Programme Area: Smart Systems and Heat

Project: WP1 Appliance Disaggregation

Title: Incorporation of Appliance and layout information

Abstract:

The deliverable addresses the problem of prediction of hot water usage and gas using online regression models. While there was no access to the quantity of hot water consumed over time by the inhabitant, the inference can be made using the temperature of the domestic hot water flow and the central heating flow. The algorithmic steps are provided in this report, and a full description of the algorithms is given in the referred papers.

Context:

The High Frequency Appliance Disaggregation Analysis (HFADA) project builds upon work undertaken in the Smart Systems and Heat (SSH) programme delivered by the Energy Systems Catapult for the ETI, to refine intelligence and gain detailed smart home energy data. The project analysed in depth data from five homes that trialed the SSH programme's Home Energy Management System (HEMS) to identify which appliances are present within a building and when they are in operation. The main goal of the HFADA project was to detect human behaviour patterns in order to forecast the home energy needs of people in the future. In particular the project delivered a detailed set of data mining algorithms to help identify patterns of building occupancy and energy use within domestic homes from water, gas and electricity data.

Project: HFADA
HIGH FREQUENCY APPLIANCE DISAGGREGATION ANALYSIS

Incorporation of Appliance and layout information

Contents

1. History

Date	Issue	Details of Change
	Version 1.0	Initial Version. Authors: Waqas Jamil Hamid Bouchachia

2. Documents Referenced

Ref	Document	Title
1	Word document that describes the HEMS data.	Data collection and data format – ELECTRIC, WATER and HEMS-V1 MONITORING
2	Word document that describes the HEMS V1 Mongo data base structure.	HEMS V1 Mongo Data Base Structure
3	Deliverable 1	HFADA_Deliverable_Ver2
4	Deliverable 2	HFADA_Deliverable2_V0
5	Deliverable 3	HFADA_Deliverable3_V0
5	Paper	Competitive Normalised Least Squares Regression
6	Paper	Competitive Online Regularised Regression

3. Glossary of Terms

Ref	Description
OSLOG	Online Shrinkage via Limit of Gibbs sampling
ORR	Online Ridge Regression
ONLMS	Online Normalised Least Mean Squares Regression

4. Executive Summary

The document addresses the problem of prediction of hot water usage and gas using online regression models. While there is no access to the quantity of hot water consumed over time by the inhabitant, the inference can be made using the temperature of the domestic hot water flow (referred to as Temperature 1) and the central heating flow (referred to as Temperature 2). The proposed new prediction algorithms operate online in line with the requirements of deliverable 5. The algorithmic steps are provided in this report, but a full description of the algorithms is given in the referred papers.

5. Introduction

As mentioned in the previous reports, the size of the utility usage data is very huge. Algorithms processing such large data must observe time and memory restrictions [1, 4]. Algorithms proposed so far have relied on the assumption that the process generating data is stochastic. On the contrary, in this report we propose new algorithms that make no assumption on the data generating process. Furthermore, these algorithms can infer and predict at the same time given the very nature of online learning. Our focus is on predicting gas and hot water temperature.

Game theoretic probability models [8] in online learning theory have grown into a backbone in machine learning. In practice, online learning is relevant to various applications such as text analysis, computer vision, time series analysis, network modelling among others. The main challenge in online models is to have a guarantee on the performance that is not much worse than the best learning strategy in hindsight. Here theoretical guarantees, expressed in terms of bounds, indicate the worst-case performance under well-specified assumptions. The statistical models often make many passes over the data and converge to a solution that minimises the loss. Thus, it is important to have nice convergence properties for a particular algorithm. The proposed algorithms go a step beyond and propose an upper bound on the performance while making only one pass over the data. Hence, having a guarantee is a much stronger and desirable property than convergence.

From the application perspective, the report looks at gas consumption and hot water temperature. The data offers the access to the quantity of gas consumed by the inhabitant, so predictive models can easily be fitted to the data at hand. Unfortunately, the access to hot water usage is not possible. The current report investigates the prediction of hot water temperature for which the data exists. Although we do not do any occupancy prediction, the idea of using hot temperature is to link it with the occupancy using some sort of thresholding.

To develop the predictive analysis (online regression) models, we relied on the attributes/features displayed in Table 1 below, which summarises water flow, gas flow and various sensor measurements such as temperature of rooms, temperature radiators and temperature of water, relative humidity of room, the state of the radiators' valve and the state of the boiler firing switch.

Table 1: Data attributes

timestampNTP	Water	Electricity			HEMS						
		Real power	Reactive power	RMS Spectrum power over different frequency ranges	Gas meter	Temperature			Humidity	Radiator _valve	Firing boiler
						Rooms	Radiators	Water			
. . . .	. . . .	. . . .	. . . .	. .	. .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .

While there is no access to the quantity of hot water consumed over time by the inhabitant, the inference can be made using the temperature of the domestic water flow (referred to as Temperature 1) and the central heating flow (referred to Temperature 2).

The organisation of this report is as follows. In the next section (Section 6) we present the pseudo-code of three online regression algorithms that we used to make prediction of gas and hot water temperature. Section 7 discusses the application of these algorithms in the context of HFADA. Section 8 concludes the report.

6. Gas Consumption and Hot Water Temperature Prediction

In the following, we briefly describe three novel algorithms used to predict gas consumption and hot water temperature. We approach the prediction problems stated as regression problems. All algorithms presented here are regression ones and operate online. These algorithms resemble the filtering models in signal processing. However, the goal is quite different. The goal in filtering is to filter the noise, while in online learning theory we are interested in solving the problem of prediction. An online regression algorithm receives an input at each step, predict the outcome and upon presentation of the actual outcome, the parameters (inference) of the model are adjusted. Notice, after processing each data we do not use it again. So, apart from the model we hold nothing in the memory, which makes online regression much more scalable than conventional statistical regression models.

6.1 Online Shrinkage via Limit of Gibbs sampling (OSLOG)

The pseudo-code of the algorithm is as follows:

Initialise: $M = 0^{n\times n}, b = 0^{n\times 1}, w = 1^{n\times 1}, a > 0$

FOR $\boldsymbol{t = 1, 2, \ldots}$

  Read $x_t \in \mathbb{R}^n$

  output $\gamma_t = w' x_t$

  $D_w = diag(w^1, \ldots, w^n)$

  $M = M + x_t x_t{}'$

  $M^{-1} = \sqrt{D_w}\left(aI + \sqrt{D_w}\, M \sqrt{D_w}\right)^{-1}\sqrt{D_w}$

Read $y_t \in \mathbb{R}$
$b = b + x_t y_t$
$w = M^{-1} b$
END FOR

If the data has significant outlier(s), then we suggest replacing $\gamma_t = w' x_t$ by the following:

$$\gamma_t = \frac{w' x_t}{1 + x_t' \sqrt{D_w} \left(aI + \sqrt{D_w} \sum_{s=1}^{t} x_s x_s' \sqrt{D_w} \right)^{-1} \sqrt{D_w}\, x_t}$$

The above replacement of prediction leads to an algorithm that is more immune to overfitting in presence of outlier(s). For more details, please see [10].

6.2 Competitive Online Iterated Ridge Regression (COIRR)

The pseudo-code of the algorithm is as follows:

Initialise: $B = aI^{n \times n}, b = 0^{n \times 1}$, $a > 0$
FOR $\boldsymbol{t = 1, 2, \ldots}$
Read $x_t \in \mathbb{R}^n$
output $\gamma_t = b' B^{-1} x_t$
$B = B + x_t x_t'$
Read $y_t \in \mathbb{R}$
$b = b + x_t y_t$
END FOR

If the data has significant outlier(s), then $\gamma_t = b' B^{-1} x_t$ is replaced with the following:

$$\gamma_t = \frac{b' B^{-1} x_t}{1 + x_t' (aI + \sum_{s=1}^{t} x_s x_s')^{-1} x_t}$$

The above replacement of prediction leads to an algorithm that is more immune to overfitting in presence of outlier(s). For more details, please see [10,11].

6.3 Online Normalised Least Mean Squares Regression (ONLMS)

The pseudo-code of the algorithm is as follows:

Initialise: $w = 0^{n \times 1}$, $-\infty < \eta < \infty$
FOR $\boldsymbol{t = 1, 2, \ldots}$
Read $x_t \in \mathbb{R}^n$
output $\gamma_t = w' x_t$
Read $y_t \in \mathbb{R}$
Normalise loss $\lambda = \frac{\gamma_t - y_t}{\eta + ||x_t||_2^2}$
$w = w + \lambda x_t$
END FOR

7. Empirical study

We used data collected from house number 24. The data has 150 attributes measured for 35 days every second. To make daily prediction we consider a moving average of each 86400 seconds (24 hours) as portrayed in the following pseudocode.

Initialise: $c = 0, S = 0^{n \times 1}$
FOR $\boldsymbol{t = 1, 2, \ldots, 86400}$
  Read $x_t \in \mathbb{R}^n$
  c = c+1
  $S = S + x_t$
  $mean = \frac{S}{c}$
END FOR

The execution of such code results in 35 data points (for Gas, Temp 1 and Temp 2) which are then used to fit the regression models COIRR, ONLMSR and OSLOG. Specifically, the regression models are trained online and allowed to make sequential prediction of the average of the next 24 hours whenever a data point (average of 24 hours) is presented. The simulation produced the results compiled in Table 2 using R^2 which measures the proportion of explained variation to the total variation of the output sequence (predicted values). R^2 is very popular measure for evaluating regression models. R^2 close to 0 indicates that the model does not explain the variability of the output (dependent variable) around its mean, while a value closer to 100% indicates that the model explains well the variability of the output around its mean.

Table 2: **R^2** results of the three algorithms

Algorithm \ R^2	Gas	Temp1	Temp2
COIRR	**0.79**	0.37	0.61
ONLMSR	0.52	0.06	0.08
OSLOG	0.11	**0.45**	**0.67**

To improve the predictive accuracy, we did feature selection in order to consider only potentially relevant features.

We use the first 86400 data points to learn the features that affect gas and hot temperature. Popular methods, the forward and backward selection, are used. In forward selection we start from a null model and add features sequentially. After adding each feature, we compute a statistical measure such as Akaike Information Criterion (AIC) and check if AIC improves. Improvement means that the added feature is important and thus worth including it into the model. Backward selection follows the same procedure, but instead of starting from null model, we start from the full model (with all features) and remove features that contribute less to the model. For further details on forward and backward feature selection, please see [12]. On the experimented 2 days of the data, the following feature(s) are retained[1]:

- Gas: 150, 147, 143, 141, 131, 123, 120, 118, 108, 53, 14

[1] The full numbering of features is provided in the Appendix.

- Temp 1: 142, 128, 118,117, 115, 88, 19
- Temp 2: 150, 147, 132, 110, 108, 105, 67, 61, 52, 36, 35, 15, 2

After performing model selection and tuning of the parameter a (see Table 3) on the first 86400 data points, we obtained the R^2 results shown in Table 4.

Table 3: Values of the parameter *a*

Algorithm \ *a*	Gas	Temp 1	Temp 2
COIRR	6	0.9	0.7
OSLOG	4	0.6	1.6
ONLMSR	0.8	0.2	0.6

Table 4: Accuracy of the algorithms

Algorithm \ R^2	Gas	Temp 1	Temp 2
COIRR	**0.93**	0.34	0.49
ONLMSR	0.78	0.03	**0.95**
OSLOG	0.10	0.65	**0.91**

Comparing these results against those portrayed in Table 2, it is clear that the selection of important features has had very good impact.

Moreover, in order to appreciate more the quality of the fit of data, the accuracy of the prediction is illustrated in Figures 1-3.

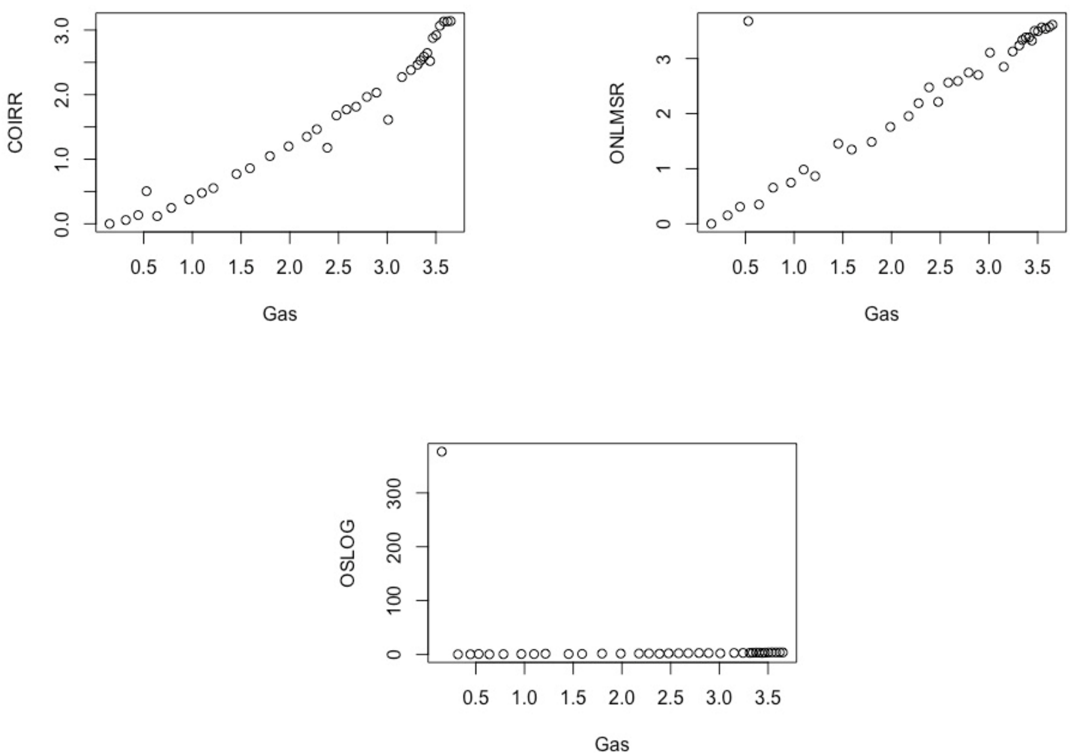

Figure 1: Gas prediction with COIRR, ONLMSR and OSLOG

Figure 1 shows that COIRR and ONLMSR perform very well on the task of predicting gas consumption, while OSLOG performs poorly. Notice, ONLMSR R^2 statistic is not as good as COIRR. This is because ONLMSR does not predict a certain observation well. After, further inspection it was discovered ONLMSR does not fit the outlier well, whereas COIRR can cope up better with the outlier.

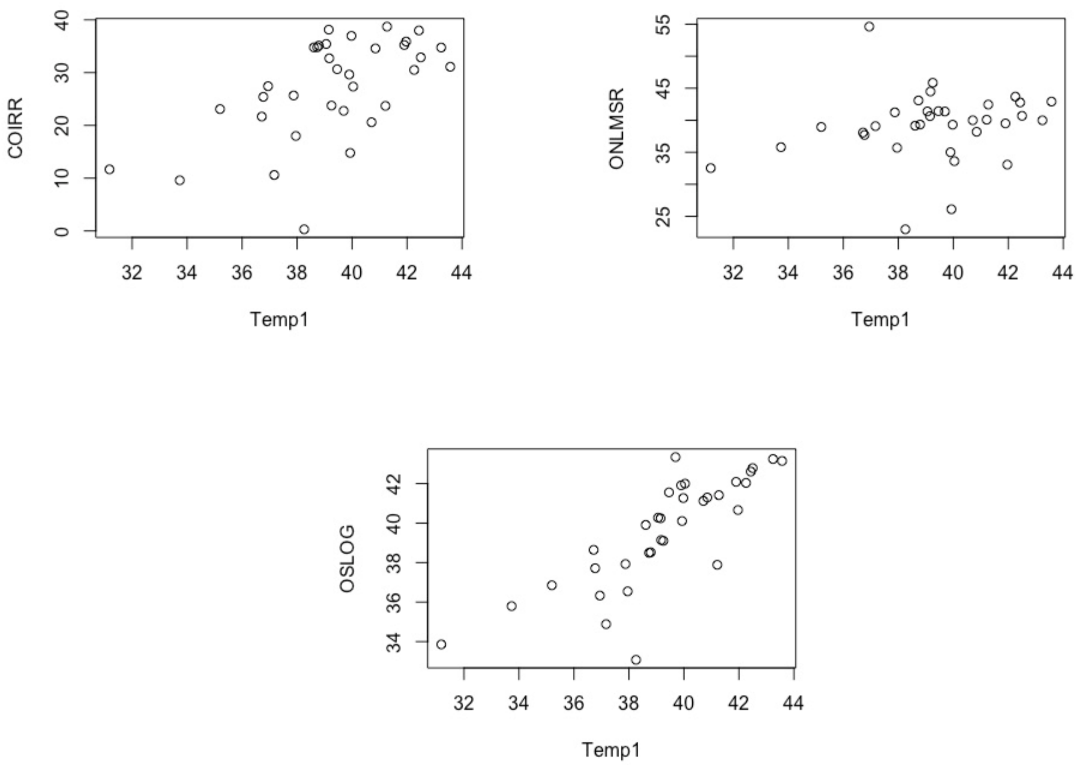

Figure 2: Temp 1 prediction with COIRR, ONLMSR and OSLOG

All algorithms did not do well on the prediction of Temp 1 as shown in Figure 2. It seems that the input features do not correlate well with the response output Temp 1 based on only the sample of data used. OSLOG is the only algorithm that explains more than 50% of the variability.

For the prediction of the hot water temperature for the central heating, Temp 2, Figure 3 indicates that ONLMSR and OSLOG perform well. Here COIRR performs poorly.

These results suggest that gas consumption and temperature of central heating can be predicted well using the proposed algorithms. The prediction of Temp1 is not as good as others. It is worth noting that OSLOG and ONLMSR are more likely to overestimate, while COIRR is likely to underestimate.

Furthermore, the simulations suggest that in presence of a substantial outlier, OSLOG can perform poorly overall, ONLMSR most likely will not predict the substantial outlier well. On the other hand, COIRR is likely to handle the outlier.

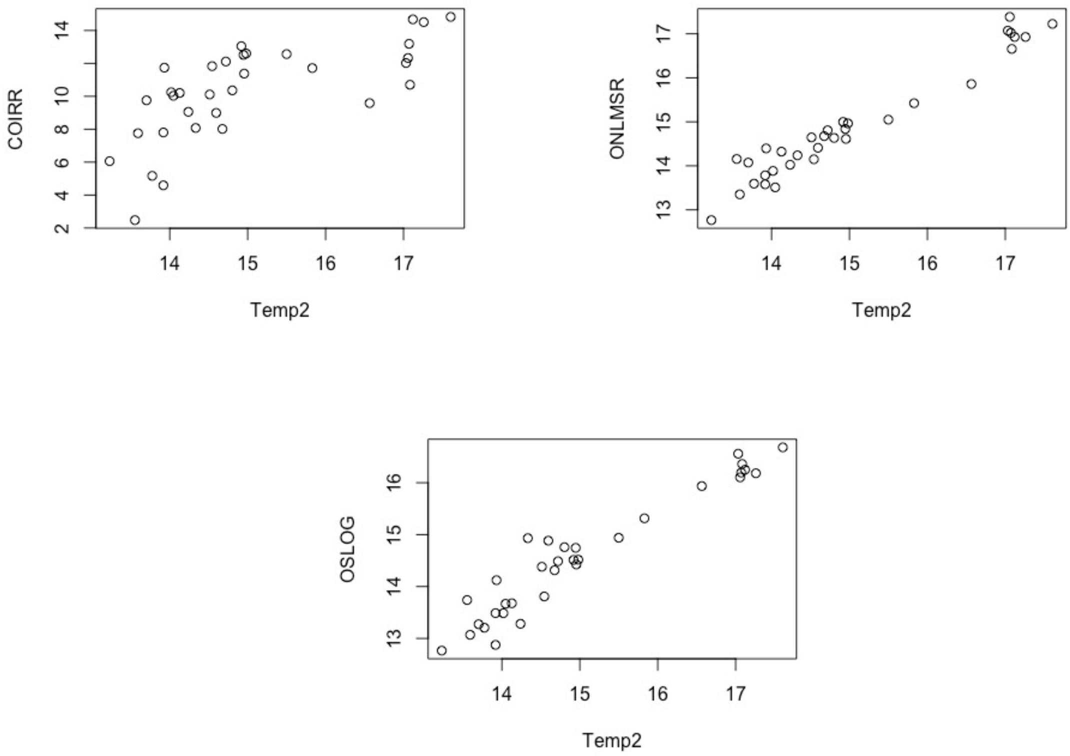

Figure 3: Temp 2 prediction with COIRR, ONLMSR and OSLOG

8. Conclusion

This report describes three novel online algorithms: Competitive Online Iterated Ridge Regression (COIRR), Online Normalised Least Mean Squares Regression (ONLMS) and Online Shrinkage via Limit of Gibbs sampling (OSLOG) to predict gas consumption and hot water usage (the hot temperature of the domestic heating and the hot temperature of the central heating flow). The proposed prediction algorithms show gas consumption and temperature of the central heating are quantities that can be predicted with a reasonable accuracy.

9. References

[1] https://www.proteus-bigdata.com/project-description/scalable-online-machine-learning/

[2] Hoffman, Matthew D., et al. "Stochastic variational inference." Journal of Machine Learning Research 14.1 (2013): 1303-1347.

[3] Mohamad, Saad, Abdelhamid Bouchachia, and Moamar Sayed-Mouchaweh. "A Bi-Criteria Active Learning Algorithm for Dynamic Data Streams." IEEE Transactions on Neural Networks and Learning Systems .2016.

[4] Gama, João, et al. "A survey on concept drift adaptation." ACM Computing Surveys (CSUR) 46.4 (2014): 44.

[5] Saad Mohamad, Abdelhamid Bouchachia "Online Gaussian LDA for unsupervised pattern mining from utility usage data". Submitted to ECML-PKDD, 2018

[6] Saad Mohamad, Damla Arifoglu, Chemseddine Mansouri, and Abdelhamid Bouchachia "Deep online hierarchical unsupervised learning for pattern mining from utility usage data". Submitted to UKCI, 2018

[7] Saad Mohamad, Damla Arifoglu, Chemseddine Mansouri, and Abdelhamid Bouchachia "Deep Online Hierarchical Dynamic Unsupervised Learning for Pattern Mining from Utility Usage Data". Submitted to PAMI, 2018

[8] Shafer, Glenn, and Vladimir Vovk. Probability and finance: it's only a game!. Vol. 491. John Wiley & Sons, 2005.

[9] Waqas Jamil and Hamid Bouchachia. 2018. Competitive Normalised Least Squares Regression. In preparation (2018)

[10] Waqas Jamil and Hamid Bouchachia. 2018. Competitive Online Regularised Regression. Machine Learning (submitted) (2018).

[11] Volodya Vovk. 2001. Competitive On-line Statistics. International Statistical Review 69, 2 (2001), 213–248.

[12] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

Appendix

"X1","Time_stemps"	"X51","RMS Spectrum power"	"X101","RMS Spectrum power"
"X2","Gas_data"	"X52","RMS Spectrum power"	"X102","RMS Spectrum power"
"X3","Real power"	"X53","RMS Spectrum power"	"X103","RMS Spectrum power"
"X4","Reactive power"	"X54","RMS Spectrum power"	"X104","RMS Spectrum power"
"X5","RMS Spectrum power"	"X55","RMS Spectrum power"	"X105","Water input (quantity)"
"X6","RMS Spectrum power"	"X56","RMS Spectrum power"	"X106","Room temperature"
"X7","RMS Spectrum power"	"X57","RMS Spectrum power"	"X107","Room temperature"
"X8","RMS Spectrum power"	"X58","RMS Spectrum power"	"X108","Room temperature"
"X9","RMS Spectrum power"	"X59","RMS Spectrum power"	"X109","Room temperature"
"X10","RMS Spectrum power"	"X60","RMS Spectrum power"	"X110","Room temperature"
"X11","RMS Spectrum power"	"X61","RMS Spectrum power"	"X111","Room temperature"
"X12","RMS Spectrum power"	"X62","RMS Spectrum power"	"X112","Room temperature"
"X13","RMS Spectrum power"	"X63","RMS Spectrum power"	"X113","Room temperature"
"X14","RMS Spectrum power"	"X64","RMS Spectrum power"	"X114","Room temperature"
"X15","RMS Spectrum power"	"X65","RMS Spectrum power"	"X115","Room temperature"
"X16","RMS Spectrum power"	"X66","RMS Spectrum power"	"X116","Room temperature"
"X17","RMS Spectrum power"	"X67","RMS Spectrum power"	"X117","Hot Water temperature"
"X18","RMS Spectrum power"	"X68","RMS Spectrum power"	
"X19","RMS Spectrum power"	"X69","RMS Spectrum power"	"X118","Hot Water temperature"
"X20","RMS Spectrum power"	"X70","RMS Spectrum power"	
"X21","RMS Spectrum power"	"X71","RMS Spectrum power"	"X119","Hot Water temperature"
"X22","RMS Spectrum power"	"X72","RMS Spectrum power"	
"X23","RMS Spectrum power"	"X73","RMS Spectrum power"	"X120","Cold Water temperature"
"X24","RMS Spectrum power"	"X74","RMS Spectrum power"	
"X25","RMS Spectrum power"	"X75","RMS Spectrum power"	"X121","Radiator temperature"
"X26","RMS Spectrum power"	"X76","RMS Spectrum power"	"X122","Radiator temperature"
"X27","RMS Spectrum power"	"X77","RMS Spectrum power"	"X123","Radiator temperature"
"X28","RMS Spectrum power"	"X78","RMS Spectrum power"	"X124","Radiator temperature"
"X29","RMS Spectrum power"	"X79","RMS Spectrum power"	"X125","Radiator temperature"
"X30","RMS Spectrum power"	"X80","RMS Spectrum power"	"X126","Radiator temperature"
"X31","RMS Spectrum power"	"X81","RMS Spectrum power"	"X127","Radiator temperature"

"X32","RMS Spectrum power" "X33","RMS Spectrum power" "X34","RMS Spectrum power" "X35","RMS Spectrum power" "X36","RMS Spectrum power" "X37","RMS Spectrum power" "X38","RMS Spectrum power" "X39","RMS Spectrum power" "X40","RMS Spectrum power" "X41","RMS Spectrum power" "X42","RMS Spectrum power" "X43","RMS Spectrum power" "X44","RMS Spectrum power" "X45","RMS Spectrum power" "X46","RMS Spectrum power" "X47","RMS Spectrum power" "X48","RMS Spectrum power" "X49","RMS Spectrum power" "X50","RMS Spectrum power"	"X82","RMS Spectrum power" "X83","RMS Spectrum power" "X84","RMS Spectrum power" "X85","RMS Spectrum power" "X86","RMS Spectrum power" "X87","RMS Spectrum power" "X88","RMS Spectrum power" "X89","RMS Spectrum power" "X90","RMS Spectrum power" "X91","RMS Spectrum power" "X92","RMS Spectrum power" "X93","RMS Spectrum power" "X94","RMS Spectrum power" "X95","RMS Spectrum power" "X96","RMS Spectrum power" "X97","RMS Spectrum power" "X98","RMS Spectrum power" "X99","RMS Spectrum power" "X100","RMS Spectrum power"	"X128","Radiator temperature" "X129","Radiator temperature" "X130","Radiator temperature" "X131","Room humidity" "X132","Room humidity" "X133","Room humidity" "X134","Room humidity" "X135","Room humidity" "X136","Room humidity" "X137","Room humidity" "X138","Room humidity" "X139","Room humidity" "X140","Room humidity" "X141","Room humidity" "X142","Boiler" "X143","Valves" "X144","Valves" "X145","Valves" "X146","Valves" "X147","Valves" "X148","Valves" "X149","Valves" "X150","Valves"