IMPLEMENTAÇÃO DE UM MÉTODO DE DETECÇÃO DE ATAQUES À INDÚSTRIA 4.0

Víctor Costa da Silva Campos*, Paulo Ricardo de Freitas†

* *Universidade Federal de Ouro Preto, Engenharia Elétrica*
João Monlevade, Minas Gerais, Brasil

† *Universidade Federal de Minas Gerais, Departamento de Engenharia Eletrônica*
Belo Horizonte, Minas Gerais, Brasil

Emails: kozttah@gmail.com, paulo94freitas@gmail.com

Abstract— Industries 4.0 characterize a technologcal revolution in critical infrastructures. The increasingly need for connectivity and access to information breaks with the standards of previous supervisory systems models. The use of cyberphysical systems allows great improvements in the control and supervision of industrial processes. On the other hand, interconnectivity is related to reduced security, which makes infrastructures vulnerable to digital attacks. This work presents a method for detecting attacks using a state estimator with minimal quadratic error.

Keywords— Industry 4.0, supervisory systems, Cyber-Physical System, Attack detection, state estimator.

Resumo— As Indústrias 4.0 caracterizam uma revolução tecnológica nas infraestruturas críticas. A crescente necessidade de conectividade e acesso a informação rompe com os padrões dos modelos de sistemas supervisórios anteriores. O aumento da utilização dos sistemas ciberfísicos possibilitam grandes melhorias no controle de processos industriais. Por outro lado, a interconectividade está relacionada à redução da segurança, o que deixa as infraestruturas vulneráveis a ataques digitais. Esse trabalho apresenta um método para detecção de ataques utilizando um estimador de estados com erro quadrático mínimo.

Palavras-chave— Indústrias 4.0, sistemas supervisórios, sistemas ciberfísico, detecção de ataque, estimador de estados.

1 Introdução

As indústrias vêm passando por um processo evolutivo constante. De tempos em tempos novas tecnologias são empregadas caracterizando uma nova Revolução. No início do século XXI as infraestruturas mais tecnológicas investem na aplicação de Sistemas Ciberfísicos (CPS — do inglês *Cyber Physical System*) e Internet das Coisas (IoT — do inglês *Internet of Things*) para integrar as indústrias inteligentes. Nessas indústrias há intensa troca de dados por meio de diferentes tecnologias: wi-fi, fibra óptica e *bluetooth* são alguns exemplos. Os dados são trocados entre os dispositivos de campo, que ganharam capacidade de realizar processamento local, algo completamente diferente do que até então era utilizado em sistemas supervisórios.

Nas indústrias inteligentes, o gerenciamento dos processos também pode ocorrer por meio da Internet e da computação em nuvem, isso representa uma grande ruptura com os modelos originais de supervisórios. Antes do surgimento da Indústria 4.0, os sistemas supervisórios eram instalados em redes isoladas, o que acarretava em sistemas protegidos dos modernos ciberataques possibilitados pela tecnologia de comunicação e informação.

Todas essas características estão presentes nas Indústrias 4.0, que compõe a Quarta Revolução Industrial, marcada por massiva troca de dados na indústria, comunicação entre diferentes dispositivos de plantas industriais, maior intimidade entre sistemas físicos e cibernéticos, *big data* e a crescente necessidade de informação em tempo mínimo. Ao mesmo tempo que essa atual Revolução traz inúmeros benefícios para os negócios, quando infraestruturas críticas aplicam tecnologias de informação e comunicação elas deixam de ser isoladas e se tornam vulneráveis a ciberataques, sejam eles específicos a CPSs ou comuns à informática.

1.1 Exploração de vulnerabilidades

Na literatura é possível constatar um vasto registro de incidentes envolvendo infraestruturas críticas que foram originados por ataques aos CPSs (Miller and Rowe, 2012). Um dos ataques mais importantes da história foi realizado em 2010 na usina de enriquecimento de Urânio de Natanz, no Irã. O incidente foi possível por meio da exploração de falhas até então desconhecidas no sistema operacional Windows para se infiltrar no supervisório. Nesse incidente foi testado o *worm* Stuxnet, que tem a capacidade de tomar controle do CPS sem que o operador tenha conhecimento. Enquanto o agente invasor toma controle dos processos, as entradas e saídas do sistema são camufladas utilizando um ataque de enganação conhecido como *replay* que será explicado na próxima seção.

Após o registro da primeira utilização do Stuxnet, vários outros incidentes com princípio de funcionamento semelhante foram registrados (Zhu et al., 2011). Alguns exemplos que esclarecem

essa semelhança são DUQU(2011) e Flame(2012), abordados com mais detalhes em (Miller and Rowe, 2012).

Levando em consideração que grande parte dos incidentes ocorridos em infraestruturas críticas se embasam na execução de um ataque *replay* iniciado previamente, uma forma de identificar a possibilidade de um ataque qualquer ao sistema é pela confirmação do ataque *replay*. Portanto, este trabalho foca no estudo comportamental de sistemas de controle e reconhecimento de parâmetros anômalos que configurem o *replay*.

1.2 Organização

Este trabalho se compromete a apresentar métodos de detecção de ataques a sistemas de controle, algo que se tornou muito mais tangível na Quarta Revolução Industrial. Na Seção 2 são apresentados brevemente os métodos mais comuns utilizados para detecção e identificação de falhas e ataques a sistemas, discutindo seus diferenciais. Em seguida é escolhido o método mais vantajoso para ser implementado em um sistema real. Os resultados obtidos pelo método serão discutidos na Seção 3 e, por último, na Seção 4 serão apresentas algumas conclusões acerca do método escolhido, no que se refere à eficiência e qualidade dos resultados, assim como melhorias que devem ser feitas para futuros projetos.

2 Proposta do método de detecção

Ao longo dos anos, um vasto número de trabalhos foram apresentados com métodos de detecção e identificação de falhas em sistemas de controle (Billings et al., 1989; Mo et al., 2014; Nishiya et al., 1982; Zhu and Martinez, 2011). Para entender melhor sobre o assunto é necessário determinar o significado de cada um dos termos expostos previamente.

As falhas de um sistema de controle se referem às divergências comportamentais entre o que se exprime na prática, que pode ser observado, e o esperado teoricamente ou por meio de simulações. Portanto, pode-se determinar como falhas do sistema mudanças súbitas dos estados do sistema, o que se dá por fenômenos não modelados que compõem o processo em si.

Como exemplo desse tipo de falha, considere uma planta de distribuição de energia elétrica. Uma vez que se obtém o modelo de todo o sistema, é possível ter uma previsão de seu comportamento enquanto ele se mantém parecido com o que era no instante em que o modelo foi estimado. Entretanto, o aumento da demanda de carga em determinado período do dia ou a ocorrência de uma falta, são eventos que modificam o comportamento do sistema e o modelo estimado passa a não ser eficiente como antes, o que caracterizam falhas do sistema (Nishiya et al., 1982).

O segundo tipo de falha de sistemas de controle está relacionado à parte física e técnica do processo. Ao passo que no primeiro caso, o operador do sistema ainda tem dados corretos dos estados do sistema, no segundo tipo de falha, prevalecem os erros de medições, falhas de transmissão de dados ou ataques ao CPS (Forti et al., 2016), o que leva a dados imprecisos ou corrompidos (Billings et al., 1989; Pasqualetti et al., 2013).

Este trabalho se baseia nas propostas de (Pasqualetti et al., 2012; Pasqualetti et al., 2013) e (Billings et al., 1989) para desenvolver um método de detecção de falhas ocasionadas por ataque *replay* em sistemas de controle. A escolha de implementar um detector para esse tipo específico de ciberataque se explica no fato dele ser uma ótima escolha de ataque de enganação executado simultaneamente a outro ataque a CPS de natureza distinta. Além disso, o conjunto de características do *replay* torna sua detecção mais simples.

Para entender o ataque *replay* é necessário considerar que os atacantes têm acesso às entradas e saídas do sistema mas desconhecem sua dinâmica. Isso é possível por meio do Stuxnet, por exemplo.

A estratégia do agente malicioso é registrar os dados de entrada e saída do sistema em regime permanente por tempo suficientemente grande para realizar as modificações desejadas no sistema. Após o término dessa etapa, o agente está pronto para sequestrar os dados de entrada e saída do sistema e trocar pela réplica dos dados salvos anteriormente. Durante a injeção de comandos, o operador da planta, que acompanha o processo à distância por meio da Interface Homem Máquina (IHM), estará monitorando apenas os dados replicados. É possível também registrar dados do sistema por um tempo pequeno e concatená-los quantas vezes for necessário (Mo et al., 2014).

A implementação de um detector de ataques é feita por meio da estimação do modelo do sistema que está sujeito inicialmente a entradas desconhecidas. Em seguida, é necessário projetar um observador de erro quadrático mínimo para estimar os estados do modelo (Pasqualetti et al., 2013). A partir das entradas e saídas do sistema são feitas análises do erro entre os estados esperados e obtidos na planta. A detecção do ataque é realizada quando o erro de estimação se torna exagerado.

Por sua vez, a identificação do ataque é feita por meio da comparação entre as saídas conhecidas e as possíveis saídas obtidas para diferentes entradas, sendo que cada conjunto de entradas caracteriza um ataque específico testado, atrelado a esse conjunto está a resposta do sistema quando se encontra sob ataque. Sendo assim, um ataque indetectável é aquele que excita única e exclusivamente o estado zero do sistema, não produzindo

saídas no sistema. Ataques não identificáveis são detectáveis mas seus conjuntos de entradas e saídas do sistema não são reconhecidos (Pasqualetti et al., 2013).

A proposta do monitor para detecção e identificação de ataques é feita a partir da estimação do modelo do sistema sob estudo. Em um primeiro momento, o modelo do sistema genérico pode ser tomado como

$$\begin{aligned} E\dot{x}[k] &= Ax[k] + Bu[k] + \mu[k] \\ y[k] &= Cx[k] + Du[k] + \nu[k] \end{aligned} \tag{1}$$

sendo $E \in \mathfrak{R}^n$, $A \in \mathfrak{R}^{n\times n}$, $B \in \mathfrak{R}^{n\times m}$, $C \in \mathfrak{R}^{m\times n}$, $D \in \mathfrak{R}^{m\times m}$. Os ruídos de processo ($\mu[k]$) e de medição ($\nu[k]$) estão presentes em qualquer processo real e portanto são considerados durante os testes. Esses ruídos são matematicamente descritos como brancos gaussianos e, além disso, possuem média zero e função de correlação $R(k) = \sigma^2\delta(k)$, em que σ^2 é a potência média ou variância do sinal.

Por se tratar de um método para detecção e identificação de ataques do tipo *replay* a CPSs, o modelo é composto por cinco etapas:

1. Estimação do modelo do sistema;
2. Calibração do Filtro de Kalman (FK);
3. Estimação dos estados e saídas;
4. Cálculo da autocorrelação do resíduo;
5. Detecção do ataque.

O método utilizado para estimar o modelo do sistema fica a cargo do projetista. Deve-se reconhecer que o modelo estimado deve ser eficiente e ter um bom desempenho, por conseguinte, a validação do modelo é de extrema importância antes de avançar para a próxima etapa. Neste trabalho foi aplicado o método dos mínimos quadrados com finalidade de estimar um modelo ARX utilizando os registros das entradas e saídas da planta estudada após ter alcançado regime permanente em malha aberta. Um ponto importante a se atentar antes e durante a coleta dos dados utilizados consiste no *setpoint* em torno do qual o modelo será linearizado. Por exemplo, caso deseja-se controlar o nível de um reservatório de água em torno de 80%, é necessário utilizar entradas que excitem o sistema a fim de obter esse nível em regime permanente.

Caso o modelo estimado não represente o sistema com uma boa eficiência, é necessário utilizar outro tipo de modelo para descrever o sistema. Os métodos para estimar os parâmetros do modelo e testar eficiência do mesmo estão bem descritos em (Aguirre, 2004).

Uma vez obtida a representação do sistema sem controlador, é natural a adição de controladores a fim de controlar o processo e obter resultados em um menor tempo, por exemplo. O fato é que a adição de controladores altera o comportamento da planta, afinal, a função de transferência depende tanto do sistema quanto do controlador adicionado. Entretanto, do ponto de vista do Filtro de Kalman, próxima etapa do método de identificação de ataque, o controlador adicionado não afeta a estimação de estados. Isso se explica pelo fato de as entradas do sistema serem precisamente conhecidas, uma vez que são as saídas do controlador.

A estimação dos estados do processo é implementada utilizando o conjunto de equações 2 adaptado de (Aguirre, 2004)

$$\begin{aligned} \hat{x}_{(k+1|k)} &= A\ \hat{x}_{(k|k)} + B\ u(k) \\ P_{(k+1|k)} &= A\ P_{(k|k)}\ A^T + B\ Q\ B^T \\ K &= P_{(k+1|k)}\ \frac{C^T}{\left[C\ P_{(k+1|k)}\ C^T + R\right]} \\ \varepsilon_{(k+1)} &= y_{(k+1)} - C\ \hat{x}_{(k+1|k)} - D\ u_{(k)} \\ \hat{x}_{(k+1|k+1)} &= \hat{x}_{(k+1|k)} + K\ \varepsilon_{(k+1)} \\ P_{(k+1|k+1)} &= P_{(k+1|k)} - K\ C\ P_{(k+1|k)} \end{aligned} \tag{2}$$

em que $\hat{x}_{(k+1|k)}$ são os estados estimados pelo FK por propagação, $\hat{x}_{(k+1|k+1)}$ são os estados corrigidos, ε é o resíduo ou erro de estimação, Q e R são as constantes do FK, P é a matriz de covariância e K é o ganho multiplicativo do resíduo (ganho de Kalman).

A calibração do Filtro deve ser feita ajustando as constantes Q e R utilizando dados confiáveis, ou seja, livres de ataques. Caso contrário toda a detecção de falhas estará comprometida. É importante ressaltar que, para a calibração do FK, busca-se obter um resíduo com características de sinal branco gaussiano.

Por último, o FK é um estimador de estados estocástico que apresenta erro quadrático mínimo e tem sua formulação matemática levando em consideração os ruídos aditivos e multiplicativos (Aguirre, 2004), portanto, espera-se que sua utilização obtenha bons resultados em testes com dados obtidos em um sistema não ideal.

Uma vez que o FK é ajustado até obter resíduo branco gaussiano e, por sua vez, o ataque *replay* consiste na concatenação de um conjunto de dados n vezes (enquanto dura o ataque ofensivo), a autocorrelação do conjunto de dados durante o ataque se torna alta.

Na proposta desse método, utiliza-se um registro dos dados de entrada e saída mais recentes do sistema. Esse registro deve ser grande o suficiente para possibilitar boa precisão nos cálculos e reduzir a incidência de falso negativos que é um problema maior que falso positivos para uma infraestrutura crítica. Por outro lado o tamanho do conjunto de dados é limitado pela quantidade de memória utilizada e pelo tempo para realizar os cálculos necessários da detecção.

Os cálculos realizados na detecção do ataque

são feitos em um processo iterativo. A cada iteração, os dados referentes às entradas e saídas do sistema são obtidos e fornecidas ao Filtro de Kalman. Por sua vez, o FK realiza a estimação dos estados e saídas e calcula do resíduo, que em seguida será armazenado no banco de dados, esse processo é chamado de atualização do registro. O banco de dados armazena os erros de estimação das N últimas iterações.

Uma vez que a Indústria 4.0 adota o armazenamento em nuvem e *cloud computing*, problemas com espaço em memória e gerenciamento de uma quantidade massiva de informações se tornam irrelevantes. Isso permite ter grande controle sobre o histórico dos processos e dos equipamentos industriais. Um exemplo dessa aplicação é abordado em (Lee et al., 2014), em que se identifica o estado dos equipamentos baseando-se em comparações com o registro de padrão de normalidade de todas variáveis relacionadas a cada equipamento. Portanto, adaptando essas tecnologias às necessidades do método apresentado, é possível construir um banco de dados suficientemente grande para conhecer todos os resíduos desde a implantação do sistema.

A partir da atualização do banco de dados é necessário calcular a correlação entre a amostra recém-armazenada e as antigas, deve ser determinado um limiar para indicar a detecção de ataque, esse valor é estabelecido empiricamente.

3 Aplicação do método

O método proposto foi aplicado à planta SMAR PD3-F em uma simulação de sistema inteligente interconectado e vulnerável a ciberataques. Foi utilizada uma malha responsável pelo controle do nível de água no tanque de aquecimento da planta (tanque esquerdo da Figura 1). A estimação do modelo do sistema foi feita considerando o mesmo como uma caixa preta e utilizando o método dos mínimos quadrados para obter um modelo ARX. Inicialmente colocou-se o sistema em malha aberta com uma entrada fixa $u_v[k] = 0.6$, sendo que esta entrada representa a porcentagem de abertura da válvula pneumática que controla a vazão para o tanque de aquecimento. Após o sistema entrar em estado estacionário, iniciou-se o envio de um sinal $u_v[k]$ pseudoaleatório de média 0.6 para o controle da válvula pneumática e foi feita a coleta dos dados referentes ao nível.

Na Figura 2 é mostrada a dinâmica do sistema para a entrada pseudoaleatória com características de sinal branco gaussiano. Os sinais de entrada e saída mostrados são parte do conjunto de dados utilizados para estimar os parâmetros do modelo utilizado nos testes de detecção.

Como foi dito na Seção 2, é de extrema importância verificar se o modelo estimado é capaz de representar o sistema. Portanto antes de avançar

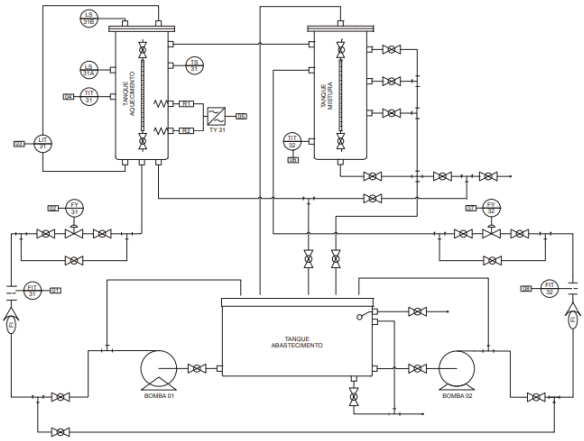


Figura 1: Diagrama P&I da planta SMAR PD3-F. Fonte: Manual de instruções SMAR

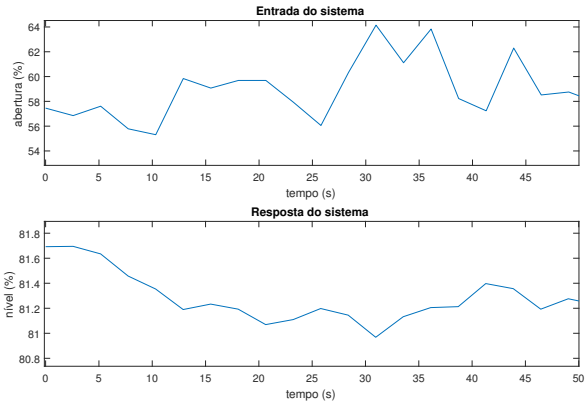


Figura 2: Representação gráfica da dinâmica do modelo mostrando parte das amostras de entrada e saída do sistema utilizadas na estimação. Fonte: O autor

nos testes, o modelo foi validado pelos métodos um passo a frente e Livre. Os resultados são mostrados na Figura 3, em que é feita a comparação entre os dados extraídos da planta e os simulados pelos diferentes métodos.

Uma vez que o modelo se mostra eficiente foi realizada a simulação com a replicação de dados coletados. Esse teste simula o ataque *replay* sendo orquestrado por meio do *hijacking* de dados, seguido da substituição dos mesmos por dados passados. Uma diferença entre as formas que ambos ataques foram simulados se dá no número de repetições do conjunto de dados salvo: a simulação mostrada na Figura 4a foi feita com quatro repetições do conjunto de dados capturado; a Figura 4b mostra a simulação de um ataque *replay* com duas repetições do conjunto de dados.

Os testes de detecção foram realizados com os dados mostrados na Figura 5, que mostram dados extraídos da planta utilizando dois controladores diferentes. Nestes dois casos específicos de testes, foram simulados ataques utilizando os dados mais recentes, ou seja, assim que o agente perpetrante

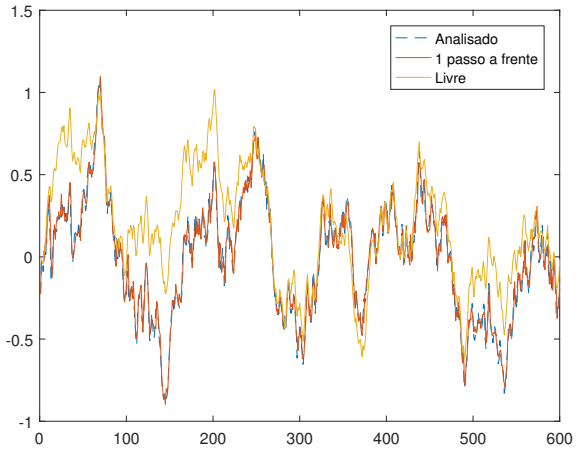
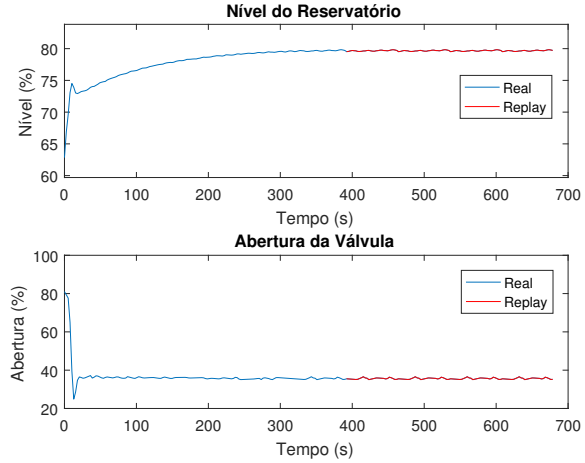
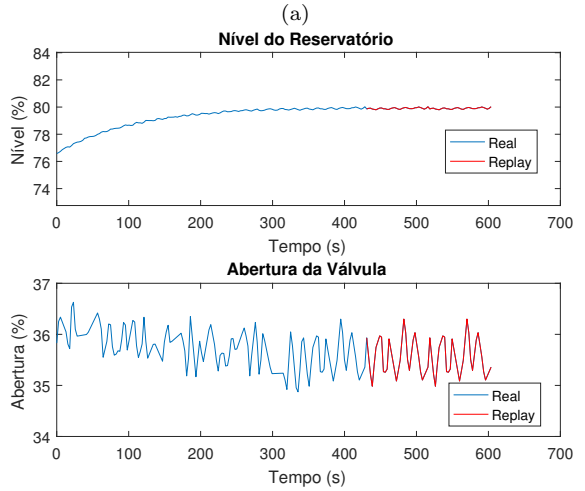


Figura 3: Validação do modelo estimado com predições um passo a frente e livre. Fonte: O autor

finalizou a captura e armazenamento dos dados, ele inicia a substituição dos novos dados.



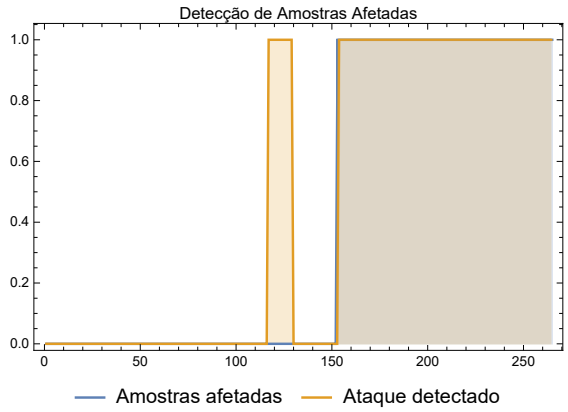
(a)



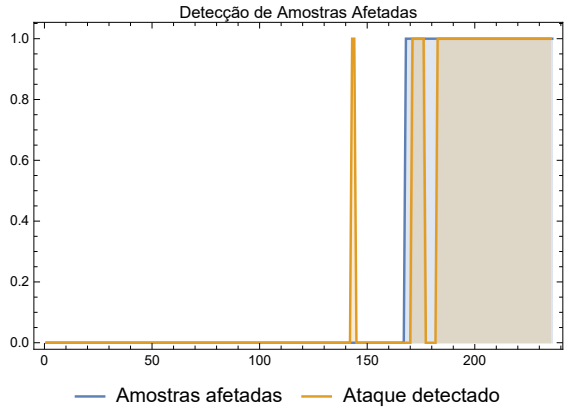
(b)

Figura 4: Simulação do ataque *replay* nos dados coletados de diferentes controladores: (a) *replay* de dados prolongado entre os instantes 312,18 e 392,16 s; (b) *replay* de dados entre os instantes 345,72 e 433,44 s. Fonte: o autor.

Por último, o algoritmo implementado foi aplicado aos dados coletados na planta sob ataque. Nas Figuras 5a e 5b, são mostradas as amostras que sofreram ataque. As curvas azuis indicam os dados que foram modificados pelo atacante, por sua vez, as curvas laranjas indicam as amostras que o algoritmo considera como modificada por ataque *replay*.



(a)



(b)

Figura 5: Resultados dos testes do ataque *replay* nos dados coletados de diferentes controladores. 1 Significa existência ou detecção de ataque, 0 significa ausência de ataque: (a) Comprometimento das amostras 152 em diante; (b) Comprometimento das amostras 168 em diante. Fonte: o autor.

Nas Figuras 5a e 5b, pode-se perceber um ponto interessante sobre o limiar de detecção que foi considerado na Seção 2. Ambos os resultados apresentam falso positivos, mas a princípio isso não é um problema, pois uma linha de produção não seria interrompida antes de utilizar outros testes mais avançados. Porém quando reduziu-se o limiar de detecção para diminuir o número de amostras falso positivas algumas amostras que haviam sido modificadas pelo atacante não foram reconhecidas e isso é preocupante pois à partir do momento que um grande número de amostras modificadas não são detectadas, o atacante se torna capaz de fazer as modificações desejadas no sistema sem levantar suspeitas. Portanto, caso o que se procura é reduzir o número de amostras falso

positivas no método, deve-se atentar bem ao número de amostras modificadas que passam como intactas. Para esse sistema, devido a sua natureza, os resultados oferecidos são satisfatórios. Em sistemas cuja resposta seja mais rápida, alguns segundos que o operador se baseia em dados falso negativos podem resultar em um incidente.

4 Conclusões

O método proposto apresenta uma forma simples e rápida de detectar ataques *replay* a um sistema ciberfísico. O grande diferencial desse método se dá nele conseguir identificar o tipo de ataque no instante da detecção devido a sua abordagem, além disso, ele foca em um ataque de enganação comum e que normalmente é utilizado juntamente com outros ataques mais agressivos ao sistema. Portanto, um algoritmo que implemente esse método pode fazer uso de outros métodos mais específicos que serão usados após uma primeira análise. Assim, quando o ataque *replay* é detectado no sistema, é importante verificar a presença de outros ataques e ponderar os riscos gerados à infraestrutura e ao processo.

Referências

Aguirre, L. A. (2004). *Introdução à identificação de sistemas–Técnicas lineares e não-lineares aplicadas a sistemas reais*, Editora UFMG.

Billings, S., Chen, S. and Korenberg, M. (1989). Identification of mimo non-linear systems using a forward-regression orthogonal estimator, *International journal of control* **49**(6): 2157–2189.

Forti, N., Battistelli, G., Chisci, L. and Sinopoli, B. (2016). A bayesian approach to joint attack detection and resilient state estimation, *Decision and Control (CDC), 2016 IEEE 55th Conference on*, IEEE, pp. 1192–1198.

Lee, J., Kao, H.-A. and Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment, *Procedia Cirp* **16**: 3–8.

Miller, B. and Rowe, D. (2012). A survey scada of and critical infrastructure incidents, *Proceedings of the 1st Annual Conference on Research in Information Technology*, RIIT '12, ACM, New York, NY, USA, pp. 51–56.

Mo, Y., Chabukswar, R. and Sinopoli, B. (2014). Detecting integrity attacks on scada systems, *IEEE Transactions on Control Systems Technology* **22**(4): 1396–1407.

Nishiya, K., Hasegawa, J. and Koike, T. (1982). Dynamic state estimation including anomaly detection and identification for power systems, *IEE Proceedings C (Generation, Transmission and Distribution)*, Vol. 129, IET, pp. 192–198.

Pasqualetti, F., Dörfler, F. and Bullo, F. (2012). Attack detection and identification in cyber-physical systems–part ii: Centralized and distributed monitor design, *arXiv preprint arXiv:1202.6049* .

Pasqualetti, F., Dörfler, F. and Bullo, F. (2013). Attack detection and identification in cyber-physical systems, *IEEE Transactions on Automatic Control* **58**(11): 2715–2729.

Zhu, B., Joseph, A. and Sastry, S. (2011). A taxonomy of cyber attacks on scada systems, *Internet of Things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing*, pp. 380–388.

Zhu, M. and Martinez, S. (2011). Stackelberg-game analysis of correlated attacks in cyber-physical systems, *American Control Conference (ACC), 2011*, IEEE, pp. 4063–4068.