Residential Demand Response Behaviour Modeling applied to Cyber-physical Intrusion Detection

Heussen, Kai; Tyge, Emil; Kosek, Anna Magdalena

Published in:
Proceedings of 12th IEEE Power and Energy Society PowerTech Conference

Link to article, DOI:
10.1109/PTC.2017.7981209

Publication date:
2017

Document Version
Peer reviewed version

Residential Demand Response Behaviour Modeling applied to Cyber-physical Intrusion Detection

Kai Heussen, Emil Tyge and Anna Magdalena Kosek
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark
Email: kh@elektro.dtu.dk

***Abstract*—A real-time demand response system can be viewed as a cyber-physical system, with physical systems dependent on cyber infrastructure for coordination and control, which may be vulnerable to cyber-attacks. The time domain dynamic behaviour of individual residential demand responses is governed by a mix of physical system parameters, exogenous influences, user behaviour and preferences, which can be characterized by unstructured models such as a time-varying finite impulse response. In this study, which is based on field data, it is shown how this characteristic response behaviours can be identified and how the characterization can be updated continuously. Finally, we propose an approach to apply this behaviour characterization to the identification of anomalous and potentially malicious behaviour modifications as part of a cyber-physical intrusion detection mechanism.**

***Index Terms*—Demand response, Cyber-physical systems, Intrusion detection, Data-driven**

I. Introduction

Residential demand response is considered a significant resource of localized flexibility, in particular in cases where the heat and cooling needs of buildings are satisfied by electric heating or heat pumps. As demand response is maturing from a vision to real-world applications, it is also becoming a potential target for cyber attacks. A real-time demand response system can be viewed as a cyber-physical system: a physical structure with a behaviour that is strongly influenced by ICT facilitated interactions. Demand response behaviour is partly governed by physical properties of the process, partly by autonomous behaviour of residents, and in part by the local control systems, which may be parametrized by local users. This combination of uncertain and in-transparent system properties leads to new challenges for reliability and security of operation. Further, the involved control systems are more diverse and open, which offers more entry points for cyber-attacks. We investigate the feasibility of an online monitoring system characterizing the dynamic response behaviour of price-controlled demand. The goal is to formulate indicators of anomalous behaviour based on the observable characteristics of individual households. The investigations are based on a data set obtained by the EcoGrid.eu project [1]. The indicators are framed by a method for cyber-physical intrusion detection system (CPS-IDS) developed in context of the SALVAGE project[1].

[1]http://www.salvage-project.com/

As demand response requires a large number of typically quite diverse individual units, one cannot cannot expect direct and manual monitoring. Further, occupant privacy should be considered in such systems, thus give preference to the applications of aggregate, purpose-build detection models, and to avoid the use of individually traceable information. Based on experimentally observed data, we therefore aim to develop a modeling approach to detect specific kinds of "anomalies" in observable response dynamics.

Whether a cyber-intervention is the actual cause of anomalous behaviour cannot be inferred from physical models alone: CPS-IDS hypotheses in the Salvage framework combine several sources of evidence, including both cyber- and physical anomaly and intrusion detection [2]. In the Salvage CPS-IDS, the developed indicators are viewed as a DER analysis component as illustrated in Fig. 1. As such anomalies may not only be caused by external interventions, but simply be a reflection of changes in the inhabitants' behaviour. We therefore propose a hypothesis-driven approach, which a) will account for more apparently goal-directed changes, and b) neglect commonly observed patterns of behavioural change.

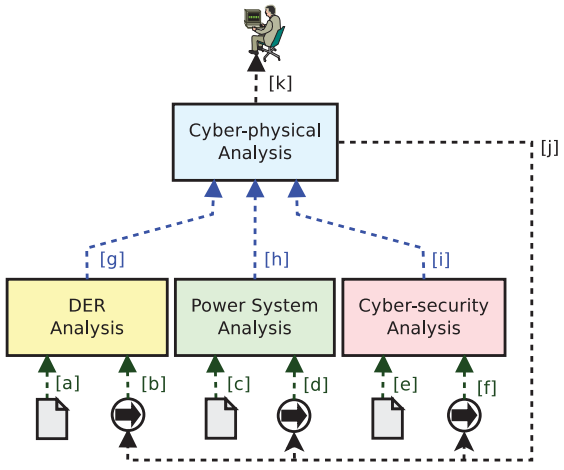

Fig. 1. Cyber-physical IDS architecture proposed in SALVAGE project [3].

In Section II the Cyber-physical IDS concept and the here proposed approach to behavioural anomaly detection in demand response is introduced. Section III reviews the

applied methods. A proof-of-concept application to field data is reported in Section IV which supports exemplifies the proposed approach to demand response modeling. The method and results are discussed in Section V.

II. Concept and Approach for IDS

A. CPS-IDS Concept

Intrusion detection systems (IDS) gather and analyze the information from a computer network or a system in order to discover malicious activities or violations of policy [4]. An IDS for application in smart grids needs to integrate both on-line and post-mortem analyses, of both the observed cyber- and physical systems: A Cyber-Physical Intrusion Detection System (CPS-IDS) [2].

The cyber-physical IDS architecture proposed in the SALVAGE project consists of two main parts: a domain-specific analysis of the behavior of the individual components, and an integrated analysis of the cyber-physical system 1. The joint cyber-physical analysis combines the information from both physical and cyber-security components and presents the outcomes to the power system operator.

The initial SALVAGE CPS-IDS design had a unidirectional structure detect anomalies in operation of both cyber and physical components [3], [5], [6]. The extended hypothesis-driven framework, presented in [2], allows the integrated analysis to configure and combine the available component analyses (arrow [j] in Fig. 1). The extended architecture supports the idea of a hypothesis testing approach to cyber-physical security. Hypotheses in the SALVAGE CPS-IDS framework are derived from hypothesis templates composed by domain experts, each addressing a particular cyber-physical power system attack goal. The hypothesis template is then matched with the present system configuration such that any applicable hypothesis can be generated and quantified.

The different parts of the entire CPS-IDS as depicted in Figure 1 perform each a different type of analysis. The Cyber-physical analysis combines performs the cyber-physical hypothesis generation [j], quantification and ranking, assessing the *ad hoc* risk of any of the active hypotheses [k]. Inputs to this hypothesis quantification [g,h,i], are labeled state- and probability data generated in domain-specific analysis modules. The module for power system analysis performs e.g. load flow based calculations, requiring network models as well as electrical measurements. The module for cyber-security analysis performs a model-based probabilistic simulation of cyber attacks within the IT infrastructure that exists alongside the physical infrastructure of the power grid and its cyber-neighbourhood. The DER analysis assesses the current state and likelihood of anomalous behaviour of different types of distributed energy resources (DER), such as photovoltaic (PV) electricity generation [3], [5] or demand response. The main purpose of this paper is to discuss a particular variant of the DER analysis module suitable for assessing price-based demand response systems.

The objective of this work has been to provide online indicators for anomalous behaviour of individual households

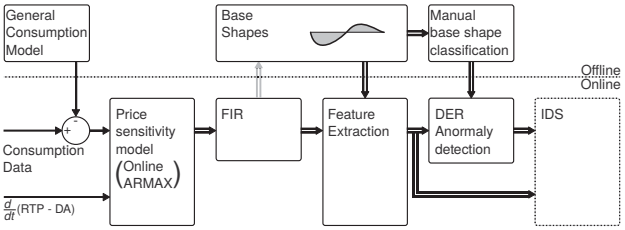

Fig. 2. Conceptual outline of detection approach.

or groups of households in a demand response portfolio. Given a certain attack hypothesis, what types of behaviour can be expected to be observed? Such an attack hypothesis allows to formulate a systematic character in the behaviour change. The here investigated systematic behaviour changes include: *(a)* Elimination of response amplitude (de-activate response); *(b)* altering response shape to worsen "Kick-back" (to cause grid overload or Cobweb effect [7]); *(c)* Inverse-response (destabilizing system).

B. Approach for DER Behaviour IDS

In previous work, we have investigated contextual model-based anomaly detection for DER analysis applied to photo-voltaic (PV) DER [3], [5], [8]. There is however a fundamental difference in the dynamical characteristics of (residential) demand response and the PV models: instead of an algebraic input-output relation, the residential demand response is characterized by *a)* unobservable (random) behaviour of residents influencing both demand volume as well as parameters of the response characteristic, and *b)* the unknown thermal dynamics of the household, influenced by further exogenous parameters.

In previous work [9], [10] on characterizing the price-change responses from houses with smart metering equipment the suggested approach has been to extract finite impulse response (FIR) coefficients from the price input to the demand. The following steps are proposed:

1. General power consumption modelling. Models the part of the system that is not sensitive to the real time prices.
2. Online price sensitivity modelling. Models the price sensitivity, to reduce features of the price sensitivity signal for anomaly detection.
3. Anomaly detection and feature extraction based on the system online estimated parameters from step 2.

Based on these principles, the data processing concept shown in Figure 2 has been developed.

The model comprises an offline phase where a general system behaviour and dynamical response characteristics are modelled, and an online phase where system parameters are continuously identified and anomaly detection parameters are computed as input to an IDS.

III. Methods and Dataset

The methods employed in this work belong to a very basic toolset of statistical methods: linear regression, ARMAX time series modeling and two clustering methods. The data set is

one recorded in a large scale field demonstration of demand response.

1) Linear Regression: Linear regression is a well known approach for modelling a linear relationship between a set of input variables and an output. It has the nice property that it will always converge to the global minimum, if there are as many or more samples than unknowns. The general form is shown in (1), where u_t, ϵ_t and y_t is the input, error and output, and θ_t is the estimated proportionality constants [11].

$$y_t + \epsilon_t = x_t \theta_x \tag{1}$$

Using this method, Larsen et al [12] and [10], shows that it is possible to extract FIRs to the real time market from another period of the Ecogrid EU data. With prices, weather and a set of Fourier terms as input to model the general behaviour and a difference model modelling the response from changes in the pricing. This provides a very condensed set of information at DER level, describing the system response to market price changes.

2) ARMAX system identification: ARMAX is a model for system identification. It stands for Auto-Regressive Moving-Average with eXogenous inputs. It estimates a linear model which on a transfer function form can be written as in (2). As an extension of the linear regression, it represents dynamics of the system by using delayed system outputs (auto-regression) and estimation of noise.

$$y_t = \frac{B(q)}{A(q)} u_t + \frac{C(q)}{A(q)} e_t \tag{2}$$

$A(q)$ is the system polynomial, $B(q)$ and $C(q)$ input polynomials of input and noise, where u_t, e_t and y_t, are the input, noise and output, respectively [13]. It is possible to estimate these parameters of the polynomials in a recursive manner, making an online system identification, that can follow a changing system.

3) k-Means algorithm: As described in [11], the k-Means algorithm is a commonly used clustering algorithm for unclassified data. The algorithm associates data with a predefined number of clusters, iteratively minimizing the total distance from all the samples to the cluster centroids. This is achieved by alternating between associating the data samples with a cluster of the closest centroid and updating the cluster centroid as the mean point of the cluster.

This algorithm is fast and does not require prior knowledge of the data. Key to successful clustering in higher-dimensional sample spaces is the choice of a distance measure to calculate the distance between each centroid and sample.

4) Cosine distance: The Cosine distance is measure applicable for the k-Means algorithm. It is not a true metric, but has proven useful as a measure of data with a high dimensionality where the direction of a sample vector is as important as the sample itself [14]. The cosine distance builds upon the dot product of vectors, thus it captures angular distance between the sample and centroid in the k-Means algorithm. The definition is shown in (3). Where $\mathbf{A}$ and $\mathbf{B}$ are the feature vectors.

$$D_c(\mathbf{A}, \mathbf{B}) \equiv 1 - cos(\theta) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||} \tag{3}$$

5) Gaussian mixture models: Gaussian mixture models take a more statistical approach to the clustering problem. The idea is to approximate n statistic sub-populations in a dataset. This is done by maximum a posteriori estimates. Like k-Means clustering, this is often done in iterative steps of ascending the likelihood function and updating the population estimates until convergence [11].

6) Dataset: The Ecogrid EU project was a research and demonstration project for a future smart grid across the European Union. It took place on the Danish island Bornholm during the years 2011 to 2015. Characteristic for Bornholm is a high penetration of renewable energy sources and the grid is almost separated from the mainland, with a single power line to Sweden. A corner stone of the Ecogrid EU, was therefore a real-time market for activating demand response from small scale DER [1]. The market generated 5-min real-time imbalance prices to which household level controllers would respond, adjusting electricity consumption of electric space heating. The houses involved in Ecogrid EU, had a smart-meter monitor their power consumption in five minute intervals. For this work about 4 months of consumption data have been assessed: in total 1736 houses and 40033 datapoints, along with local weather data. There are, to our knowledge, no actual attacks on the power grid, local controllers or real-time market in the data used for this project.

IV. Results

Here, the DER modeling and monitoring approach is outlined and exemplified on field data.

A. Stationary responsiveness

In the offline characterization we follow the methods and approach presented in [9], [10], applying linear regression. The results for the baseline consumption models are very similar to the original work and will not be further discussed here. The goal of the price sensitivity model is to extract the demand response FIR from the power consumption data on DER level. Following [10], the price-responsiveness of household power consumption was modeled in form of a finite impulse response (FIR) to the real-time price variations. The input, u_t for these models is the derivative of the difference between the real time price (RTP_t) and the day ahead price (DA_t), the output y_t is the consumption data of the household subtracted the prediction from a baseline model x_t. The baseline model for the data set has been created following [10] and is not further discussed here.

$$u_t = \frac{d}{dt}(RTP_t - DA_t) \tag{4}$$

$$y_t = c_t - x_t \tag{5}$$

The price-changes u_t are then time-lagged as input $\mathbf{u}_t = [u_t \dots u_{t-T_L}]^T$ computed from (4) where T_L is the time lag here chosen to be 200 minutes, or 40 samples; as output the

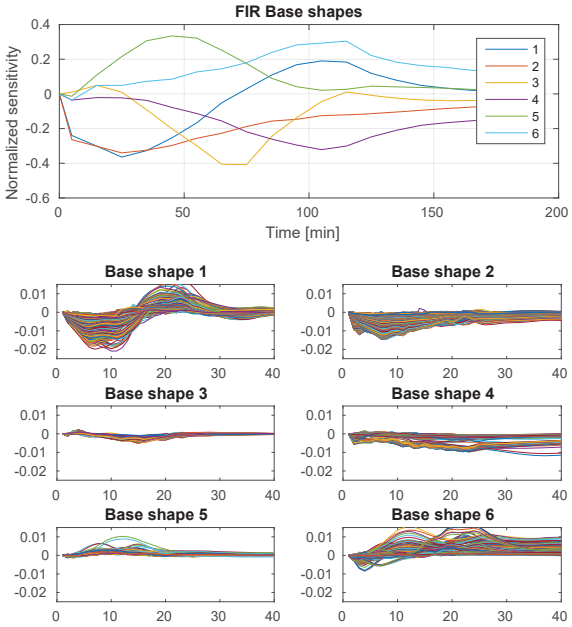

Fig. 3. *Top*: Normalized FIR coefficients for each of the 6 cluster centroids (k=6). *Bottom*: Sample FIRs from each cluster (sample selection based on $\text{conf}_{controlled} > 0.5$ and shape similarity $\delta_{shape,i} > 0.5$). Note the x-axis corresponds to the coefficient index (40*5min = 200min) and y-axis is the absolute sensitivity.

(5) of each of the houses to a linear regression, the linear model (6) appears. From θ_{FIR} the FIRs of each house can be extracted one at a time, note that .

$$y_t + \epsilon_t = \mathbf{u}_t^T \theta_{FIR} \tag{6}$$

The response characterization by θ_{FIR} is a stationary model for an individual household.

B. Dynamic responsiveness characterization

An ARMAX model was implemented to emulate an online estimation of the dynanmic responsiveness. The order of the ARMAX model is chosen by trial and error on test data to be 16 for the system ($A(q)$), spanning 80 minutes; 24 for the input polynomial ($B(q)$), spanning 2 hours; 2 for the error dynamics, spanning 10 minutes. The characteristic time of the recursive parameter estimation algorithm is set to 14 days. This choice balances noise with the ability to follow changing system properties. The output of the ARMAX model is a set of parameters for the three polynomials at each point in time. From each parameter set, a FIR can be approximated as the superposition of a FIR response with the input through the parameters in $B(q)$ and an infinite impulse response (IIR) of the system through $A(q)$.

Figure 4 illustrates the dynamic FIR as computed from the online ARMAX estimation for a single household. The upper graph plots the FIR coefficients by lag time, as presented in Fig. 3, and allows visual comparison with the stationary response shape of the linear regression model. The lower

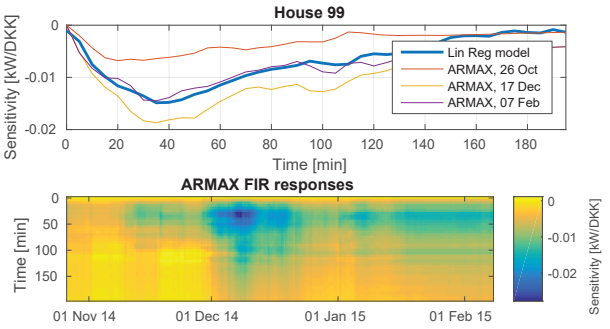

Fig. 4. Illustration of ARMAX-based online FIR estimation. *Top*: comparison of stationary response with 3 ARMAX sample responses; *bottom*: variation of ARMAX FIR over observation time (vertical axis: delay-time corresponding to x-axis in top-plot). The response amplitude is intensified, but the characteristic shape remains.

timeseries illustration provides a colourmap of the evolution of the ARMAX-estimated FIR parameters (vertical) along the time period of the modeling (horizontal).

This "online" ARMAX was only applied to a subset of 150 houses of the original data set. These were chosen by selecting: a) the 50 least and b) 50 most *responsive* houses, along with c) 50 random draws among the remaining houses. The responsiveness was measured by three amplitude parameters: the integral, the maximum and the minimum value of the FIR.

C. Clustering of FIR

The set of stationary responses has been clustered by application of k-Means using a cosine-distance measure applied to normalized and down-sampled FIR parameters. The respective cluster centroids (normalised) are plotted in presented in Fig. 3 (top). The bottom plots in the same figure illustrate the (unscaled) FIR responses associated with the respective cluster. Each line represents one household FIR.

D. Characterization of Anomaly and Intent

The above outlined models are aimed to serve the characterization and identification anomalous or even malicious behaviour in the demand response systems. To mitigate the high uncertainties and variance in the response behaviour, two types of metrics are applied: i) a measure of the response amplitude or volume $\log(||\delta_A||)$; ii) a similarity assessment, comparing the observed behaviour shape to benevolent or malicious behaviours. Using Gaussian Mixture (GM) models, the statistics of these metrics are then modelled for both 'normal' and 'undesired' behaviour.

1) Responsiveness measure: A histogram of the log magnitude of the three amplitude features for these 150 houses, reveals three proto-distributions: the least controlled (uncontrolled) are separated from the most controlled group, with the random selection in the middle as expected (Figure 5). As first measure, a filter was estimated fitting a Gaussian mixture model to the data set containing only the 100 most/least

responsive houses. This filter yields the quantity $\text{conf}_{controlled}$, plotted as the black line in Fig. 7.

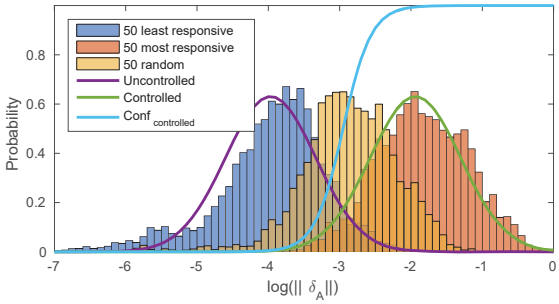

Fig. 5. Histogram of $\log(||\delta_A||)$ with with characterization of the price-responsiveness amplitude. Overlayed is the Gaussian mixture (GM) model which generates the $\text{conf}_{controlled}$ parameter.

2) Similarity to base shapes: Using the cosine-distance of the normalized response shapes, a measure similar to the $\text{conf}_{controlled}$ can be derived based on the summation of distance features associated with the respective shapes. Two approaches have been formulated to assess similarity to base shapes.

The first approach is based on the measure feature $\delta_{shape,i}$, defined as the inverse cosine distance to the base shape, normalized with the sum of the distance to all the shapes. The inverse relationship on the distance is penalizing long distances, while the normalization ensures that the features can be compared. This approach has been applied in the cluster allocation in Fig. 3. By manually grouping the base shapes into a *desirable* and *undesirable*, this measure has been applied to compute the affinity with that response type: *desirable* (green: BS-1 to BS-4) and *undesirable* (red: BS-5 and BS-6), as illustrated in Fig. 7, upper plots. For the stable case ("House 99"), here only little change in the response type is observed, even though the response amplitude is changing over time (Fig. 4). For the house undergoing an intervention ("Augmented house 23"), the measure becomes stable for higher amplitude but is sensitive at low FIR amplitudes.

The second approach uses a Gaussian Mixture Model to characterize a confidence for cluster allocation, similar to the responsiveness measure introduced above. To identify the 2D GMM with two centers, the cluster centroids were again associated either with *desirable* or *undesirable* base shapes; two features were then computed based on the sum of log-distances to either undesirable, x_1, or desirable, x_2, base shapes. The identified GMM produces a more informed characterization of the classification of response samples by offering a confidence-level of the classification. Note that in preparation of the GMM base shapes, the k-Means clustering has been applied to a subset of the total data set, including only the houses with 20% highest responsiveness range to extract only significant response contributors, yielding a different set of cluster centroids (base shapes) than utilized in application of the first approach. The base shape numbers here are therefore also different.

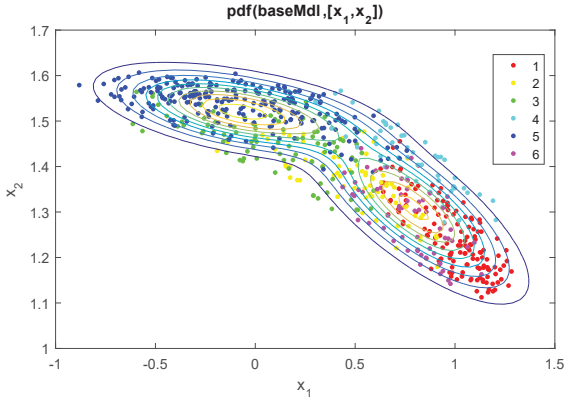

Fig. 6. Scatterplot of sum of log-cos-distance to undesirable shapes x_1 vs. sum of log-cos-distance to desirable shapes x_2; the data points are coloured by associated base shape clusters (*desirable BS*: {1, 2, 4, 6} ; *undesirable BS*: {3, 5}). The data is overlayed with the two-parameter GM to be used for response type identification: upper-left: undesirable response shape, lower-right: desirable response shape.

In the chosen approach to clustering, the engineering choices in sample selection and the manual base shape classification directly influence the detection outcomes. To evaluate the final detection effectiveness, reference cases would be required, characterizing the systematic behaviour changes (a)-(c) listed in Section II-A. With such reference cases, or analytic metrics, an automatic base shape classification could be developed.

3) Anomaly Detection: The implemented anomaly detection evaluates changes in the feature vector δ_{FIR}. Every point in time is assigned a probability of it being an anomaly. Assuming difference of the features is normal distributed, their standard deviations is estimated based on all 150 houses. These parameters are then used to calculate an anomaly probability at each timestamp per house.

In Figure 7, this feature is marked by a diamond, as can be seen in the lower plot set. Here, it is apparent that this anomaly detection is overly sensitive for low response amplitudes (observed false positives before Dec.1s). Along the same lines, the similarity measure (separation line between red/green areas) reacts very sensitive in the low responsiveness period before Dec. 1st ($\text{conf}_{controlled} < .5$). On the contrary, it is rather stable in combination with high responsiveness, as can be observed in both examples in Figure 7. This suggests a combination of the measures, e.g. by reducing the confidence in the similarity measure in dependence of another measure such as $\text{conf}_{controlled}$.

Here it is worth noting, that in case of a CPS-IDS integration, the anomaly detection component will not be applied independently. Here the statistical measures outlined previously are better applied, as they deliver a continuous probability value, to be employed by the CPS-IDS hypothesis quantification component [2].

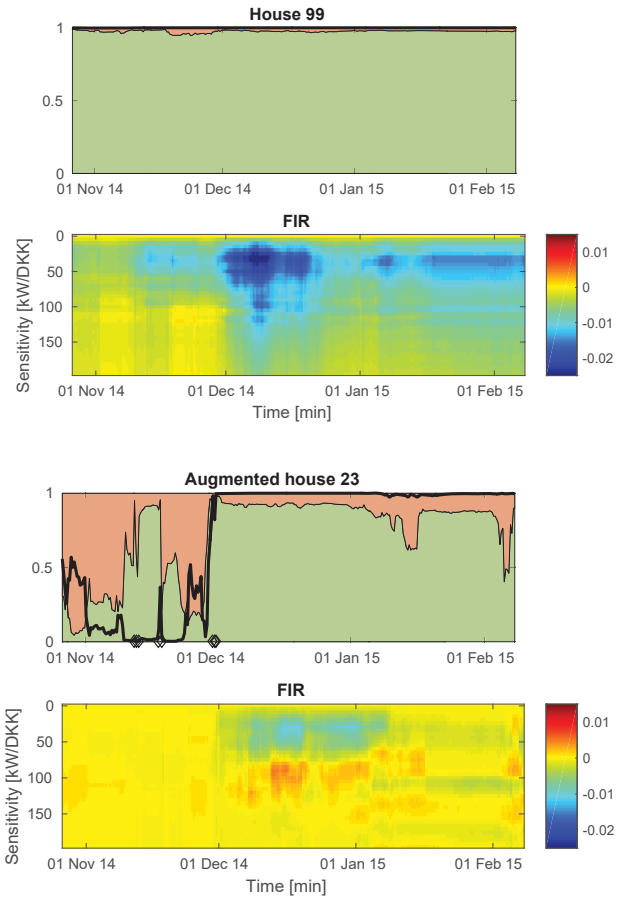

Fig. 7. The ARMAX and anomaly detection applied to two houses: "House 99" from Fig. 4 and "Augmented house 23" with artificially mixed response data: shifting from low price-responsivenes to high price-responsiveness on Dec.1st.

V. Discussion and Conclusion

Based on smart metering data, the behaviour of price-responsive control of loads can be monitored, and these observations may be integrated in a cyber-physical intrusion detection system (CPS-IDS).

The load response behaviour was characterized by the Finite Impulse Response (FIR) the behaviour. The wide variety of response shape indicates that relevant 'anomalies' are not easily identified in the time domain behaviour of a group of loads, but engineering intuition was applied to classify the shapes of time domain behaviours observable in the data. This expert-based approach was employed to intuitively classify response types into "desirable" and "undesirable" features. Statistical methods were then applied to detect and classify behaviour anomalies. Behaviour change has been formulated as criterion for anomalies using to independent features: response amplitude ($\mathrm{conf}_{controlled}$) and a similarity measure. Both measures are formulated as a probability metric using statistically identified distributions, so that the observed probabilities can be employed in the further probabilistic reasoning step in the CPS-IDS for risk analysis.

The results demonstrate a feasibility of a statistical approach to integrating cyber-physical observations in demand response oriented intrusion detection system. Parameter identification using the chosen ARMAX technique takes about 24h of observation until convergence, which puts limitations on the integration in online CPS-IDS systems, but is in line with the time-scale of typical smart-metering data acquisition. The validity and accuracy of the developed statistical models has to be evaluated in future studies. Approaches to avoid the manual classification of response types should also be replaced by more principled metrics based from attack goals.

The reported monitoring for normal and anomalous demand response behaviour offers a number of possible applications beyond the CPS-IDS application outlined here, such as: system supervision and decision support, monitoring of an aggregator's portfolio to estimate flexibility or monitor user behaviour, or validation of a contracted response.

Acknowledgment

This work was supported in part by the SALVAGE project, under ERA-Net SmartGrids through the Danish ForskEL programme (Grant No. 12254), and under EU FP7 grant no. ENER/FP7/268199/EcoGrid EU.

References

[1] N. P. Lund, P., R. D. Grandal, S. H. Sørensen, M. F. Bendtsen, G. L. Ray, E. M. Larsen, J. Mastop, F. Judex, F. Leingruber, K. J. Kok, and P. A. MacDougall, "EcoGrid EU - A Prototype for European Smart Grids, Overall evaluation and conclusion," 2015.

[2] O. Gehrke, K. Heussen, and M. Korman, "Integrated multi-domain risk assessment using automated hypothesis testing," in *Workshop on Cyber-physical security and reliability in smart grids (CPSR-SG)*. ACM, 2017.

[3] A. M. Kosek and O. Gehrke, "Ensemble regression model-based anomaly detection for cyber-physical intrusion detection in smart grids," in *2016 IEEE Electrical Power and Energy Conference*. IEEE, 2016.

[4] P. E. Proctor, *Practical intrusion detection handbook*. Prentice Hall PTR, 2000.

[5] A. M. Kosek, "Contextual anomaly detection for cyber-physical security in smart grids based on an artificial neural network model," in *2016 Joint Workshop on Cyber-physical Security and Resilience in Smart Grids*. IEEE, 2016.

[6] M. A. Kosek and K. Heussen, "SALVAGE D 2.1 Description of existing and extended smart grid component models for use in the intrusion detection system," DTU, Tech. Rep., 2015.

[7] E. Larsen, P. Pinson, J. Wang, and Y. Ding, *The Cobweb Effect in Balancing Markets with Demand Response*. IEEE, 2015.

[8] M. A. Kosek, M. Korman, K. Heussen, E. Tyge, and A. H. Jonasdottir, "SALVAGE D 2.2 Description of the developed algorithms for intrusion detection in smart grid components," DTU, Tech. Rep., 2016.

[9] E. Larsen, P. Pinson, G. le Ray, and G. Giannopoulos, "Demonstration of market-based real-time electricity pricing on a congested feeder," in *Proceedings of European Electricity Market Conference*. IEEE, 2015, pp. 886–890.

[10] E. Larsen, "Demand response in a market environment," Ph.D. dissertation, 2016.

[11] S. Marsland, *Machine Learning: an algorithmic perspective*. CRC Press, 2009.

[12] E. M. Larsen, P. Pinson, S. Member, F. Leimgruber, and F. Judex, "From Demand Response Evaluation to Forecasting - Methods and Results from the EcoGrid EU Experiment," pp. 1–9.

[13] Poulsen, Niels Kjølstad, *Stokastisk adaptiv regulering*. Polyteknisk forlag, 1997.

[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education, 2014.