Uma Interface de Analítica Visual para Identificação de Perdas Não-Técnicas na Rede de Distribuição Elétrica

Filipe Marçal
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas
Campinas, Brasil
filipemarcal07@hotmail.com

Luís Felipe Granello de Souza
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas
Campinas, Brasil
lfgranellosouza@gmail.com

***Abstract*—In this article, since we are interested in algorithmic power consumption fraud detection applications, such as changepoint detection and *XGBoost*, we present a visual analytics approach in order to identify two kinds of electricity user profiles: (i) the ones with high probability of exhibiting irregular consumption, apart from (ii) those with low probability. Context and relevance are brought to the problem of losses in electrical energy since this issue reaches annually around 8 million Brazilian reais in terms of infrastructure and public service. Mathematical basis which ground multiple changepoint detection and ensemble regression methods, exemplified by *XGBoost*, is shown.**

The qualification of the present work as visual analytic is made by combining (i) a 45,000 rows by 98 columns of a proof of concept dataset of time series and registry information processed by the presented methods; (ii) their respective visualizations in order to highlight consumption anomalies and suspicious profiles in an interactive interface built from the 'shiny' and 'flexdashboard' R libraries and (iii) the decision making and insight generation processes made by a committee composed of field technicians and office experts. Finally, we conclude on the increments that a visual analytics system is able to offer to the dispatch of faults and fraud inspection teams in electric energy consumption end points.

***Index Terms*—Visual analytics, non-technical losses, changepoint detection, *XGBoost*.**

I. Introdução

As perdas de energia elétrica referem-se à energia consumida e não faturada pela distribuidora, devendo-se (i) ao efeito Joule, ou seja, ao fato de o percurso de uma corrente elétrica em um condutor dissipar energia na forma de calor; (ii) a furtos, que são ligações clandestinas desviadas dos medidores de energia; (iii) a fraudes, feitas por manipulação dos medidores e (iv) a falhas, de origem material, pessoal ou processual na medição, leitura e faturamento dessa energia.

Se excluirmos o efeito Joule, típico das perdas técnicas, estima-se que as perdas comerciais das 59 principais distribuidoras de energia elétrica do Brasil sejam da ordem de 5% da energia injetada nas redes de distribuição, o que corresponderia a cerca de 15 TWh, todo consumo do estado de Santa Catarina em 1 ano. Isso, em efeitos tarifários e tributários a R$ 542,00 / MWh, equivaleria a um déficit anual de R$ 8,2 bilhões de reais [1].

Neste artigo, interessados nas aplicações algorítmicas de detecção de fraudes, apresentamos uma abordagem tanto analítica quanto visual para detectar dois tipos principais de perfis de consumidores[1] de energia elétrica: os de alta e os de baixa propensão a apresentar consumo com vestígios de irregularidade. Na primeira seção, são trazidos contexto e relevância para o problema de perdas em energia elétrica, qualificando a abordagem escolhida como uma de analítica visual; na seção "*Trabalhos Relacionados*" são visitados tanto um modelo de probabilidades *a priori* e *a posteriori* para detecção de fraudes de consumo quanto uma interface que estabelece métricas para a fácil distinção de anomalias em consumos elétricos; na seção "*Métodos Empregados*", métodos mais recentes também são visitados, como o de detecção de pontos de mudança e o de classificação através do *XGBoost*, passando-se por seus fundamentos matemáticos e concedendo-se referência adicional especializada; na seção "*Resultados: Uma Interface para Analítica Visual de Perdas Não-Técnicas*", a interface para visualização, interação e suporte à tomada de decisão, criada na linguagem *R*, é detalhada, mostrando-se como os métodos discutidos foram embarcados e, finalmente, conclui-se sobre os incrementos que um sistema de analítica visual é capaz de oferecer ante a decisão de fiscalizar ou não pontos finais de consumo de energia elétrica a fim de se buscar avarias ou fraudes.

A. Contexto e Relevância

Perdas energéticas, no contexto de distribuição de energia elétrica, é um assunto extensamente abordado por sua relevância em infraestrutura e tecnologia. Economicamente, os índices de perdas de energia elétrica são frequentemente relacionados à qualidade de gestão de qualquer agente da cadeia de geração, transmissão e distribuição de energia elétrica por serem consequência tanto de processos comerciais internos – como leitura de medidores, faturamento e arrecadação –

[1]Embora "consumidor" soe como um termo mais preciso para se referir ao usuário final da cadeia de transformação de energia elétrica em um mercado cativo, a adoção de "cliente" também é justificável para os casos em que a cultura organizacional é centrada no cliente.

quanto de investimentos em tecnologia para blindagem da rede de distribuição e de sistemas de medição, a fim de prevenir o manuseio não autorizado. A gestão das perdas em energia elétrica é transversal dentro dos departamentos de uma distribuidora, de modo que o universo tomado aqui para estudo, em caráter de prova conceitual, diz respeito às fraudes externas, localizadas nos pontos finais de consumo (residências, indústrias, estabelecimentos comerciais etc.), generalizados como "unidades consumidoras". Sendo também atribuição da distribuidora o combate às perdas, deve existir uma estratégia bem embasada e otimizada de seleção e apontamento preditivo de unidades consumidoras com alta propensão a fraudar seu consumo, a fim de que sejam submetidas à fiscalização técnica de suas instalações elétricas.

Como nossa abordagem agrega modelos matemáticos de classificação baseados em comportamento à experiência conjunta de um comitê de pessoas especializadas, que envolve desde analistas administrativos a agentes de campo, ela se enquadra na proposta de Keim et al. [2] acerca de analítica visual: a combinação, através da visualização, de esforços distintos entre processamento humano e eletrônico para se alcançar os resultados mais efetivos. Nesse contexto, propomos uma interface visual que exibe a tal comitê os principais resultados analíticos obtidos pelo processamento dos dados de um conjunto de unidades consumidoras, a fim de que decida, ponto a ponto, se uma investigação mais a fundo é cabível ou não. Um conjunto com cerca de 45.000 observações com 98 atributos foram os dados utilizados para esse fim, compostos por: histórico de consumo dos últimos 64 meses, coordenadas georreferenciadas, variáveis categóricas de agrupamento como tipo de fase elétrica, subgrupo de tensão elétrica e classificação da atividade fim da unidade consumidora, dentre outros. As duas técnicas envolvidas no processamento desses dados foram a detecção de ponto de mudança e o *extreme gradient boosting*.

II. Trabalhos relacionados

Chavarro [3] propõe um modelo para classificação entre clientes de consumo de energia elétrica regular e irregular utilizando uma abordagem bayesiana com probabilidades *a priori* e *a posteriori* para definir o índice de potencialidade de infração IPI, reunindo, em seguida, os consumidores em quatro grandes grupos: usuários de alto de IPI – possíveis fraudadores, usuários com consumo muito inferior ao consumo médio do grupo, usuários com consumo muito superior ao consumo médio do grupo e usuários sem consumo. Dentre suas conclusões, constam a relevância estatística do tipo de tarifa e de classe[2] de consumo para predição de valores consumidos, além de uma acurácia de aproximadamente 30% em seu modelo de detecção, podendo ser incrementada de acordo com o ajuste de parâmetros dados por probabilidades *a priori*.

Janetzko et al. [4] propõem um sistema de analítica visual interno para a detecção de mudanças inesperadas no padrão de consumo de corporações eletrointensivas, justificando que o custo com energia elétrica compõe grande parte dos custos operacionais dessas organizações. Os autores apresentam também, a partir de séries temporais de valores de consumo, métricas dadas pela diferença entre consumos previstos e observados, normalizadas pelo desvio médio entre consumos gerados por um modelo de predição e os observados em tempo real. Uma análise de similaridade em conjuntos compostos por essas métricas é realizada nos moldes propostos por Bellala et al. [5], transportando a série temporal dessa métrica para o domínio de frequência por meio da Transformada de Fourier. A densidade da distribuição da métrica no espaço MDS[3] reduzido é então mapeada em uma matiz de cores a fim de que ela seja visualizada por analistas e tenha instantes de potencial interesse realçados.

III. Métodos empregados

Em certa medida, este trabalho utiliza substrato semelhante ao dos supracitados por propor um outro sistema de analítica visual, desta vez, externo, direcionado à detecção de mudanças no padrão de consumo das unidades consumidoras abastecidas por uma distribuidora, mudanças essas capazes de sinalizar uma utilização de energia com vestígios de fraudes ou avarias. Todavia, nossa proposta difere-se da de Janetzko et al. [4] por não se basear em uma visualização por padrão recursivo da série temporal de consumo mas sim por identificar intervalos com padrões de consumo bem definidos e possibilitar, interativamente ao usuário, a escolha da quantidade aproximada desses padrões especificando-se métodos para a detecção bayesiana de pontos de mudança. Nesse sentido, esta abordagem é mais próxima à de Chavarro [3], além de apresentar um significativo incremento em relação a esta, já que embarca o método bayesiano em um painel interativo. Ademais, é incluído também um modelo de regressão multivariável, obtido pelo método de impulso de gradiente extremo (*XGBoost*), que tem se destacado no cenário de detecção de anomalias. Por meio de um valor-limiar (*threshold*), também ajustável pelo usuário, aplicado sobre a pontuação que o *XGBoost* calcula automaticamente a partir dos dados obtidos por unidade consumidora, o modelo passa a funcionar também como um classificador de consumidores regulares ou irregulares. Enquanto a pontuação de um consumidor, dada pelo *XGBoost*, indica sua propensão de estar incorrendo em uma série de consumo irregular, a diferença entre padrões de consumo possibilita estimar o montante energético não faturado pela distribuidora, o que subsidia as decisões de seus analistas quanto ao despacho de uma fiscalização, somando-se outras variáveis como rota, acessibilidade ao local, disponibilidade de equipes de campo, risco-retorno e etc.

[2]Categorização segundo a atividade fim de uma unidade consumidora, exemplificada por residencial, industrial, comercial, poder público etc.

[3]Segundo Mead [6], *Multidimensional scaling* (MDS) é uma forma de visualizar o nível de similaridade dentre as ocorrências de um conjunto de dados, sendo utilizada para representar informações sobre a distância entre pares em um conjunto de n objetos através da configuração de n pontos mapeados no espaço cartesiano abstrato.

A. *Detecção Bayesiana de Ponto de Mudança*

Detecção de pontos de mudança é a identificação de mudanças abruptas nos parâmetros geradores de dados sequenciais. Em séries temporais, ocupa-se em encontrar o instante t em que mudanças na média, variância, correlação, densidade espectral[4] e etc. tornam-se significativamente perceptíveis.

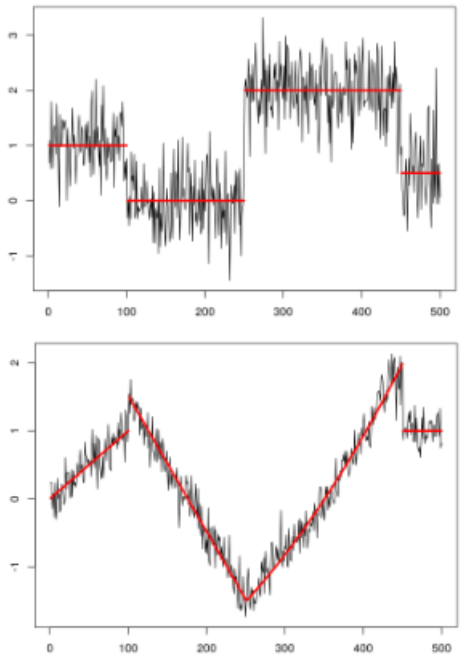

Fig. 1. Exemplos de duas séries temporais com três pontos de mudança identificados em cada uma. O primeiro exemplo ilustra mudanças na média ao longo da primeira série e o segundo exemplo ilustra mudanças na variância ao longo da segunda série. Extraído de Killick [7].

Para tal, pode-se ser utilizado o teste de razão de verossimilhanças[5]. Seja $\boldsymbol{x}(\tau)$ com τ um índice discreto e $\boldsymbol{x}_{a:b}$ o vetor dado por: $\begin{bmatrix} x(a) & x(a+1) & \dots & x(b) \end{bmatrix}$. O teste de razão de verossimilhanças LR em $\tau = t$ em uma série de tamanho $n-1$ é expresso por:

$$LR(t) = \ell(\boldsymbol{x}_{1:t}) + \ell(\boldsymbol{x}_{t+1:n}) - \ell(\boldsymbol{x}_{1:n}) \tag{1}$$

O t^* em que há a mudança mais significativamente perceptível pode ser calculado por:

$$t^* = argmax\{\ell(\boldsymbol{x}_{1:t}) + \ell(\boldsymbol{x}_{t+1:n}) - \ell(\boldsymbol{x}_{1:n})\} \tag{2}$$

[4]A densidade espectral de um sinal $x(t)$ descreve-o em termos de seus componentes de frequência. De acordo com a análise de Fourier, qualquer sinal pode ser decomposto em um número discreto ou um espectro contínuo de frequências.

[5]A verossimilhança de um conjunto de parâmetros Θ, dado um conjunto X de n observações, é igual a função de densidade de probabilidade das n observações X dado seu conjunto de parâmetros Θ. Por exemplo, a verossimilhança associada a distribuição gaussiana é:

$$\ell(\mu, \sigma|x) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Se, contanto, estivermos interessados em outros pontos de mudança, pode-se definir a penalidade λ de modo que t é um ponto de mudança se $LR(t) > \lambda$. O teste de razão das verossimilhanças também pode ser adaptado para k pontos de mudança somando-se a ele uma função $f(k)$ multiplicada por λ. Esse tipo de comparação é a essência por trás de todas as técnicas de detecção de ponto de mudança, mesmo que o teste de razão das verossimilhanças não seja utilizado.

Adam e McKay [8] categoricamente apresentam um algoritmo bayesiano de inferência, em tempo real, focado em filtragem causal preditiva e geração de distribuições precisas, do próximo dado desconhecido de um sinal $x(t)$, dada apenas a sequência temporal já observada, o que é uma situação recorrente em aplicações de inteligência de máquina. Os autores assumem que os dados para cada partição ρ de uma série temporal são independente e identicamente distribuídos (i.i.d.) conforme alguma função densidade de probabilidade $P(x_t|\eta_\rho)$. Os parâmetros η_ρ, $\rho = 1, 2, ...$, também são tomados como i.i.d. e a probabilidade *a priori* sobre um intervalo $g_i = \boldsymbol{x}_{a:(b-1)}$ entre pontos de mudança a e b é denotada por $P_{gap}(g_i)$.

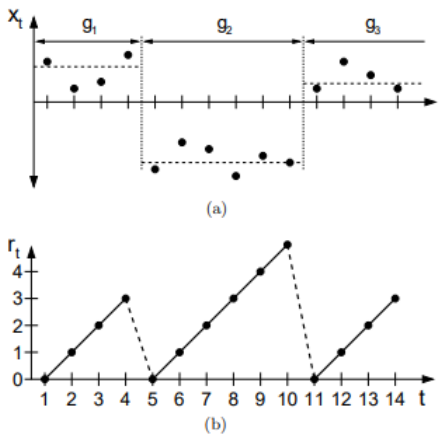

Fig. 2. (a) Dados univariados divididos em três partições e g_1, g_2 e g_3. (b) "Comprimentos percorridos" r_t como funções do tempo que vão a zero quando um ponto de mudança ocorre. Extraído de Adam e McKay [8].

Como um dos objetivos é a estimativa da probabilidade *a posteriori* sobre o "comprimento percorrido" r_t em questão da série temporal – o tempo desde o último ponto de mudança, considerando-se os dados observados até o momento –, as observações em r_t são denotadas por $\boldsymbol{x}_t^{(r)}$ e a probabilidade *a posteriori* do próximo termo da sequência conhecendo-se seu histórico é denotada por:

$$P(x_{t+1}|\boldsymbol{x}_{1:t}) = \sum_{r_t} P\left(x_{t+1}|r_t, \boldsymbol{x}_t^{(r)}\right) P(r_t|\boldsymbol{x}_{1:t}) \tag{3}$$

e também:

$$P(r_t|\boldsymbol{x}_{1:t}) = \frac{P(r_t, \boldsymbol{x}_{1:t})}{P(\boldsymbol{x}_{1:t})} \tag{4}$$

Como a distribuição preditiva $P(x_t|r_{t-1}, \boldsymbol{x}_{1:t})$ é função somente das observações $\boldsymbol{x}_t^{(r)}$, pode-se gerar um algoritmo recursivo baseado nos cálculos (1) da probabilidade *a priori* de r_t dado r_{t-1} e (2) da distribuição preditiva dos novos dados observados a partir do histórico desde o último ponto de mudança. Esse esquema algorítmico é integrado pelos autores a um modelo conjugado exponencial, particularmente conveniente para tal. Detalhes com maior profundidade podem ser conferidos em Adam e McKay [8].

O pacote '*changepoint*', elaborado em linguagem *R*, é extensivamente explorado por Killick [7], que também elenca referências correlatas. Graças a ele, a implementação de métodos de detecção de ponto de mudança na interface de analítica visual tornou-se possível e consistente, já que possibilitou o aproveitamento de funções bem formuladas, estruturas de dados como a *S4*[6] e três diferentes métodos de busca, desenvolvidos a partir do teste da razão das verossimilhanças adaptado para k pontos de mudança posicionados em $t = (t_0, t_1, ..., t_{k+1})$, acrescido de uma penalidade λ e uma função penalidade adequada $f(k)$:

$$\min_{k,t} \left\{ \sum_{i=1}^{k+1} \left[-\ell \left(\boldsymbol{x}_{t_{i-1:i}} \right) \right] + \lambda f(k) \right\} \tag{5}$$

1) Uma mudança no máximo (*AMOC – At Most One Change*): encontra um único ponto de mudança, que maximiza a diferença de parâmetros nas duas partições da série temporal;
2) Segmentação binária (Scott e Knott [9]): encontra vários pontos de mudança de modo aproximado mas computacionalmente rápido, com ordem $\mathcal{O}(n \cdot \log(n))$;
3) Tempo linear exato podado (*PELT – Prunned Exact Linear Time*) (Killick et al. [10]): encontra vários pontos de mudança de modo exato e rápido, com sua pior ordem sendo $\mathcal{O}(n^2)$.

.

Esses três métodos podem ser explorados pelo usuário final na interface construída.

B. XGBoost - Extreme Gradient Boosting

O algoritmo de aprendizado de máquina denominado *Extreme Gradient Boosting*, comumente abreviado para *XGBoost*, é considerado por muitos o estado da arte para a solução de diversos tipos de problemas de regressão, classificação e ranqueamento, incluindo detecção de anomalias, predição de comportamento de clientes e detecção de fraudes de energia [11]. Dentre os fatores que fazem desse pacote de código-aberto ser frequentemente aplicado a conjuntos de dados estruturados, estão:

- Capacidade de realizar o treinamento[7] de um grande volume de dados em apenas uma máquina;
- Alta velocidade de processamento, uma vez que realiza processamento distribuído e paralelo;
- Tratativa automática de dados faltantes ou esparsos;
- Interação com as características do modelo encontrado.

O modelo é composto por K árvores de decisão, sendo cada árvore representada por uma função que calcula a predição de um evento ocorrer a partir de determinadas características carregadas em cada ramo da árvore. O diferencial do modelo proposto por Chen [11] é que, a partir desse conjunto de árvores, encontram-se funções regularizadas, que, com baixa complexidade, contribuem para predições com baixa variância e mais assertivas. A função-objetivo Obj do modelo pode ser descrita por:

$$Obj = \sum_{i=1}^{N} \ell \left(y_i, \bar{y}_i \right) + \sum_{k=1}^{K} \Omega \left(f_k \right), \tag{6}$$

em que N é igual ao número de amostras e K é a quantidade de árvores de decisão utilizadas no modelo.

O primeiro termo da função representa o erro a ser minimizado entre a predição $\bar{y}_i$ e o valor-alvo y_i. O termo adicional, por sua vez, realiza a regularização das funções, ajudando na suavização dos pesos de cada ramo e, consequentemente, evita o *overfitting*[8]. Caso esse termo seja ajustado para zero, o modelo final torna-se a representação do modelo tradicional *Gradient Boosting*.

Neste trabalho, utilizou-se o pacote '*xgboost*' disponibilizado em *R* para realizar o treinamento do modelo e, consequentemente, obter a probabilidade de irregularidade para cada unidade consumidora. Para selecionar as variáveis utilizadas para a construção do modelo, foi feita uma análise exploratória dos dados de forma a identificar aquelas que mais se mostraram relevantes para a definição do perfil do cliente. A **Tabela I** apresenta as variáveis utilizadas no modelo de classificação utilizado:

TABLE I
VARIÁVEIS DO MODELO *XGBoost*

Nome da Variável	**Valor Esperado**	**Tipo Variável**
Média de consumo dos últimos 12 meses	-	Inteira
Tipo de Fase	MO/BI/TR	Categórica
Tipo de Local	UB/RR	Categórica
Grupo de Tensão	A/B	Categórica
Classe de Consumo	Residêncial/Industrial	Categórica
Carga Instalada	-	Inteira
Perfil	Irregular/Regular	Categórica

IV. RESULTADOS: UMA INTERFACE PARA ANALÍTICA VISUAL DE PERDAS NÃO-TÉCNICAS

Para realizar a construção do sistema de analítica visual, utilizou-se o *software RStudio* como ambiente de desenvolvimento. A criação da interface analítica foi realizada a partir de

[6]Sistema da linguagem *R* orientado a objeto e representado por uma lista de atributos (*slots*.) com aplicações mais restritas e mais semelhantes à orientação a objetos de outras linguagens.

[7]Atribuição de valores aos parâmetros do modelo a partir da exposição intensiva do algoritmo a dados e saídas rotuladas no caso de aprendizado supervisionado

[8]Sobretreinamento de um modelo com perda de sua capacidade de predição generalizada.

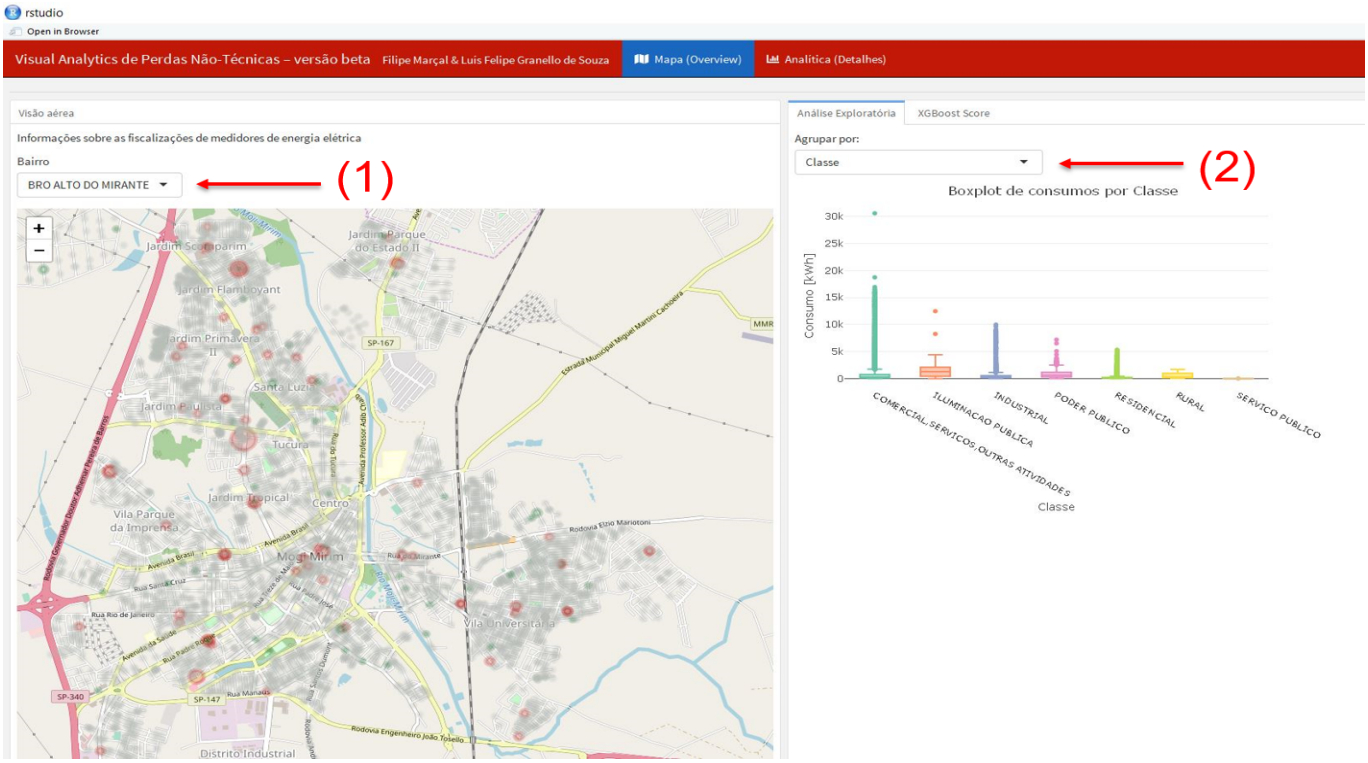

Fig. 3. Sistema de Analítica Visual – Página "Mapa (*Overview*)", abas "Visão Aérea" e "Análise Exploratória", essa mostrando agrupamentos de valores consumidos por classe de consumo

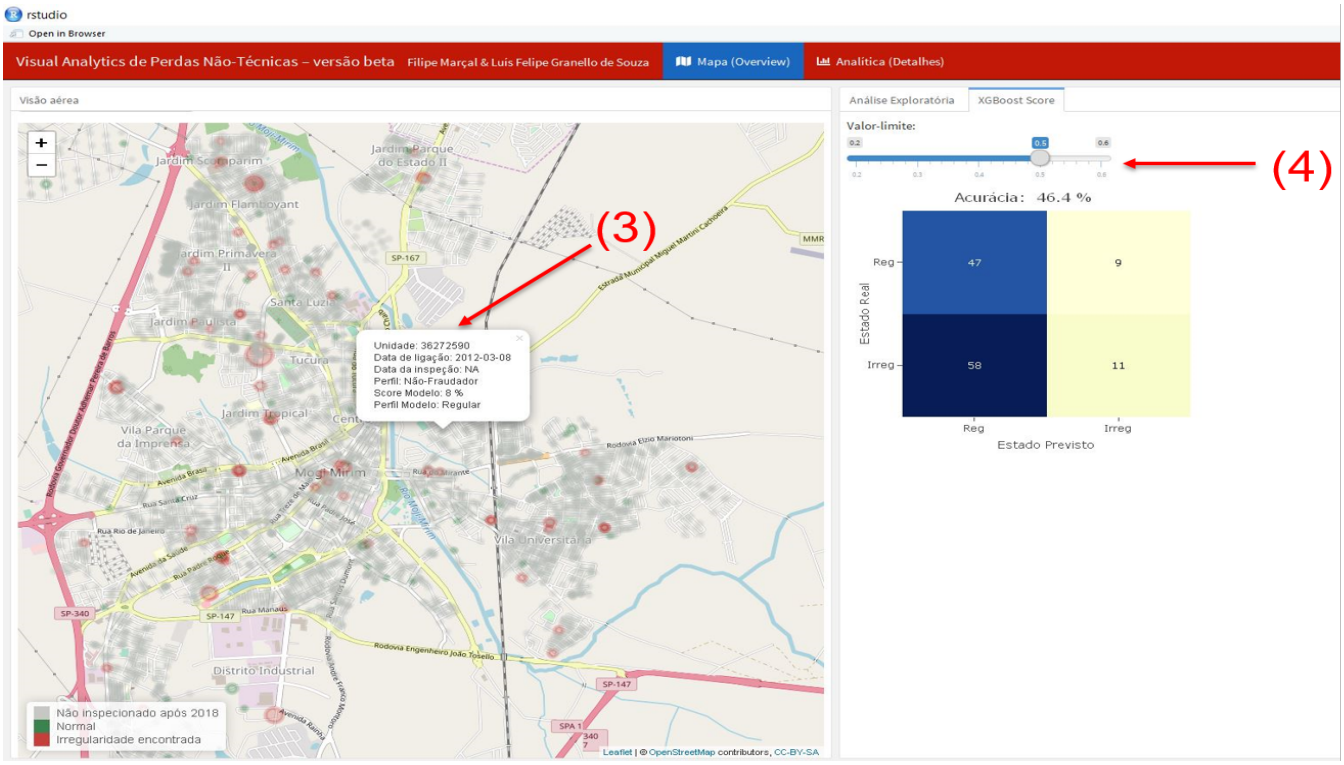

Fig. 4. Sistema de Analítica Visual – Página "Mapa (*Overview*)", aba "*XGBoost Score*"

dois pacotes principais, '*flexdashboard*' e '*shiny*' e a junção de ambos permitiu a criação de *frames* com diferentes tipos de informações e formas de interatividade com o usuário, como a utilização de filtros seletores, abas, painéis, ícones e etc.

Para plotar os gráficos presentes no sistema, como histogramas, *boxplots*, mapa de calor e séries temporais, utilizou-se o pacote '*plotly*' como ferramenta, uma vez que ele permite a interação do usuário com os gráficos de acordo com o posicionamento do *mouse*, facilitando o seu entendimento e exploração. O mapa presente em uma das abas do sistema teve suas camadas renderizadas através do pacote '*leaflet*'.

Além disso, como citado anteriormente neste artigo, os pacotes '*changepoint*' e '*xgboost*' foram utilizados para permitir, respectivamente, a análise exploratória dos dados de consumo e o cálculo da probabilidade de irregularidade de uma unidade consumidora. Vale ressaltar que o tempo de processamento do sistema depende do volume de dados que se deseja analisar. Isso posto, para a importação da base de dados apresentada neste trabalho, o tempo aproximado de processamento de *backend* e de *frontend*, testado diversas vezes em um computador de 8 GB de RAM, 12 MB de cache e 2,7 GHz de clock e placa de vídeo *onboard*, foi de 5 minutos. Esse tempo pode variar em função das especificações técnicas do computador designado para executar o *script* construído.

Na Fig. 3 temos a visão geral do sistema, inspirados pelo mantra atualizado de B. Schneiderman: "Visão geral (com destaque aos pontos de interesse) primeiro, detalhes sob demanda". À esquerda da tela, na página "Mapa (*Overview*)", a visão da localidade em que se deseja realizar a análise dos dados. É possível selecionar em (1) o bairro de interesse ou ainda, a partir da manipulação do mapa, explorar outras regiões. Na parte direita da tela, representada na Fig.3, temos duas abas: a primeira, intitulada "análise exploratória", permite a visualização dos *boxplots* de consumos dos clientes com diferentes tipos de agrupamentos já que, em (2), é possível agrupar clientes por classe principal, bairro, tipo de fase, tensão e classe de consumo.

Na aba "*XGBoost*", mostrada na Fig. 4, temos uma matriz de confusão do modelo utilizado para definir o *score* dos clientes em análise. Em (4), a fim de que o modelo funcione como um classificador, pode-se variar o valor de *threshold* e, pelo teste lógico de que, se o *score* de um consumidor é maior ou igual ao valor de *threshold*, seu perfil é irregular, caso contrário, é regular. O desempenho do modelo, em função do *threshold* escolhido, ao se comparar perfis previstos e reais para as unidades consumidoras já inspecionadas é expresso nas variações de sua acurácia[9]. Ainda na Fig. 4, observamos em (3) que, ao clicar em uma unidade consumidora presente no mapa, são disponibilizadas, ao usuário, informações sobre o cliente: data de ligação, data de inspeção, perfil, *score* e classificação em regular ou irregular conforme o *XGBoost*.

A página nomeada como "Analítica (detalhes)" que está apresentada na Fig.5, exibe mais detalhes para os dados que estão sendo analisados. O primeiro recurso à disposição do usuário está destacado em (5), pelo qual, através da seleção de uma unidade consumidora, tem-se a possibilidade de aplicar um dos três métodos de detecção do ponto de mudança aplicado à série temporal dos últimos 64 meses de consumo da unidade selecionada.

Além disso, em (6), tem-se a opção de seleção da largura da faixa de consumo para o histograma de consumo do cliente. Apresenta-se também abaixo, logo abaixo, o histograma de consumo de toda a classe de consumo a que unidade selecionada pertence, sendo viável comparar as curvas a fim de detectar possíveis anomalias no comportamento de consumo do cliente que está sendo analisado em relação à sua classe. Em (7), por sua vez, o usuário poderá visualizar o *score* do cliente e, consequentemente, o tipo de perfil indicado pelo modelo para aquela unidade: regular ou irregular.

Por fim, como mostrado na Fig. 6 a aba indicada em (8), nomeada como "Dados adicionais", permite a visualização de mais informações para a unidade consumidora, como endereço, carga elétrica instalada, em kW, *score* do modelo e média e desvio padrão do consumo de energia elétrica, em kWh.

CONCLUSÃO

O estudo de perdas não-técnicas sempre foi um tema importante e de alta complexidade no setor de distribuição de energia elétrica no Brasil. Assim sendo, este trabalho proporcionou, através de análise exploratória de dados, a utilização de três recursos essenciais para sua qualificação em analítica visual: (i) visualização de dados, (ii) criação de modelos e, principalmente, (iii) o ganho de conhecimento e *insights* sobre o problema.

A partir da utilização da interface, é possível explorar diversos tipos de visualização dos dados. Os recursos gráficos utilizados são de variados tipos e permitem a comparação de dados para diferentes agrupamentos. Além disso, graças aos pacotes '*shiny*' e '*plotly*', os gráficos são interativos, tornando simples a exploração e oferecendo mais atratividade ao usuário final. As técnicas de detecção de ponto de mudança utilizadas se mostraram uma excelente ferramenta de visualização evidenciando diferenças relevantes nos padrões no histórico de consumo dos clientes.

O modelo de análise preditiva do problema, que alcança por volta de 40% de acurácia para intervalos ótimos de *threshold*, foi construído a partir da exploração prévia dos dados. O ambiente de visualização, como um todo, permite que o usuário consiga correlacionar as várias formas de visualização apresentadas com o resultado final do *XGBoost* e, dessa forma, também proporciona a retroalimentação do modelo de classificação, abrindo caminho para o refinamento de seus parâmetros e, consequentemente, para a melhoria de sua performance. Vale ressaltar que a busca por um tempo de resposta razoável durante as interações do usuário com a interface refletiram na otimização de *scripts* e constituíram uma das partes desafiadoras do trabalho.

[9]Acurácia é expressa pela razão entre verdadeiros positivos e verdadeiros negativos resultantes da classificação sobre o total de classificações feitas

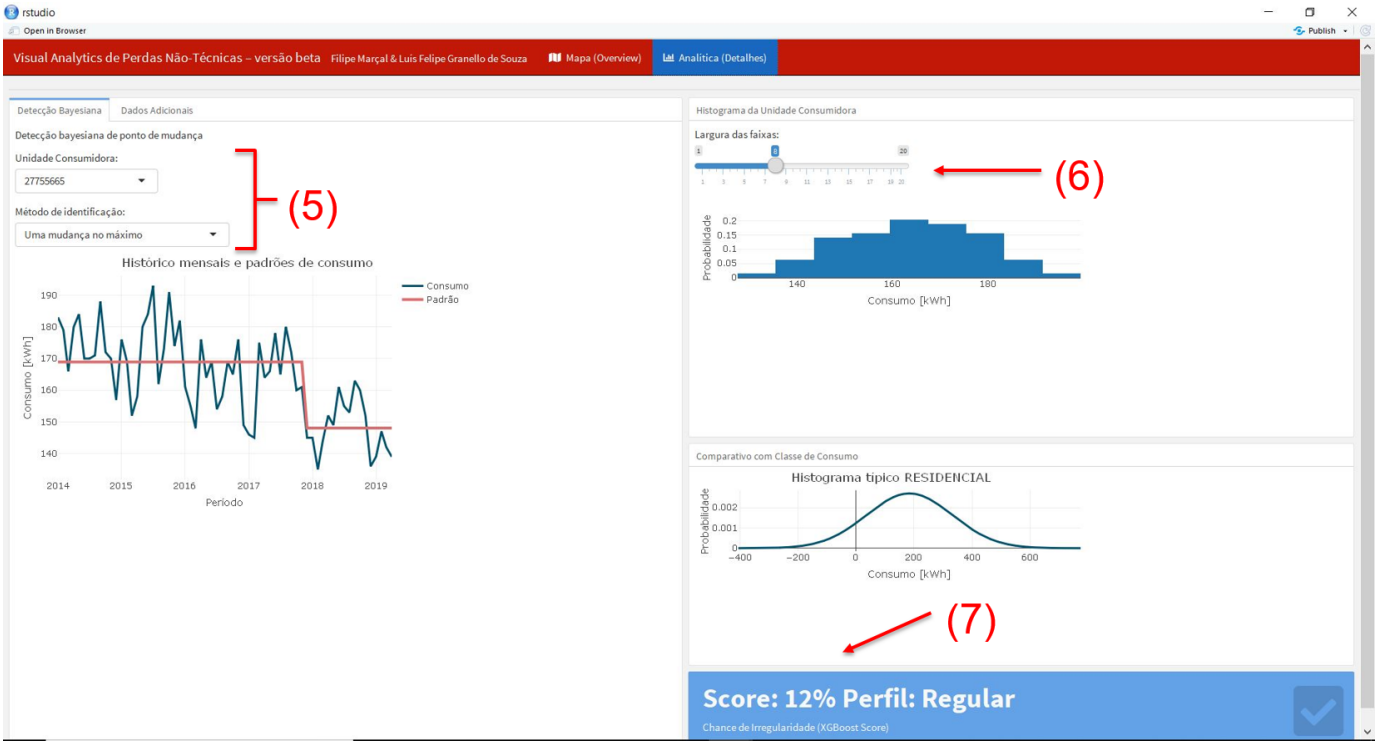

Fig. 5. Sistema de Analítica Visual – Página "Analítica (*Detalhes*)", aba "Detecção Bayesiana"

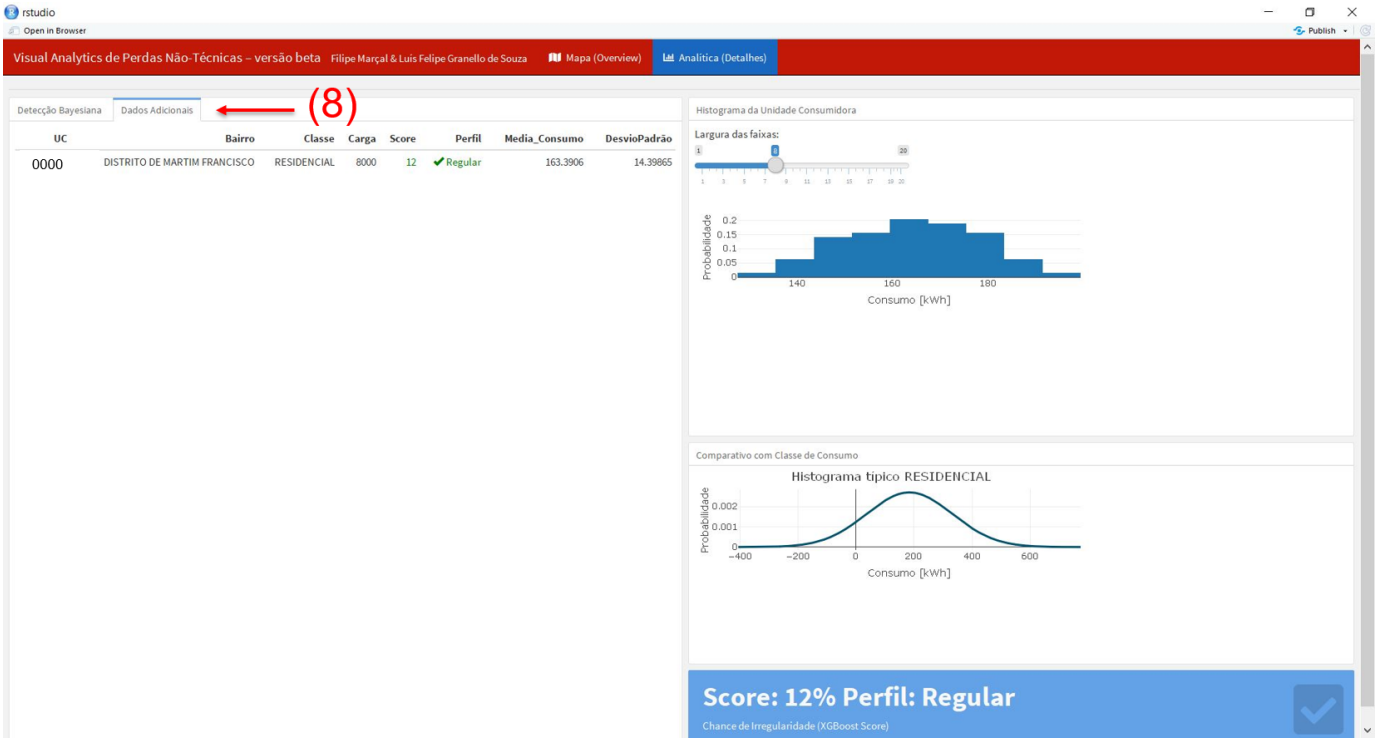

Fig. 6. Sistema de Analítica Visual – Página "Analítica (*Detalhes*)", aba "Dados Adicionais"

Um dos principais recursos de um sistema como esse é a geração de *insights* para a tomada de decisão. Para que isso ocorra, o usuário precisa obter conhecimento sobre o problema. Esse conhecimento é adquirido seja a partir dos gráficos apresentados ou partir do modelo desenvolvido e constantemente aprimorado. A junção desses fatores, segundo Keim [2], irá formar um ciclo capaz de suportar os interessados na solução do problema na escolha das melhores estratégias para o negócio.

Como trabalhos futuros, este trabalho poderá ter sua aplicabilidade testada em um ambiente de planejamento e recuperação de energia de uma distribuidora de energia elétrica. Sua utilização, em caráter de prova de conceito, viabilizará a avaliação, com maiores detalhes, de sua efetividade e direcionará eventuais pontos principais de melhoria como: tipos de gráficos, relevância de filtros, enriquecimento com novos tipos de dados, refinamento do modelo de predição, utilização de técnicas aprimorada para processamento de séries temporais e detecção de pontos de mudança, aperfeiçoamentos de *design* e usabilidade, dentre outros.

REFERENCES

[1] Instituto Acende Brasil (2017). "Perdas Comerciais e Inadimplência no Setor Elétrico". White Paper 18, São Paulo, 40 p

[2] D. A. Keim, "Mastering the Information Age – Solving Problems with Visual Analytics". 2010.

[3] A. Chavarro (2017). "Probabilistic Model for Determining Non-Technical Losses in a Power Distribution System. Consultoría Colombiana S.A, Cra 20 No. 37-28, Bogotá, Colombia.

[4] H. Janetzko, F. Stoffel, S. Mittelstädt, D. A. Keim (2014). "Anomaly Detection for Visual Analytics of Power Consumption Data". Computers e Graphics. 38. 27–37.

[5] G. Bellala, M. Marwah, M. Arlitt, G. Lyon, C. E. Bash. "Towards an understanding of campus-scale power consumption". In: Proceedings of the 3rd ACM workshop on embedded sensing systems for energy-efficiency in buildings. ACM; 2011. p. 73–8.

[6] A. Mead (1992). "Review of the Development of Multidimensional Scaling Methods". Journal of the Royal Statistical Society. Series D (The Statistician). 41 (1): 27–39.

[7] R. Killick (2017). "Introduction to optimal changepoint detection algorithms. Disponível em: http://members.cbio.mines-paristech.fr/ thocking/change-tutorial/RK-CptWorkshop.html. Acesso em 29 jun. 2019

[8] R. P. Adams, D. MacKay (2007). "Bayesian Online Changepoint Detection". Arxiv preprint arXiv:0710.3742

[9] A. J. Scott, M. Knott (1974). "A cluster analysis method for grouping means in the analysis of variance". Biometrics, 30(3):507–12.

[10] R. Killick, P. Fearnhead, I. A. Eckley (2012). "Optimal Detection of Changepoints With a Linear Computational Cost", Journal of the American Statistical Association, 107:500, 1590-8.

[11] T. Chen, C Guestrin, "XGBoost: A Scalable Tree Boosting System" Universidade de Washington. Washington, Junho 2016.