) < rank_{\mathcal{S}_1}(q)$.
2. Não existir um outro subespaço $\mathcal{S}_3 \subset \mathcal{S}_1$, tal que $rank_{\mathcal{S}_3}(q) = rank_{\mathcal{S}_1}(q)$.

Isolation Path (iPath) (VINH *et al.*, 2016) é um algoritmo de pontuação-e-procura para OAM desenvolvido como uma adaptação ao método *isolation forest* (LIU *et al.*, 2008), originalmente idealizado para a detecção de anomalias. Em comum, ambas as técnicas constroem árvores binárias dividindo aleatoriamente o conjunto de dados até que a amostra de interesse seja isolada dos demais elementos.

Uma das diferenças entre *isolation forest*, desenvolvido para a detecção de *outliers*, e *isolation path*, elaborado para OAM, é a necessidade ou não de se isolar todos os pontos no conjunto de dados. Como em OAM busca-se uma explicação para a anormalidade de um ponto específico, somente o caminho até este elemento precisa ser calculado e o restante dos ramos gerados a cada partição pode ser descartado. O mesmo não pode ocorrer na aplicação de *isolation forest*, uma vez que a pontuação de todos os objetos é necessária para que, por comparação, os *outliers* possam ser determinados.

Conforme salientado por Liu *et al.* (2008) no contexto do *isolation forest* e mantido para o *isolation path*, árvores binárias dispensam cálculos custosos comuns a métodos populares, como medidas de distância e densidade. Neste sentido, a complexidade média esperada para *isolation path* é $O(n_S)$ em que n_S é uma subamostra do conjunto de dados e o pior cenário consiste em $O(n_S^2)$. No primeiro caso, considera-se que, em média, cada particionamento irá dividir os dados restantes ao meio. No segundo caso, tem-se que cada particionamento irá remover somente um elemento do conjunto a cada iteração.

2.3.1 Desafios na avaliação de desempenho

Uma análise da bibliografia demonstrou que ainda não há consenso quanto a métricas para avaliação de desempenho de algoritmos de OAM, principalmente em se tratando de dados reais. Concordamos que o estabelecimento de tais métricas pode ser de fato complicado pois

seria, por exemplo, necessária uma explicação/interpretação dos atributos diferentes/anômalos de todos os pontos, não só dos *outliers* em um conjunto de dados, dado que OAM pode ser aplicado a qualquer ponto.

Além disso, tal interpretação é usualmente ser fornecida por especialistas, o que pode tornar a análise tendenciosa. Duan *et al.* (2015), por exemplo, observaram esta parcialidade ao obter resultados surpreendentes em um estudo sobre jogadores de basquete da NBA. Na investigação, os atributos apontados como relevantes pelo algoritmo de OAM empregado não indicavam a qualidade do jogador, somente anormalidade. Assim, muitas desses critérios eram desconsiderados por comentaristas esportivos nas avaliações de desempenho.

Devido a estas dificuldades, muitas vezes os algoritmos são testados em dados sintéticos. Neste sentido, o conjunto de dados criado por Keller *et al.* (2012) foi reutilizado por Vinh *et al.* (2014), Duan *et al.* (2015), Vinh *et al.* (2016) e serviu de base para Liu *et al.* (2018). No conjunto proposto, são gerados *clusters* de alta densidade com 2 a 5 atributos. Nestes subespaços, são então distorcidos cinco objetos de modo que eles sejam significativamente diferentes dos grupos originais. Ademais, garantiu-se que estes *outliers* fossem não-triviais, i.e. não visíveis a partir de projeções de somente um atributo. Os algoritmos sob avaliação devem então ser capazes de identificar os subespaços modificados ao buscarem por uma explicação para os *outliers*.

Uma estratégia diferente adotada por Liu *et al.* (2018) foi adicionar dimensionalidades sem nenhuma informação relevante (ruído) a um conjunto de dados reais. Coube ao método testado diferenciar os novos atributos dos atributos originais. O desempenho foi então apurado com Precisão, *Recall* e *F1-Score*. O problema deste tipo de análise é considerar que todas as variáveis originais contribuíam de fato para a interpretação de anormalidade dos elementos de forma equivalente.

Além de utilizaram ROC e AUC assim como em Dang *et al.* (2014), Micenková *et al.* (2013) também propuseram duas alternativas para avaliação de resultados em dados reais. Primeiramente, geraram uma classificação a partir de LOF e compararam este resultado com o ordenamento gerado pelas técnicas sob análise. Ao aplicar esta estratégia a dados mais complexos, no entanto, notou-se uma piora geral no desempenho de todos os métodos comparados. Com isso, os resultados foram dados como inconclusivos, pois não era possível definir se os algoritmos sob análise de fato pioraram ou se LOF não era mais adequado ao novo conjunto de dados.

Na segunda estratégia sugerida por Micenková *et al.* (2013), os autores primeiramente agruparam *outliers* em *clusters* usando LARS-lasso. Sob a perspectiva de que elementos de um

mesmo *cluster* devem apresentar características parecidas, comparou-se a similaridade entre subespaços apontados como significativos para objetos pertencentes a um mesmo *cluster*.

Uma metodologia semelhante a Micenková *et al.* (2013) foi adotada em Vinh *et al.* (2016). Neste trabalho, os autores propõem o *Consensus Index*, uma técnica que também se baseia na hipótese de que elementos de uma mesma classe devem ter características relevantes semelhantes. Para cada classe faz-se uma votação com os atributos apontados como mais significativos e mede-se entropia/semelhança dentre os atributos votados para cada classe. Observamos que, tanto para Micenková *et al.* (2013) como para Vinh *et al.* (2016), a divisão entre classes/*clusters* tem potencial de influenciar muito nos resultados.

Com base neste trabalhos, observamos que a diversidade de estratégias para avaliação de desempenho atrapalha o *bechmark* das soluções e ainda é um campo aberto para pesquisa futuras. Outra questão que precisa ser considerada é a necessidade da criação de tabelas da verdade não tendenciosas para dados reais e que contemplem não somente *outliers*, mas também amostras "normais". Estas limitações no entanto não afetam a presente pesquisa porque nela OAM será empregado com base em um modelo já existente e os resultados obtidos serão avaliados frente a sua capacidade de fornecer subsídios à tomada de decisões por especialistas.

2.4 VISUAL ANALYTIC E PARALLEL COORDINATE PLOTS

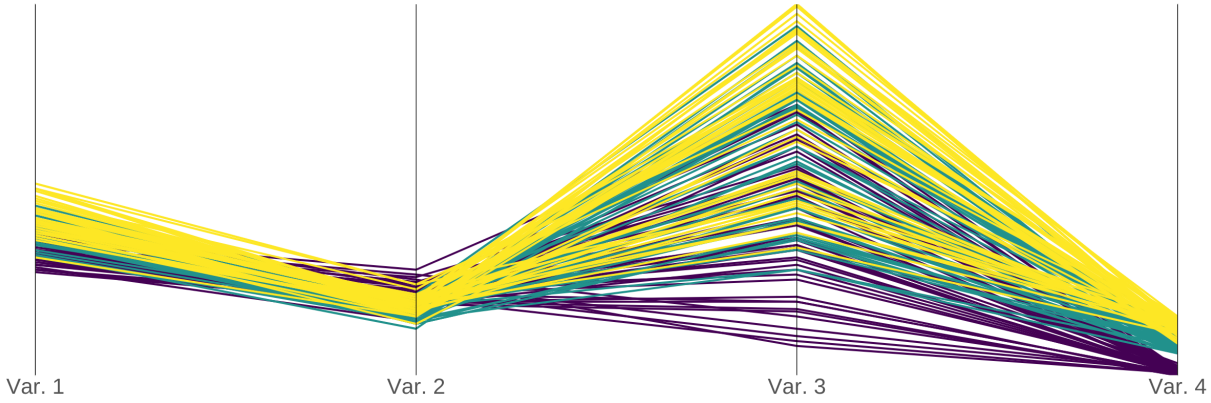
Mesmo com o aumento da disponibilidade dos dados e de técnicas para extração de conhecimento, a obtenção de modelos causais de forma automática continua sendo um desafio em sistemas complexos. Por isso, habilidades comuns aos seres humanos como criatividade, conhecimento teórico e flexibilidade ainda são fundamentais. No entanto, a quantidade de amostras e variáveis que uma pessoa consegue processar é limitada e, por isso, os elementos precisam ser apresentados de forma adequada à capacidade cognitiva humana.

Neste sentido, para Chen *et al.* (2011) a Análise Visual (*Visual Analytics*) procura, através do uso de interface visual, criar um ambiente propício para a interação humana com os dados e, assim, gerar análises mais aprofundadas. A Análise Visual é mais ampla que a mera visualização das informações e engloba também transformação e interação com os dados.

Dentre as técnicas de visualização de dados, Parallel Coordinate Plot (PCP) ou Gráficos de Coordenadas Paralelas (INSELBERG, 1985) têm sido muito explorados para a representação de dados multidimensionais. Esta técnica na sua forma clássica, conforme mostrado na Figura 3, utiliza eixos paralelos verticais (opcionalmente horizontais) para representar as dimensões e

polilinhas para indicar as amostras. O ponto em que uma polilinha intersecta um eixo reflete o valor daquela amostra para o respectivo atributo. Cores diferentes também podem ser atribuídas às polilinhas para salientar grupos ou mesmo amostras individuais. Dentre as vantagens desta forma de visualização, salienta-se a possibilidade de representação teoricamente ilimitada de dimensões e a exibição da geometria das amostras em vez de somente as pontos (Blaas *et al.*, 2008). Isto torna PCPs adequados a OAM.

Figura 3 – Exemplo de gráfico de coordenadas paralela (PCP).



Fonte: O Autor

Nestes mais de trinta anos desde a concepção, diversas variações de PCPs foram propostas com o objetivo de agregar elementos à forma clássica e facilitar a interpretação. Por exemplo, as polilinhas já foram substituídas por curvas (interpolações), representações de densidades e polígonos (HEINRICH; WEISKOPF, 2013; JOHANSSON *et al.*, 2007). Mais recentemente, imagens em 3D começaram a ganhar destaque, com (TADEJA *et al.*, 2019) ou sem (ZHONGHUA; LINGDA, 2016) realidade virtual. Além disso, as aplicações se estendem por diversos setores, como a análise de dados de sensores para atividade física (TONG *et al.*, 2019), exploração de expressão gênica (DIETZSCH *et al.*, 2009) ou eventos de microssísmicos (MOSTAFA *et al.*, 2012).

Um dos problemas mais comuns nos gráficos de coordenadas paralelas é o excesso de polilinhas, comumente chamado de *overplotting*. Isto ocorre porque cada amostra em um conjunto de dados teoricamente pode ser representado por uma polilinha. No entanto, uma grande quantidade destes elementos pode tornar qualquer interpretação inviável, dada a confusão de formas.

A seleção de subconjuntos de dados, ou *brushing*, é uma das formas mais comuns e simples de enfrentar esta limitação. Normalmente isto é feito por meio de plataformas interativas, em que os usuários podem salientar elementos ou mesmo criar novas imagens a partir de

subamostras de interesse (Blaas *et al.*, 2008; SANSEN *et al.*, 2017a; TONG *et al.*, 2019). No entanto, este tipo de análise é pouco escalável e requer conhecimento prévio dos dados.

Outra alternativa é a introdução de técnicas de clusterização. Neste âmbito, Novotny e Hauser (2006) aplicam *binning* aos dados *inliers* ao mesmo tempo que preservam *outliers* em sua forma original. Sansen *et al.* (2017b) vão mais além ao exibir no próprio PCP elementos visuais que indicam a distribuição de amostras e intervalos em um mesmo cluster.

Quanto à preocupação com a representação de *outliers*, pode-se dizer que ela é bastante recorrente em PCPs. Em diversos trabalhos, anomalias são exibidas de forma diferenciada, o que graficamente se materializa através de cores chamativas ou texturas próprias (Blaas *et al.*, 2008; NOVOTNY; HAUSER, 2006; GLENDENNING *et al.*, 2016).

Por sua vez, adaptações aos gráficos de coordenadas paralelas que possibilitem a compreensão da evolução temporal dos elementos é mais rara, sendo portanto ainda um campo em aberto. Johansson *et al.* (2007), por exemplo, propõem duas técnicas para a visualização de séries temporais com PCPs. Na primeira, *Temporal Density Parallel Coordinates*, as polilinhas são substituídas por polígonos, que de acordo com sua tonalidade comunicam a densidade dos valores através do período considerado, i.e. regiões com alta reincidência de valores através do tempo são representadas de forma mais vívida no gráfico. Na segunda técnica, *Depth Cue Parallel Coordinates*, um gradiente de cores é usado para indicar a diferença temporal das amostras, assim, amostras mais recentes são exibidas com cores mais fortes e aparentam estar no primeiro plano.

Diferentemente, Blaas *et al.* (2008) não exibem dados coletados em instantes distintos em um mesmo PCP. Entretanto, reconhece a relevância desta análise e por isso possibilita que o usuário de sua interface interativa avance/regrida no tempo deslizando um marcador. Similarmente, em Barlow e Stuart (2004) permite-se que sejam escolhidos os instantes para os quais PCP é exibido ou mesmo que seja observada a evolução de valores por meio de uma animação.

Zhonghua e Lingda (2016) exibem simultaneamente múltiplos PCPs referentes a instantes diferentes por meio de visualizações 3D. Para tanto, os PCPs são alinhados paralelamente ao ponto de vista do observador, com os gráficos que ilustram momentos mais antigas exibidos mais distantes e com cores mais transparentes. Os eixos para cada atributo foram substituídos por planos que permitem a visualização da transformação dos valores para cada atributo com o tempo. Na opinião dos próprios autores, o excesso de elementos gráficos e sua consequente

confusão de formas, dificultam a interpretação e obtenção de análises relevantes na solução proposta.

Mostafa *et al.* (2012) utilizam um dos eixos dos PCPs para exibição da variável temporal, juntamente dos outros atributos. Lee e Shen (2009) também propõem PCPs para a representação de tendências temporais. Para isso, modificam a ideia original de PCP mais drasticamente ao dividir o conjunto de dados em estados, sendo cada estado representado nos eixos paralelos. A inclinação das linha entre os eixos serve de indicação do tempo que cada estado persistiu.

Mesmo com esta grande diversidade de topologias e aplicações de PCPs apresentados, desconhecemos outros trabalhos que proponham a representação de séries temporais juntamente com anomalias em uma único gráfico de coordenadas paralelas. Além disso, esta contribuição mostra-se relevante no contexto de OAM uma vez que, a partir de uma perspectiva multidimensional, permitiria a análise da degradação temporal do sistema resultando em uma falha. Outro diferencial desta proposta é a integração de PCP em um contexto de *Visual Analytics*, permitindo que o especialista interaja com a interface e alterne entre PCP e gráficos de séries temporais tradicionais.

3 EXPERIÊNCIAS PRÁTICAS NA APLICAÇÃO DE MÉTODOS DE CONTEXTUALIZAÇÃO DE ANOMALIAS EM DADOS OPERACIONAIS DE LOCOMOTIVAS

Dado o carácter incipiente dos estudos de OAM, a utilização destas técnicas em aplicações reais ainda é rara. Por isso, o experimento descrito nesta seção, em que uma técnica de OAM, *isolation path*, é empregada na interpretação de falhas em locomotivas mostra-se relevante.

De forma a avaliar os resultados gerados, decidiu-se pela utilização de anomalias previamente identificadas com base em regras elaboradas por especialistas (fornecida em conjunto com a base de dados de testes), cabendo à técnica de OAM identificar os atributos empregados na composição das regras. Por exemplo, uma falha de lubrificação do motor diesel ocorre quando a pressão do lubrificante estiver abaixo de um valor mínimo em situações de alta temperatura do óleo e alta potência. Desta forma, para uma amostra classificada como anômala em relação a estes critérios, o algoritmo de OAM deve indicar os atributos *pressão do lubrificante*, *temperatura do óleo antes do arrefecimento* e *potência da máquina* como relevantes para a interpretação da falha.

3.1 OBTENÇÃO E TRATAMENTO DOS DADOS

Os dados foram obtidos entre os anos de 2017 e 2018, a partir de uma frota de 120 máquinas de um mesmo modelo de locomotiva. Cada máquina da frota gera novos registros a cada um segundo. No entanto, o envio aos servidores ocorre em bateladas a cada 1,5 horas, o que impede uma análise em tempo real dos eventos.

Quanto à escolha de eventos para análise, buscou-se por um tipo de falha específico e que apresentasse uma reincidência mínima em locomotivas diferentes. Além disso, verificações consideradas muito simples (e.g. dependentes de somente um atributo) não justificavam uma análise de especialistas e, por isso, foram descartadas. Com base nestes critérios, o estudo foi resumido a eventos de problema de lubrificação do motor diesel, com vinte amostras selecionadas: dez eventos de falha e dez amostras normais.

Inicialmente considerou-se utilizar os 136 atributos disponíveis para o teste (66 binários e 70 contínuos). No entanto, testes iniciais demostraram a impossibilidade desta estratégia, tanto do ponto de vista computacional quanto pelo resultado fornecido.

Tabela 2 – Variáveis da locomotiva analisadas ou controladas.

Equipamento	Variável contínua
Motor diesel	Temperatura óleo antes do arrefecimento (TOA) (^{o}C) Temperatura óleo depois do arrefecimento (TOD) (^{o}C) Temperatura da água (TH2O) (^{o}C) Pressão lubrificante (PLUB) (psi) Pressão combustível (PCOMB) (psi) Rotação (ROT) (rpm) / Potência (POT) (HP)
Gerador Principal	Tensão Gerador Principal (VGP) (V) Corrente Gerador Principal (IGP) (A)
Gerador Companheiro	Corrente Gerador Principal (VGA) (V) Corrente Gerador Auxiliar (IGA) (A)
Motores de tração	Corrente Motor de tração (MT1) (A)
Geral	Velocidade de deslocamento (GPSVEL) (km/h) Ponto de operação (PTO)

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

Primeiramente, esta quantidade de atributos geraria uma combinação de 419.356 subespaços. Dado que o tempo médio de cálculo da métrica para um único subespaço é de 3,69 segundos em um computador Intel(R) Core i7-2720QM CPU @2.20 GHz, seriam necessários quase 18 dias para o processamento de todas as combinações de subespaços possíveis para uma única amostra. Mesmo com a utilização dos oito núcleos disponíveis no processador, o tempo necessário para construção de uma amostragem estatisticamente significante inviabilizaria a pesquisa.

Além disso, também foi verificado que o desempenho do *isolation path* frente a variáveis binárias é bastante ruim. Esta degradação já era esperada porque somente uma interação é possível com atributos deste tipo, o que distorce as medidas. Por exemplo, para um subespaço com 3 variáveis binárias, existem apenas 8 ramificações possíveis para a árvore construída e a amostra analisada será impreterivelmente isolada em 3 divisões. Desta forma, a comparação entre as métricas calculadas para variáveis de diferentes tipos seria tendenciosa, dado que o número de divisões possíveis para variáveis contínuas é teoricamente ilimitado. Ademais, assumindo-se um distribuição uniforme para este subespaço com três atributos binários, ao final das divisões o objeto de interesse ainda estaria junto de 12,5% das amostras, quantidade muito superior à ocorrência de qualquer falha.

Por isso, decidiu-se pela remoção de todos os atributos binários da análise. Como os sistemas críticos de uma locomotiva também possuem variáveis contínuas para seu monitoramento, os estados indicados pelos atributos binários continuam representados. Por exemplo, o acionamento do motor diesel pode ser deduzido tanto por variável binária específica para este fim, quanto pela medida de rotação do motor. Por isso, a utilização de somente atributos contínuos

Tabela 3 – Correlação entre atributos

	PLUB	MT1	POT	PCOMB	PTO	VGA	VGP	TH2O	TOD	TOA	GPSVEL	IGA	IGP	ROT
PLUB	1	0.64	0.17	0.19	0.83	0.02	0.07	0.03	0.11	0.02	0.48	-0.75	-0.07	0.34
MT1	0.64	1	0.34	-0.04	0.63	0.02	0.07	0.39	0.46	0.36	0.26	-0.6	0.32	0.27
POT	0.17	0.34	1	-0.72	0.2	0	0.9	0.51	0.61	0.57	0.22	-0.37	0.86	0.21
PCOMB	0.19	-0.04	-0.72	1	0.07	0.01	-0.69	-0.38	-0.43	-0.6	-0.04	-0.07	-0.66	-0.5
PTO	0.83	0.63	0.2	0.07	1	0.04	0.03	0.32	0.43	0.37	0.48	-0.76	-0.09	0.48
VGA	0.02	0.02	0	0.01	0.04	1	-0.01	0.01	0.01	0	0	-0.07	0	0.01
VGP	0.07	0.07	0.9	-0.69	0.03	-0.01	1	0.38	0.47	0.47	0.38	-0.27	0.7	0.18
TH2O	0.03	0.39	0.51	-0.38	0.32	0.01	0.38	1	0.93	0.81	0.22	-0.36	0.41	0.3
TOD	0.11	0.46	0.61	-0.43	0.43	0.01	0.47	0.93	1	0.93	0.32	-0.46	0.46	0.35
TOA	0.02	0.36	0.57	-0.6	0.37	0	0.47	0.81	0.93	1	0.3	-0.31	0.43	0.53
GPSVEL	0.48	0.26	0.22	-0.04	0.48	0	0.38	0.22	0.32	0.3	1	-0.47	-0.17	0.31
IGA	-0.75	-0.6	-0.37	-0.07	-0.76	-0.07	-0.27	-0.36	-0.46	-0.31	-0.47	1	-0.15	-0.4
IGP	-0.07	0.32	0.86	-0.66	-0.09	0	0.7	0.41	0.46	0.43	-0.17	-0.15	1	0.04
ROT	0.34	0.27	0.21	-0.5	0.48	0.01	0.18	0.3	0.35	0.53	0.31	-0.4	0.04	1

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

não compromete a análise dos principais sistemas de uma locomotiva.

Removidas as variáveis binárias, restaram 70 atributos contínuos, ou 57.225 subespaços, valor ainda proibitivo. Por isso, optou-se pela seleção manual de variáveis: foram mantidas somente aquelas que participavam da composição de alguma das regras usadas para indicação falha. Desta forma, a dimensionalidade ficou reduzida a 14 atributos ou 469 combinações de subespaços. A Tabela 2 apresenta as variáveis utilizadas neste trabalho que indicam as leituras de sensores para monitoramento dos motores e geradores. O comando de aceleração que o maquinista passa para o sistema chama-se Ponto (PTO) e varia entre valores de 0 a 9. A Tabela 3 exibe a correlação dentre os principais atributos.

3.2 IDENTIFICAÇÃO DE SUBESPAÇOS

Por meio da técnica *isolation path* foi possível classificar os subespaços de acordo com o número de iterações necessárias para que cada objeto fosse isolado. Como critério, foi adotado que menores pontuações indicam subespaços mais relevantes, pois quanto menor o número de iterações, mais *outlier* é o elemento e, portanto, maior potencial para este subespaço ser uma boa explicação para a anormalidade de um ponto.

Foram analisados dez pontos em que foi identificada falha de lubrificação do motor diesel em locomotivas diferentes (anomalias) e dez pontos tidos como normais. A partir da aplicação de *isolation path* para os subespaços de cada um destes vinte elementos, obteve-se a pontuação média para os subespaços em amostras *outliers* e *inliers*. Com base nestas médias, foi elaborada uma classificação para os subespaços, exibida na Tabela 4 para os cinco subespaços mais relevantes.

A partir da classificação exibida, algumas conclusões podem ser obtidas. Primeiramente, é possível notar que a pontuação dos elementos tidos como falhos foi significativamente menor que de *inliers*, ou seja, *isolation path* foi capaz de diferenciar amostras normais de *outliers*. Outra observação refere-se à diferença entre os atributos que formam os subespaços melhor avaliados. Enquanto nas amostras normais estes atributos estão relacionados a configurações determinadas pelo operador (e.g. *Ponto de Operação* e *Potência*), nas amostras anômalas, as variáveis utilizadas para proteção do sistema (temperatura) foram evidenciadas. Neste sentido, apesar de não ter sido capaz de identificar os três atributos usados para determinação da falha, *isolation path* apontou medidas relacionadas à temperatura do motor diesel como explicação para as anomalias. Observa-se também que estas variáveis de temperatura apresentam alta correlação com a *temperatura antes do arrefecimento*, variável utilizada na composição da regra elaborada pelos especialistas.

3.3 DESAFIOS

A implementação de uma técnica de OAM, *isolation path*, a dados reais reforçou alguns pontos abordados na literatura. Por exemplo, o crescimento exponencial das combinações de subespaços mostrou-se um empecilho para a utilização de todo o espaço vetorial disponível, como já abordado em outros trabalhos acadêmicos (VINH *et al.*, 2016; ZIMEK *et al.*, 2012). Dentre as alternativas existentes, experimentou-se a implementação do algoritmo Beam Search (VINH *et al.*, 2016), que busca filtrar os subespaços a partir de resultados anteriores (por exemplo, somente variáveis bem avaliadas em subespaços com duas dimensionalidades são consideradas para cálculo nos subespaços com três dimensionalidades). No entanto, esta técnica acabou degradando os resultados por deixar de considerar atributos importantes mas que não obtiveram uma boa classificação em dimensionalidades menores.

A dificuldade em avaliar adequadamente o resultado também se encontra presente na

Tabela 4 – Subespaços melhor classificados pelo método *isolation path*.

Amostras normais		Amostras com falha	
Subespaço	Pontuação	Subespaço	Pontuação
[PTO]	8.84 ± 1.08	[TH2O]	5.25 ± 0.75
[POT, PTO, ROT]	8.89 ± 1.51	[TOD]	5.37 ± 0.72
[PCOMB, PTO]	8.92 ± 1.20	[TH2O, TOD]	5.47 ± 0.81
[PCOMB, PTO, ROT]	8.93 ± 1.10	[TH2O, TOD, TOA]	5.62 ± 0.64
[PCOMB]	8.95 ± 1.62	[TH2O, TOA]	5.65 ± 0.65

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

literatura. Duan *et al.* (2015), por exemplo, ao obter resultados surpreendentes em uma análise sobre as características de jogadores de basquete da NBA, salienta que os atributos apontados por sua técnica como os mais relevantes não indicavam necessariamente as maiores qualidades dos jogadores, mas sim a anormalidade desses atributos. Neste sentido, uma tabela verdade elaborada por especialistas provavelmente ignoraria tais atributos. No experimento realizado com dados de locomotiva, o algoritmo foi capaz de indicar uma direção de estudo para o analista. No entanto, não foi possível estabelecer uma nota para o resultado de modo a viabilizar uma comparação com resultados fornecidos por outras técnicas.

Por fim, apesar de ter sido proposta para variáveis contínuas (VINH *et al.*, 2016), as dificuldades de utilização do *isolation path* com variáveis de baixa cardinalidade (e.g. binárias) não se encontram documentadas. Conforme demonstrado, a utilização de atributos com esta distribuição impossibilita o correto isolamento dos objetos de interesse e a comparação da métrica entre diferentes subespaços.

3.4 DISCUSSÃO

Por mais que não tenha sido capaz de identificar todos os atributos considerados pela regra para determinar a falha em questão, ao determinar que a anormalidade estava relacionada à temperatura do motor diesel, *isolation path* indicou uma direção para que especialistas se aprofundassem na interpretação da anomalia. Ademais, mesmo que a quantidade reduzida de *outliers* utilizados neste experimento seja insuficiente para uma análise qualitativa da técnica empregada, considerações relevantes sobre a aplicação de OAM merecem ser destacadas uma vez que anomalias são por definição raras.

Primeiramente, a facilidade de implementação do algoritmo chama a atenção: a técnica funciona independentemente da escala das variáveis, não demanda cálculos complexos e pode ser apurada somente para o objeto de interesse. Quanto à configuração dos parâmetros de execução, o número de amostras por árvore (subamostragem) e quantidade de árvore binárias construídas para o cálculo da pontuação média de um subespaço, podem ser determinados observando-se a convergência da média, ou seja, aumento dos valores destes parâmetros deixa de surtir efeito no resultado a partir de um certo nível.

Outro ponto positivo que deve ser ressaltado é a possibilidade de paralelização do cálculo: as métricas para cada subespaço e a construção de cada uma das árvores binárias podem ser feitas de forma independente. No experimento realizado, foram usados oito núcleos de

processamento e paralelizou-se somente o cálculo da pontuação entre subespaços diferentes. No entanto, nem esta estratégia de concorrência de processamento e nem a simplicidade dos algoritmos foram suficientes para que resultados fossem obtidos em tempo razoável devido à explosão combinatória de subespaços. Por isso, fez-se necessária também a filtragem manual dos atributos. Com esta redução do espaço vetorial, levou-se aproximadamente meia-hora para a obtenção dos resultados com os 14 atributos remanescentes. Tempo este que permitiria a análise dos dados de até 150 locomotivas uma vez que uma frota deste tamanho gera em média 40 situações que requerem a atenção de especialistas por dia.

Podem ser citadas como propriedades desfavoráveis do *isolation path* a falta de linearização das métricas e sua susceptibilidade frente a atributos dominantes. A primeira dificulta a comparação dos resultados, sobretudo em tarefas com variáveis de domínios e escalas diferentes. Por sua vez a sensibilidade frente a atributos dominantes pode mascarar subespaços relevantes. No experimento realizado, variáveis consideradas semelhantes (com alto grau de correlação) apareceram repetitivamente nas primeiras posições da classificação realizada.

Os desafios identificados acima são porém considerados fora do escopo deste trabalho pois soluções práticas demandarão atenção da comunidade por um longo tempo.

Dentre as melhorias que podem ser desenvolvidas em futuras implementações no contexto da aplicação descrita, enfatiza-se que a filtragem de atributos altamente correlatos seria um importante avanço. Desta forma, espera-se evitar o mascaramento das demais variáveis, como ocorreu no experimento, em que atributos semelhantes dominaram todas as primeiras posições da classificação.

Além disso, seria interessante que a comparação entre as amostras fosse baseada em condições de operação semelhantes. No experimento, pontos de falha existentes somente em alta potência foram comparados frente a amostras normais obtidas aleatoriamente, o que pode resultar em uma análise tendenciosa. Como melhoria, estratégias de clusterização poderiam ser adotadas para separação das amostras usando critérios como localização espacial, potência dos motores etc.

Quanto aos resultados, a listagem dos subespaços mais relevantes mostrou-se de difícil utilização sem um ambiente para posterior exploração. Por exemplo, análise do comportamento da temperatura da água (atributo apontado como relevante) não faz parte da abordagem de OAM, que preocupa-se somente com a identificação dos atributos mas não indica se eles são maiores ou menores que os valores encontrados em *inliers*. Por isso, a adição elementos visuais poderia

complementar a estratégia adotada neste capítulo e auxiliar o especialista na interpretação de anomalias. Assim, a experiência obtida neste caso de uso foi usada para substanciar e guiar o desenvolvimento da ferramenta de VA apresentada no próximo capítulo.

4 VISUAL ANALYTICS PARA OAM

Os resultados obtidos a partir da aplicação de uma técnica de OAM serviram de motivação para o desenvolvimento de uma estratégia de VA adequada a este contexto de estudo. Desenvolveu-se então uma plataforma com o objetivo fornecer subsídios ao especialista para a tomada de decisões a partir da explicação dos motivos que levaram a uma falha.

Para demonstrar a aplicabilidade da plataforma proposta, escolheu-se uma situação de falha em locomotiva relacionada à temperatura do motor diesel. Esta anomalia normalmente começa a aparecer à medida que o ciclo de vida dos trocadores de calor se aproxima do fim.

Dentre as variáveis disponíveis para análise (reapresentadas na Tabela 5), estudos anteriores de OAM indicaram corretamente que a temperatura do óleo antes/depois do arrefecimento (TOA e TOD) e temperatura da água (TH2O) seriam mais relevantes à interpretação desta falha.

Tabela 5 – Variáveis da locomotiva analisadas ou controladas.

Equipamento	Variável contínua
Motor diesel	Temperatura óleo antes do arrefecimento (^{o}C) Temperatura óleo depois do arrefecimento (^{o}C) Temperatura da água (^{o}C) Pressão lubrificante (psi) Pressão combustível (psi) Rotação (rpm) / POT (HP)
Gerador Principal	Tensão Gerador Principal (V) Corrente Gerador Principal (A)
Gerador Companheiro	Corrente Gerador Principal (V) Corrente Gerador Auxiliar (A)
Motores de tração	Corrente Motor de tração (A)
Geral	Velocidade de deslocamento (km/h) Ponto de operação

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

4.1 VISÃO GERAL

Conforme mencionado anteriormente, *Visual Analytics* não se restringe à criação de gráficos, também engloba por exemplo o gerenciamento e análise de eventos (KEIM *et al.*, 2010). Assim, no âmbito deste trabalho VA faz parte de um sistema idealizado para o tratamento e análise de dados de locomotivas, conforme apresentado na Figura **??**. Nele, as informações são recebidas por um sistema de telemetria e busca-se por *outliers*. No caso de algum evento raro ser detectado, cabe à técnica de OAM determinar uma interpretação para a anomalia. Os resultados desta análise ficam então disponíveis para a posterior análise de especialistas. Cabe salientar,

entretanto, que a etapa de identificação de *outliers* não fez parte deste trabalho, focando-se assim na contextualização (OAM) e fornecimento ferramentas para análise (VA) de falhas já conhecidas.

Uma vez realizados os processamentos para identificação e interpretação de anomalias, especialistas podem explorar as condições de falha por meio de uma interface que enfatiza: (i) exibição temporal de múltiplas variáveis, (ii) marcação clara do subconjunto de dados no momento da anomalia, (iii) sugestão de subconjuntos de atributos relevantes e (iv) possibilidade de mudança para visualização de séries temporais. A dinâmica de utilização da interface do ponto de vista de utilização é apresentada na Figura 4(b).

Para exibição multidimensional usamos PCP, como pode ser visto na Figura 5, em que cada eixo vertical representa uma variável. Para indicar o momento de detecção do *outlier*, usamos um esquema de cores em que linhas verde representam o período antes da anomalia e vermelho os instantes posteriores. Além disso, estas cores são representadas com diferente grau de saturação, de modo que amostras mais próximas temporalmente ao instante da falha fiquem mais evidentes. Uma legenda ao lado direito do gráfico permite que o usuário relacione temporalmente a cor com que a amostra foi representada no gráfico com a data da amostragem.

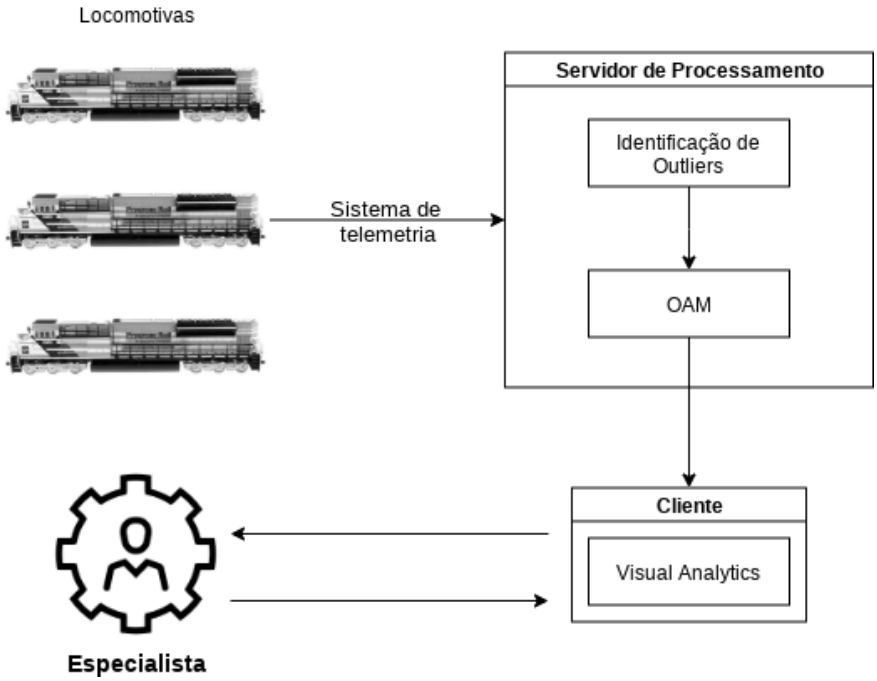
Em casos complexos pode ser interessante analisar os dados como séries temporais, pois estes gráficos propiciam a visualização de padrões e tendências, além de serem de simples utilização e bastante usuais (AIGNER *et al.*, 2007). Um exemplo adequado para visualização neste formato é a degradação das condições de operação de um equipamento. Porém, para isto seria preciso reduzir a dimensionalidade de modo a evitar que a visualização fique sobrecarregada. Neste caso oferecemos ao especialista subconjuntos de variáveis identificadas pelo OAM. Ao escolher um subconjunto, o especialista passa a visualizar os dados numa série temporal como na Figura 6.

4.2 VISUALIZAÇÃO TEMPORAL

A representação de séries temporais é extremamente comum, de fácil de interpretação e manipulação, sendo por isso incluída na plataforma. Para adequá-la ao contexto de OAM, o usuário pode escolher o subespaço de variáveis a ser representado. Além disso, como complemento, também exibe-se a evolução de uma métrica de anormalidade calculada diariamente para o subespaço analisado, no caso, optou-se pelo LOF (BREUNIG *et al.*, 2000), muito aplicada a OAM (VINH *et al.*, 2016). O valor obtido para LOF foi normalizado para o intervalo de 0

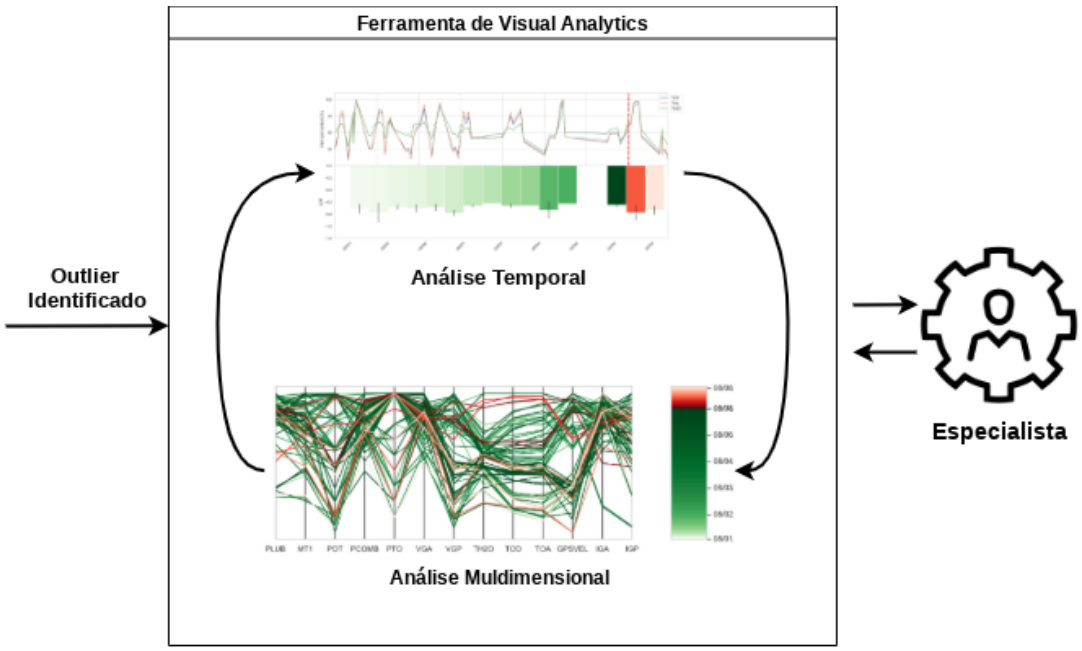
Figura 4 – Visão geral da plataforma proposta.

(a) Obtenção e processamento dos dados.



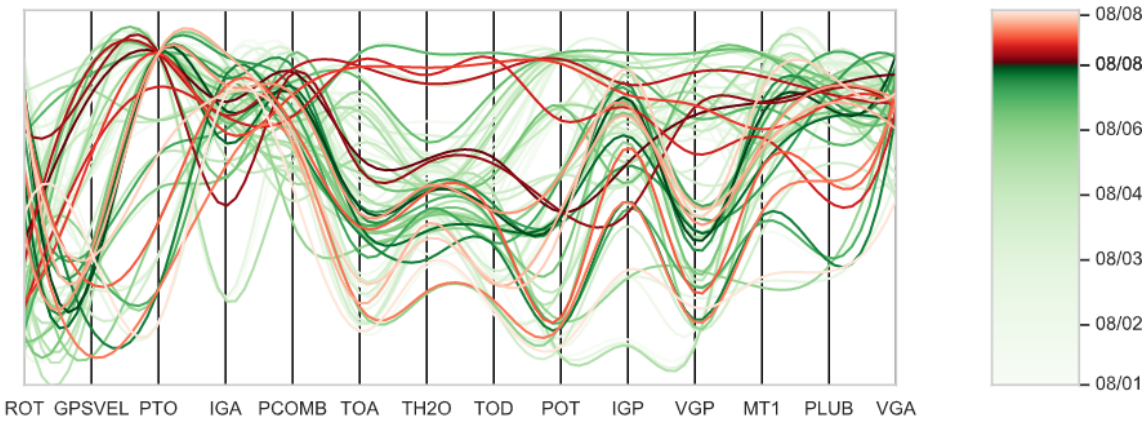
Fonte: O Autor

(b) Perspectiva do especialista.



Fonte: O Autor

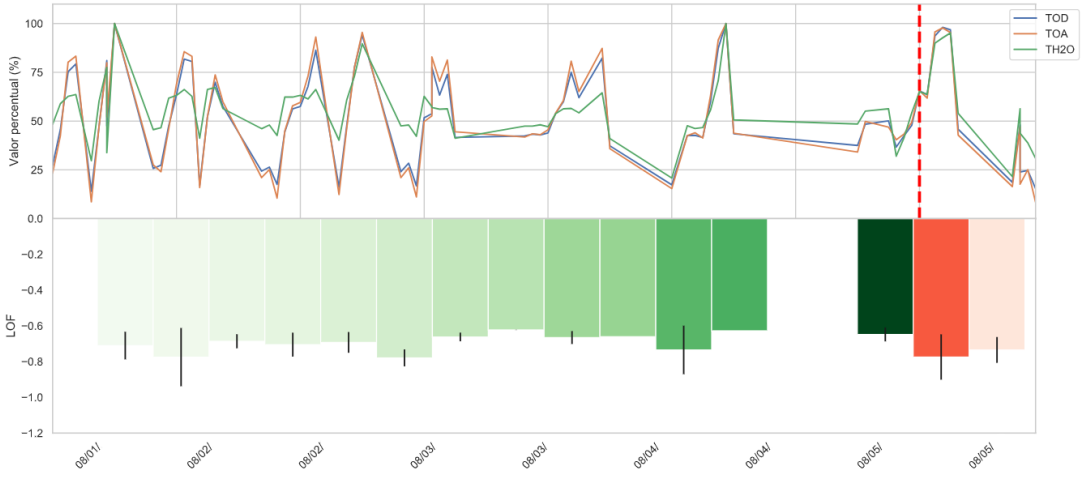
Figura 5 – PCPs com linhas curvas.



Fonte: O Autor

a -1, em que -1 representa o máximo de anormalidade para o subespaço. O gradiente de cores utilizado para a exibição da métrica de anormalidade segue o mesmo padrão adotado para a representações de amostra no gráfico de coordenadas paralelas, com valores amostrados antes da falha em verde e, após a falha, em vermelho.

Figura 6 – Modo de visualização temporal



Fonte: O Autor

Dentre os parâmetros que o usuário pode determinar, destacam-se as janelas de tempo prévio ou posterior ao momento de detecção da falha. Estas configurações são importantes porque, de acordo com o tipo de anomalia analisada, diferentes intervalos são necessários para que padrões fiquem evidentes. Por exemplo, um problema relacionado a um evento extremo pode ser visível imediatamente. Por sua vez, a degradação de certos equipamentos pode durar anos.

A janela de tempo escolhida possui consequências diretas na quantidade de amostras a serem exibidas. Dentre as ferramentas disponíveis para reduzir o *overplotting*, destaca-se

a possibilidade de sub-amostrar as variáveis com granularidade de segundos até anos. Além disso, pode-se optar pela utilização de funções como máximo, mínimo e média dos valores conforme a necessidade. Nos testes realizados, as funções de máximo e mínimo mostraram-se mais adequadas devido à natureza extrema das anomalias.

Para as imagens exibidas nesta seção, foi considerado o período de sete dias antes da identificação da falha (o que ocorreu em 08/08) e um dia após esta detecção. Quanto à amostragem dos dados, optou-se por 90 minutos, representando-se o valor máximo para cada variável no intervalo amostrado.

Na Figura 6, as três variáveis exibidas foram selecionadas tomando-se por base resultados anteriores fornecidos pelo OAM: temperatura do óleo depois do arrefecimento (TOD), temperatura do óleo antes do arrefecimento (TOA) e temperatura da água do trocador de calor (TH2O). As lacuna existentes na representação do LOF ocorrem para dias em que a locomotiva não gerou dados. Ademais, o momento em que a falha foi detectada é indicado através de uma linha pontilhada no gráfico.

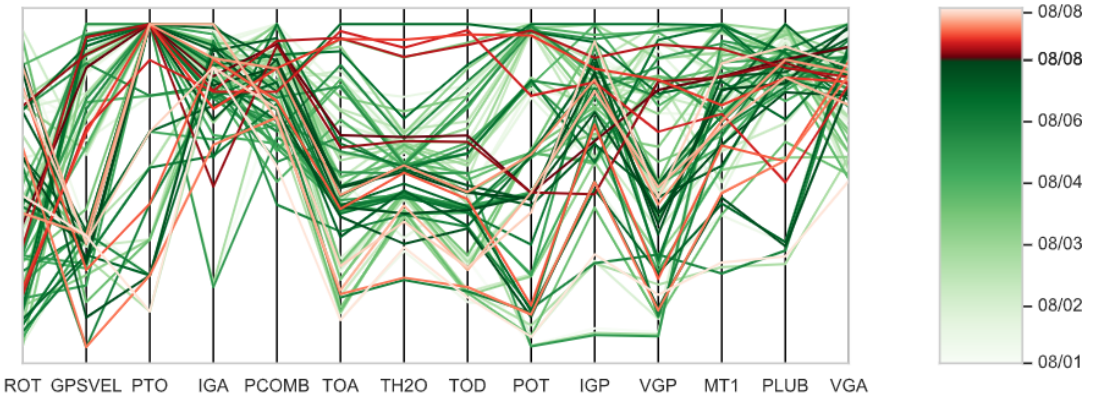
4.3 VISUALIZAÇÃO MULTIDIMENSIONAL

Uma das inovações deste trabalho é comunicação em um gráfico de coordenadas paralelas da evolução temporal. Neste sentido, amostras obtidas antes da identificação do *outlier* são representadas em verde e, após a detecção, em vermelho. No entanto este critério não é suficiente para comunicação da passagem do tempo de forma contínua. Por isso, como complemento, adaptou-se a estratégia adotada por (JOHANSSON *et al.*, 2007; ZHONGHUA; LINGDA, 2016), em que amostras mais antigas são exibidas de forma mais apagada. Contudo, como para esta aplicação a representação se estende também para após o instante de detecção da anomalia, decidiu-se pelo aumento gradual da transparência das linhas após a falha também. Desta forma, salientam-se as amostras situadas temporalmente mais próximas ao momento crítico em que a anomalia foi encontrada.

Diferentes funções para o gradiente de saturação das cores foram testadas para o período anterior à falha: $f(x) = x$, $f(x) = log(x)$ e $f(x) = 1 - log(x)$, conforme pode ser observado na Figura 7. Por enfatizar o período mais próximo à anomalia e reduzir a confusão entre as linhas, a função $f(x) = 1 - log(x)$ foi a preferida. Para o período após a falha, foi adotado um decaimento linear, pois o tempo representado costuma ser mais reduzido. Na plataforma desenvolvida, todas as alternativas apresentadas estão disponíveis para seleção do usuário.

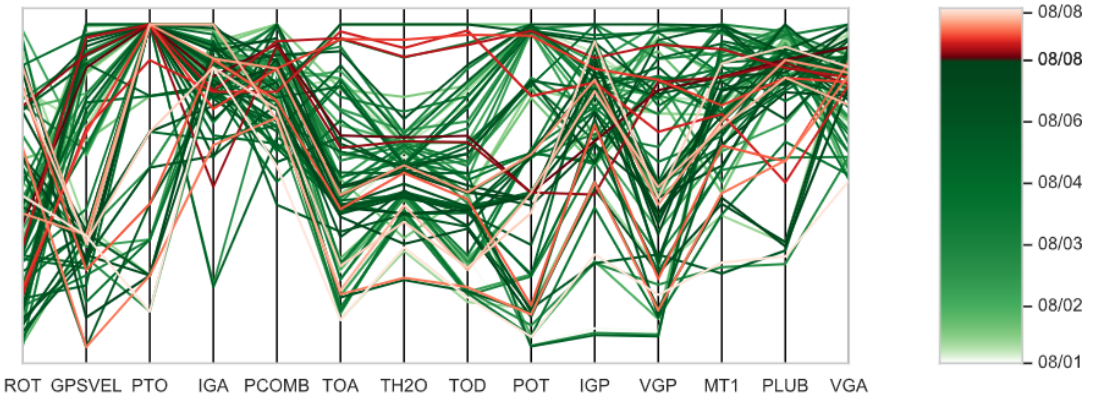
Figura 7 – PCP: diferentes gradientes para saturação de cores.

(a) Saturação de cores f(x) = x



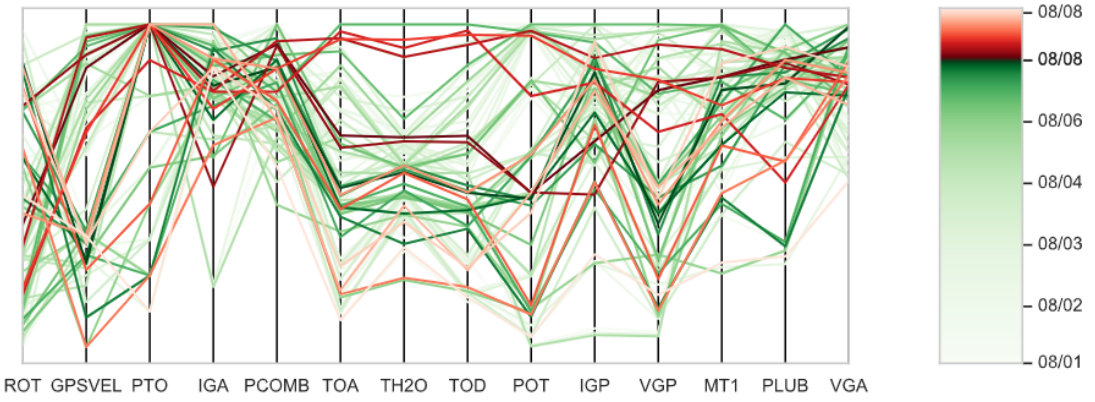
Fonte: O Autor

(b) f(x) = log(x)



Fonte: O Autor

(c) f(x) = 1-log(x)



Fonte: O Autor

Conforme demonstrado em (HEINRICH; WEISKOPF, 2013), se os valores de duas amostras forem iguais para um mesmo atributo, o senso de continuidade da polilinha pode ser perdido. Dentre as alternativas existentes para combater esta ambiguidade, ressaltam-se a utilização de diferentes cores e a substituição de linhas retas por curvas. No modelo proposto, apesar de se adotar um gradiente de cores, dependendo da proximidade temporal, é possível que os elementos gráficos se confundam. Por isso, o usuário pode alternar entre as duas formas de exibição: linhas curvas e retas. A representação com linhas retas foi mantida, pois a interpolação pode gerar distorções dos valores exibidos. Dentre as funções de interpolação testadas, preferiu-se a spline. A Figura 5 é um exemplo de PCP elaborado com linhas curvas.

4.3.1 Ordenação dos atributos

Quanto à ordem de exibição dos atributos nos eixos dos gráficos de coordenadas paralelas, propomos posicionar as variáveis mais importantes mais perto do centro. Assim, quanto mais irrelevante um atributo for, mais para as bordas ele será representado.

O critério estabelecido para determinar a relevância dos atributos e, consequentemente sua posição no gráfico, baseou-se na pontuação fornecida pela técnica de OAM para os subespaços dos quais as variáveis fazem parte. No caso da técnica empregada, iPath, quanto menor o valor, mais representativo um subespaço.

Dada a existência de atributos dominantes e a ausência de normalização para as medidas fornecidas pela técnica iPath, optou-se pela média das menores pontuações em lugar de valores puros ou médias globais. No caso, para o espaço vetorial de 14 atributos utilizados nesta análise, a média dos três menores valores mostrou-se mais adequada por evitar que pontuações extremas, baixas ou altas, distorcessem os cálculos.

É comum em OAM que seja elaborada uma classificação para os atributos, i.e. estabelecer o 1º, 2º,..., n^o atributo mais relevante para a contextualização da anomalia. A partir desta classificação, propomos a equação abaixo para se determinar em qual dos eixos cada variável deve ser alocada em um PCP:

$$IndiceEixo = \mathit{floor}\left(\frac{num.atributos}{2}\right) + (-1)^{classificao} \times \mathit{floor}\left(\frac{classificao}{2}\right)$$

em que:

$$1 \leq classificao \leq num.atributos$$

Para as figuras aqui representadas, considerou-se que o eixo de índice 1 estaria na borda

Tabela 6 – Posicionamento dos atributos nos eixos do PCP a partir de métrica obtida por OAM.

Classificação	Atributo	Pontuação	Índice eixo
1	TH20	5.45	7
2	TOD	5.49	8
3	TOA	5.64	6
4	POT	6.01	9
5	PCOMB	6.04	5
6	IGP	6.14	10
7	IGA	6.67	4
8	VGP	6.75	11
9	PTO	7.08	3
0	MT1	7.54	12
11	GPSVEL	7.93	2
12	PLUB	8.06	13
13	ROT	9.13	1
14	VGA	9.52	14

Fonte: Dados fornecidos pela empresa Progress Rail Equipamentos e Serviços Ferroviários do Brasil LTDA.

da esquerda e o índice 14 localiza-se na borda da direita. A Tabela 6 sintetiza as pontuações, classificação e eixo para cada atributo. Além disso, salienta-se que qualquer outra métrica de OAM capaz de gerar uma classificação dos atributos poderia ser utilizada no lugar do *iPath*.

4.4 DISCUSSÃO

As imagens apresentadas na seção anterior demonstram como a plataforma desenvolvida tem potencial para auxiliar na contextualização de anomalias. A Figura 6, por exemplo, demonstra como o subespaço representado na série temporal de fato apresenta valores altos para a métrica de anormalidade (LOF) nos instantes próximos à identificação da anormalidade. Além disso, o PCP também indica com clareza que as variáveis relacionadas à temperatura do motor diesel (TOA, TOD e TH20) tiveram valores elevados próximos à detecção da falha, indo ao encontro dos resultados fornecidos pela técnica de OAM e à condição de falha. Desta forma, a ferramenta proposta demostra potencial de auxiliar especialistas na investigação de eventos e nas tomadas de decisões.

Dentre as dificuldades para uso do gráfico de coordenadas paralelas, salientamos a necessidade de se escolher taxas de amostragem adequadas para se evitar o *overplotting*. Uma alternativa seria a substituição da frequência de amostragem contínua por uma análise que levasse em consideração a relevância das amostras, possivelmente baseada na variação dos dados. Com isso, amostras muito semelhantes não seriam representadas e, tanto os períodos analisados quanto a granularidade das amostras, poderiam ser expandidos sem perda de informações ou prejuízo à identificação de padrões visuais.

Quanto à utilização de gradiente de cores para a representação temporal nos gráficos de coordenadas paralelas, consideramos que esta estratégia comunicou com sucesso a evolução do estado da máquina através do tempo. No entanto, testes com usuários reais ainda precisariam ser realizados para determinação da curva de aprendizado ou mesmo da efetividade deste modelo.

5 CONCLUSÕES E PERSPECTIVAS

A pesquisa apresentada consiste em uma das poucas aplicações de *Outlying Aspect Mining* a dados reais, no caso, falhas em locomotivas. Além disso, uma plataforma de *Visual Analytics* para subsídio à tomada de decisões com base nos resultados fornecidos por OAM foi proposta.

Quanto à aplicação de OAM, observamos que a explosão combinatória de subespaços e as limitações da técnica de OAM empregada, que não lida bem com variáveis de baixa cardinalidade, impediram a análise simultânea de todos os atributos do sistema, como almejado inicialmente. No entanto, mesmo com estas restrições, o algoritmo foi capaz de indicar uma direção para a análise de especialistas.

Dentre as melhorias que podem ser implementadas em futuras aplicações, salientamos a filtragem de atributos altamente correlatos. Com isso, esperamos impedir que uma mesma variável ou atributos muito semelhantes dominem a lista de subespaços mais importantes, o que não aumenta a compreensão do especialista sobre o sistema. Além disso, neste contexto de locomotivas, seria interessante verificar o efeito que aspectos locais e ambientais possuem sobre o resultado. Por exemplo, se um trecho de subida de serra, altamente inclinado, seria comparável a ambientes mais usuais.

Repetir o experimento com mais exemplos de falha e com diferentes técnicas de OAM (LOF por exemplo), seria importante para uma avaliação mais aprofundada dos resultados e da robustez dos algoritmos. Entretanto, como citado anteriormente, a comparação de resultados de OAM ainda é uma questão aberta e pode dificultar esta análise.

No âmbito de *Visual Analytics*, ambas as visualizações propostas mostraram-se capazes de auxiliar na interpretação dos resultados por especialistas. A série temporal, por exemplo, evidenciou altos valores para a métrica de anormalidade logo após o instante em que a falha foi identificada. Por sua vez, o PCP exibiu um comportamento diferenciado de alguns atributos nos momentos críticos à análise. Assim, ambas as abordagens evidenciaram a importância que um subespaço específico possui frente a todo o espaço vetorial disponível, ajudando assim na contextualização da falha e fornecendo subsídios para que um especialista compreendesse o problema e tomasse as medidas necessárias.

Os gráficos de coordenadas paralelas utilizaram um gradiente de saturação das cores para a representação temporal de dados multidimensionais. Consideramos que esta estratégia

mostrou com clareza a evolução temporal dos elementos, ao mesmo tempo que limitou a confusão de linhas (*overplotting*).

Futuramente, pretendemos implementar nos PCPs um sistema que substitua a amostragem periódica por critérios que levem em conta a variação dos atributos. Deste modo, pontos muito parecidos não precisariam ser representadas múltiplas vezes, o que diminuiria o número de linhas e possibilitaria o aumento dos intervalos representados.

Apesar de ter sido avaliado em um contexto específico, i.e. falhas ocorridas na operação de locomotivas, salientamos que a solução apresentada é genérica e pode ser replicada para outros cenários. Julgamos, entretanto, que a estratégia descrita é mais adequada a problemas que careçam de uma análise individual. Este é o caso de demais situações industriais (e.g. mineração e naval) em que os equipamentos ficam sujeitos a intemperes climáticas, diferentes operadores e mudanças na forma de utilização, apresentando alta variabilidade nas origens das falhas. No entanto, maquinários que fazem parte de uma mesma linha de produção e estão sujeitos sempre ao mesmo tipo de forças tendem a apresentar falhas similares, não justificando a análise individual proposta por OAM.

A análise da evolução de doenças também poderia se beneficiar da estratégia apresentada neste trabalho. Neste caso, o estudo da evolução temporal de atributos específicos poderia auxiliar na compreensão das causas e na prevenção de doenças futuras. Contudo, a grande quantidade de variáveis observáveis no corpo humano e nos hábitos de uma pessoa poderiam ser um impedimento, dada a quantidade de combinação de subespaços possíveis.

REFERÊNCIAS

AIGNER, Wolfgang; MIKSCH, Silvia; MÜLLER, Wolfgang; SCHUMANN, Heidrun; TOMINSKI, Christian. Visualizing time-oriented data—a systematic view. **Computers & Graphics**, Elsevier, v. 31, n. 3, p. 401–409, 2007.

ANGIULLI, Fabrizio; FASSETTI, Fabio; MANCO, Giuseppe; PALOPOLI, Luigi. Outlying property detection with numerical attributes. **Data Min. Knowl. Discov.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 31, n. 1, p. 134–163, jan. 2017. ISSN 1384-5810. Disponível em: https://doi.org/10.1007/s10618-016-0458-x.

BARLOW, N; STUART, Liz J. Animator: A tool for the animation of parallel coordinates. *In*: IEEE. **Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.** *[S.l.]*, 2004. p. 725–730.

BATZEL, T.; SWANSON, D. C. Prognostic health management of aircraft power generators. **IEEE Transactions on Aerospace and Electronic Systems**, v. 45, 2009.

Blaas, J.; Botha, C.; Post, F. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1436–1451, 2008.

BREUNIG, Markus M; KRIEGEL, Hans-Peter; NG, Raymond T; SANDER, Jörg. Lof: identifying density-based local outliers. *In*: **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**. *[S.l.: s.n.]*, 2000. p. 93–104.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 41, n. 3, p. 1–58, 2009.

CHEN, Gang; SHI, Juan; CAI, Yuan Li. Ordinal outlier detection based on recursive uniform partitioning. **Transactions of the Institute of Measurement and Control**, v. 35, n. 7, p. 940–948, 2013. Disponível em: https://doi.org/10.1177/0142331211431272.

CHEN, Hongtian; JIANG, Bin; LU, Ningyun. A newly robust fault detection and diagnosis method for high-speed trains. **IEEE Transactions on Intelligent Transportation Systems**, PP, p. 1–11, 09 2018.

CHEN, Min; TREFETHEN, Anne; BANARES-ALCANTARA, Rene; JIROTKA, Marina; COECKE, Bob; ERTL, Thomas; SCHMIDT, Albrecht. From data analysis and visualization to causality discovery. **Computer**, IEEE, n. 10, p. 84–87, 2011.

CHEN, Woruo; YUN, Yong-Huan; WEN, ming; LU, Hongmei; ZHANG, Zhimin; LIANG, Yizeng. Representative subset selection and outlier detection via isolation forest. **Analytical methods**, v. 8, 10 2016.

CIELEN, Davy; MEYSMAN, Arno; ALI, Mohamed. **Introducing data science: big data, machine learning, and more, using Python tools**. *[S.l.]*: Manning Publications Co., 2016.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: https://doi.org/10.1023/A:1022627411411.

DANG, Xuan-Hong; ASSENT, Ira; NG, Raymond; ZIMEK, Arthur; SCHUBERT, Erich. Discriminative features for identifying and interpreting outliers. *In*: . *[S.l.: s.n.]*, 2014. p. 88–99. ISBN 978-1-4799-2555-1.

DIETZSCH, Janko; HEINRICH, Julian; NIESELT, Kay; BARTZ, Dirk. Spray: A visual analytics approach for gene expression data. *In*: IEEE. **2009 IEEE Symposium on Visual Analytics Science and Technology**. *[S.l.]*, 2009. p. 179–186.

DUAN, Lei; TANG, Guanting; PEI, Jian; BAILEY, James; CAMPBELL, Akiko; TANG, Changjie. Mining outlying aspects on numeric data. **Data Mining and Knowledge Discovery**, Springer, v. 29, n. 5, p. 1116–1151, 2015.

FROSINI, Lucia; BASSI, Ezio. Stator current and motor efficiency as indicators for different types of bearing faults in induction motors. **Industrial Electronics, IEEE Transactions on**, v. 57, p. 244 – 251, 02 2010.

GAO, Zhiwei; CECATI, Carlo; DING, Steven X. A survey of fault diagnosis and fault-tolerant techniques—part i: Fault diagnosis with model-based and signal-based approaches. **IEEE Transactions on Industrial Electronics**, v. 62, p. 3757–3767, 2015.

GARRAMIOLA, Fernando; LARRAñAGA, Jon del Olmo; POZA, Javi; MADINA, Patxi; ALMANDOZ, Gaizka. Integral sensor fault detection and isolation for railway traction drive. **Sensors**, v. 18, p. 1543, 05 2018.

General Motors, Electro-Motive Division. **SD40-2 Operators Manual**. 5th. ed. La Grange: Service Department - EMD & GM, 1978. 61 p.

General Motors, Electro-Motive Division. **SD40-2 Locomotive Service Manual**. 8th. ed. La Grange: Service Department - EMD & GM, 1997. 460 p.

GENG, Zunmin; CHEN, Jin; HULL, J. B. Analysis of engine vibration and design of an applicable diagnosing approach. 2003.

GLENDENNING, Kurtis; WISCHGOLL, Thomas; HARRIS, Jack; VICKERY, Rhonda; BLAHA, Leslie. Parameter space visualization for large-scale datasets using parallel coordinate plots. **Electronic Imaging**, Society for Imaging Science and Technology, v. 2016, n. 1, p. 1–8, 2016.

GLODEK, Michael; SCHELS, Martin; SCHWENKER, Friedhelm. Ensemble gaussian mixture models for probability density estimation. **Comput. Stat.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 28, n. 1, p. 127–138, fev. 2013. ISSN 0943-4062. Disponível em: http://dx.doi.org/10.1007/s00180-012-0374-5.

GOODALL, Roger; KORTüM, W. Mechatronic development for railway vehicles of the future. *In*: . *[S.l.: s.n.]*, 2002. v. 10, p. 21–33.

GUHA, Sudipto; MISHRA, Nina; ROY, Gourav; SCHRIJVERS, Okke. Robust random cut forest based anomaly detection on streams. *In*: **Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48**. JMLR.org, 2016. (ICML'16), p. 2712–2721. Disponível em: http://dl.acm.org/citation.cfm?id=3045390.3045676.

HAPEMAN, Martin J.; LONG, James; PLETTE, David L. Diesel electric locomotive propulsion systems-a look into the future. **Industry Applications, IEEE Transactions on**, v. 23, p. 495 – 501, 06 1986.

HARIRI, Sahand; KIND, Matias Carrasco. Batch and online anomaly detection for scientific applications in a kubernetes environment. *In*: **Proceedings of the 9th Workshop on Scientific Cloud Computing**. New York, NY, USA: ACM, 2018. (ScienceCloud'18), p. 3:1–3:7. ISBN 978-1-4503-5863-7. Disponível em: http://doi.acm.org/10.1145/3217880.3217883.

HAWKINS, D. **Identification of Outliers**. 1st. ed. *[S.l.]*: Springer Netherlands, 1980.

HEINRICH, Julian; WEISKOPF, Daniel. State of the art of parallel coordinates. *In*: **Eurographics (STARs)**. *[S.l.: s.n.]*, 2013. p. 95–116.

INSELBERG, Alfred. The plane with parallel coordinates. **The visual computer**, Springer, v. 1, n. 2, p. 69–91, 1985.

ISERMANN, Rolf. Model-based fault-detection and diagnosis – status and applications. **Annual Reviews in Control**, v. 29, n. 1, p. 71 – 85, 2005. ISSN 1367-5788. Disponível em: http://www.sciencedirect.com/science/article/pii/S1367578805000052.

JOHANSSON, Jimmy; LJUNG, Patric; COOPER, Matthew D. Depth cues and density in temporal parallel coordinates. *In*: **EuroVis**. *[S.l.: s.n.]*, 2007. v. 7, p. 35–42.

KEIM, Daniel A.; MANSMANN, Florian; THOMAS, Jim. Visual analytics: How much visualization and how much analytics? **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 11, n. 2, p. 5–8, maio 2010. ISSN 1931-0145. Disponível em: https://doi.org/10.1145/1809400.1809403.

KELLER, Fabian; MULLER, Emmanuel; BOHM, Klemens. Hics: High contrast subspaces for density-based outlier ranking. *In*: **Proceedings of the 2012 IEEE 28th International Conference on Data Engineering**. Washington, DC, USA: IEEE Computer Society, 2012. (ICDE '12), p. 1037–1048. ISBN 978-0-7695-4747-3. Disponível em: http://dx.doi.org/10.1109/ICDE.2012.88.

LEE, Teng-Yok; SHEN, Han-Wei. Visualization and exploration of temporal trend relationships in multivariate time-varying data. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 15, n. 6, p. 1359–1366, 2009.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation forest. *In*: IEEE. **2008 Eighth IEEE International Conference on Data Mining**. *[S.l.]*, 2008. p. 413–422.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. On detecting clustered anomalies using sciforest. *In*: **Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part II**. Berlin, Heidelberg: Springer-Verlag, 2010. (ECML PKDD'10), p. 274–290. ISBN 3-642-15882-X, 978-3-642-15882-7. Disponível em: http://dl.acm.org/citation.cfm?id=1888305.1888324.

LIU, Ninghao; SHIN, Donghwa; HU, Xia. Contextual outlier interpretation. *In*: **Proceedings of the 27th International Joint Conference on Artificial Intelligence**. AAAI Press, 2018. (IJCAI'18), p. 2461–2467. ISBN 978-0-9992411-2-7. Disponível em: http://dl.acm.org/citation.cfm?id=3304889.3305002.

LUWEI, Kenisuomo; KALTUNGO, Akilu yunusa; SHAABAN, Yusuf. Integrated fault detection framework for classifying rotating machine faults using frequency domain data fusion and artificial neural networks. **Machines**, v. 6, p. 59, 11 2018.

MAO, Zehui; TAO, Gang; JIANG, Bin; YAN, Xing-Gang. Adaptive compensation of traction system actuator failures for high-speed trains. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 18, n. 11, p. 2950–2963, 2017.

MICENKOVÁ, Barbora; NG, Raymond T.; DANG, Xuan-Hong; ASSENT, Ira. Explaining outliers by subspace separability. **2013 IEEE 13th International Conference on Data Mining**, p. 518–527, 2013.

MOSTAFA, Ahmed; CARPENDALE, Sheelagh; BRAZIL, Emilio; EATON, David; SHARLIN, Ehud; SOUSA, Mario Costa. **Visualizing highly multidimensional time varying Microseismic Events**. *[S.l.]*, 2012.

NAJAFABADI, Tooraj Abbasian; SALMASI, Farzad R.; JABEHDAR-MARALANI, P. Detection and isolation of speed-, dc-link voltage-, and current-sensor faults based on an adaptive observer in induction-motor drives. **Industrial Electronics, IEEE Transactions on**, v. 58, p. 1662 – 1672, 06 2011.

NATEGH S., Lindberg D. Aglen O. Brammer R. Boglietti A. Review and trends in traction motor design: Electromagnetic and cooling system layouts. *In*: IEEE. **2018 XIII International Conference on Electrical Machines (ICEM)**. *[S.l.]*, 2018. p. 2600–2606.

NOVOTNY, Matej; HAUSER, Helwig. Outlier-preserving focus+context visualization in parallel coordinates. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 12, n. 5, p. 893–900, 2006.

PUGGINI, Luca; MCLOONE, Sen. An enhanced variable selection and isolation forest based methodology for anomaly detection with oes data. **Eng. Appl. Artif. Intell.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 67, n. C, p. 126–135, jan. 2018. ISSN 0952-1976. Disponível em: https://doi.org/10.1016/j.engappai.2017.09.021.

SANSEN, Joris; RICHER, Gaëlle; JOURDE, Timothée; LALANNE, Frédéric; AUBER, David; BOURQUI, Romain. Visual exploration of large multidimensional data using parallel coordinates on big data infrastructure. *In*: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. **Informatics**. *[S.l.]*, 2017. v. 4, n. 3, p. 21.

SANSEN, Joris; RICHER, Gaëlle; JOURDE, Timothée; LALANNE, Frédéric; AUBER, David; BOURQUI, Romain. Visual exploration of large multidimensional data using parallel coordinates on big data infrastructure. *In*: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. **Informatics**. *[S.l.]*, 2017. v. 4, n. 3, p. 21.

SCHERMAN, Maja; BüLOW, Joakim. **Insider Threat detection using Isolation Forest**. 2018. Student Paper.

SCOTT, David W. **Multivariate density estimation: theory, practice, and visualization**. *[S.l.]*: John Wiley & Sons, 2015.

SINGH, G.K; KAZZAZ, Sa'ad Ahmed Saleh Al. Induction machine drive condition monitoring and diagnostic research—a survey. **Electric Power Systems Research**, v. 64, n. 2, p. 145 – 158, 2003. ISSN 0378-7796. Disponível em: http://www.sciencedirect.com/science/article/pii/S0378779602001724.

SUSTO, Gian Antonio; BEGHI, Alessandro; MCLOONE, Sean. Anomaly detection through on-line isolation forest: An application to plasma etching. *In*: . *[S.l.: s.n.]*, 2017. p. 89–94.

TADEJA, Slawomir Konrad; KIPOUROS, Timoleon; KRISTENSSON, Per Ola. Exploring parallel coordinates plots in virtual reality. *In*: **Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems**. *[S.l.: s.n.]*, 2019. p. 1–6.

TAX, David M. J.; DUIN, Robert P. W. Support vector data description. **Mach. Learn.**, Kluwer Academic Publishers, Norwell, MA, USA, v. 54, n. 1, p. 45–66, jan. 2004. ISSN 0885-6125. Disponível em: https://doi.org/10.1023/B:MACH.0000008084.60811.49.

THOMAS, James; COOK, Kristen A. Illuminating the path: The r&d agenda for visual analytics national visualization and analytics center. **National Visualization and Analytics Center**, 2005.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246. Disponível em: http://www.jstor.org/stable/2346178.

TONG, Chuxuan; ZHANG, Jinglan; CHOWDHURY, Alok; TROST, Stewart G. An interactive visualization tool for sensor-based physical activity data analysis. *In*: **Proceedings of the Australasian Computer Science Week Multiconference**. *[S.l.: s.n.]*, 2019. p. 1–4.

VINH, Nguyen Xuan; CHAN, Jeffrey; BAILEY, James. Reconsidering mutual information based feature selection: A statistical significance view. *In*: **AAAI**. *[S.l.: s.n.]*, 2014.

VINH, Nguyen Xuan; CHAN, Jeffrey; ROMANO, Simone; BAILEY, James; LECKIE, Christopher; RAMAMOHANARAO, Kotagiri; PEI, Jian. Discovering outlying aspects in large datasets. **Data Mining and Knowledge Discovery**, v. 30, n. 6, p. 1520–1555, nov 2016. ISSN 1573-756X. Disponível em: https://doi.org/10.1007/s10618-016-0453-2.

WEBGA, Kodzo; LU, Aidong. Discovery of rating fraud with real-time streaming visual analytics. *In*: IEEE. **2015 IEEE Symposium on Visualization for Cyber Security (VizSec)**. *[S.l.]*, 2015. p. 1–8.

XUE, Feng; YAN, Weizhong; RODDY, Nicholas; VARMA, Anil. Operational data based anomaly detection for locomotive diagnostics. *In*: **MLMTA**. *[S.l.: s.n.]*, 2006. p. 236–241.

ZHAO, Haipeng; ZHANG, Jinjie; JIANG, Zhinong; WEI, Donghai; ZHANG, Xudong; MAO, Zhiwei. A new fault diagnosis method for a diesel engine based on an optimized vibration mel frequency under multiple operation conditions. **Sensors**, v. 19, n. 11, 2019. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/19/11/2590.

ZHONGHUA, Yao; LINGDA, Wu. 3d-parallel coordinates: Visualization for time varying multidimensional data. *In*: IEEE. **2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)**. *[S.l.]*, 2016. p. 655–658.

ZIMEK, Arthur; SCHUBERT, Erich; KRIEGEL, Hans-Peter. A survey on unsupervised outlier detection in high-dimensional numerical data. **Statistical Analysis and Data Mining**, John Wiley & Sons, Inc., New York, NY, USA, v. 5, n. 5, p. 363–387, out. 2012. ISSN 1932-1864. Disponível em: http://dx.doi.org/10.1002/sam.11161.