Igor Trotskii and Jukka Pulkkinen

Unsupervised machine learning model for heat flow monitoring in a geothermal energy storage in a near-zero-energy-building

Abstract: With a fast-paced development in IoT, information processing and process monitoring techniques, building automation systems become more and more complex and advanced. Evolution of these technologies allows to greatly improve buildings' energy performance and substantially decrease maintenance costs by utilizing such ideas as Condition Based Maintenance (CBM) and Machine Learning, which CBM heavily relies on.

As most of the building are still maintained through normal means, such as reactive and schedule-based maintenance, because of lack of property managers' interest in investing in more efficient approaches, this paper's aim is to cover this gap by providing short overview of most basic machine learning and data processing algorithms used in building maintenance domain and by providing case study results. The case study was conducted in a near-zero-energy building, Sheet Metal Center in Visamäki, Hämeenlinna by constructing Principal Component Analysis based solution for monitoring energy flows inside geothermal energy storage for further usage in CBM. The solution helps to quickly evaluate the state of the system and allows to simplify diagnosis and faults localization.

Keywords: Condition Based Maintenance, Principal Component Analysis, near-zero-energy building

Igor Trotskii: HAMK, E-mail: igor.trotskii@hamk.fi

Jukka Pulkkinen: HAMK, E-mail: jukka.pulkkinen@hamk.fi

1 Introduction

CBM has been an interesting topic since long ago as it allows to greatly reduce downtime of the process without the need for frequent equipment checks and replacements as scheduled maintenance dictates. [19] Nevertheless, today maintenance is still done mainly based on traditional scheduled maintenance. [18] The main limitation of CBM adoption is a high initial cost for modern equipment [17], which has embedded self-diagnostic capabilities, therefore could report about its own condition. On the other side, Internet of Things has been developing rapidly, bringing the ease of obtaining enormous amount of data from monitored process without any big investment costs. This makes it possible to utilize measured data in conjunction with machine learning to build inexpensive CBM systems, avoiding costly investments in a new equipment by moving part of expenses to the software by monitoring and analyzing parameters which have indirect impact on the monitored equipment. [8] As mentioned above, CBM heavily relies on machine learning in order to forecast building and equipment performance to be able to evaluate best possible repairment schedule. The reason for that is the complexity of the data from modern automation systems. Even the simplest system can have dozens of different metrics and measurements, which have to be analyzed in order to estimate devices and equipment status and if they should be replaced or fixed. One way to solve that problem is by applying statistical methods and machine learning.

The main purpose of this study is to determine possible data analytics techniques to utilize in building maintenance domain. Even though, many algorithms for fault and anomaly detection are already present, it is still hard to select one for simplifying CBM implementation.

This paper provides a brief overview of available techniques, which are widely used in building maintenance domains and their possible applications. The usefulness of the CBM and machine learning approach for building maintenance is further proved by conducting a case study in Sheet Metal Center in Visamäki, Hämeenlinna, which belongs to near-zero-energy buildings class.

2 Literature review

Numerous techniques are known, which can be applied for improving building maintenance efficiency by helping to adopt key principals of CBM. Most of these techniques belong to anomaly detection domain as the information about faults, abnormalities or their possibilities is of utmost importance for performing repairment in economical way. The list of the available algorithms includes, but is not limited to, Principal Component Analysis, Support Vector Machine and Neural Networks.

The problem encountered by the authors is the lack of research in building maintenance and especially in maintenance and monitoring of near-zero-energy building domains. Even though, amount of developed techniques related to anomaly detection and Condition Based Maintenance is high, their possible applications and implementation for nZEB is still under study.

2.1 Principal Component Analysis

Principal component analysis is one of the most widely used multivariate statistical techniques used for dimensionality reduction and anomaly detection. PCA is based on an orthogonal decomposition of original correlated measurements into new space of lower dimensionality, which consists of new uncorrelated variables called principal components. New variables are selected in a way to maximize explained variance, what allows to significantly reduce amount of measurements by preserving as much information from original data as possible. [21]

PCA is a well-known technique in a process monitoring domain and even in its simplest form can be very useful for automatic fault detection. For example researchers have used this approach in asynchronous generators [14], wastewater treatment plant [7] and reciprocating compressors. [1]

Anomaly detection applications based on PCA often use such metrics as SPE and Hotelling's T^2 metrics (Mujica, Rodellar, Fernández, & Güemes, 2011), which are relatively simple to calculate and give accurate estimation if situation anomalous or not. However, it can be very difficult to localize the source of the fault. Fortunately, various techniques were developed to handle this problem, e.g. contribution plots. [16]

2.2 Other techniques

Principal component analysis is not the only popular tool in process monitoring and fault detection. Two other popular techniques in building maintenance domain is support vector machine (SVM) and Artificial Neural Networks (ANN).

Support vector machine is a supervised learning algorithm mainly used for classification, however, it is possible to perform regression tasks as well. The main idea behind SVMs is transforming a dataset to a higher dimensional space in order to make initial complex non-linear relations into simple linear separable clusters, but of higher dimensionality. [3]

SVMs are widely used for building energy efficiency evaluation [4], predicting electrical energy consumption [5] and forecasting of cooling/heating load for HVAC systems. [13] These methods can be useful for improving building management or tuning building automation system, however, they are not directly relevant to maintenance efficiency. More direct approach is applying SVM for fault detection. [2, 9]

Artificial Neural Networks have already revolutionized many industries and so it is only natural that they are a powerful tool in maintenance domain. The benefit of utilizing ANNs is their ability to solve both classification and regression problems. Neural Networks are widely used for forecasting building energy consumption and for CBM directly. [6, 12]

ANNs provide superior accuracy of prediction or fault detection, but they usually require big amount of recorded measurements in order to achieve high efficiency and so their usefulness can be severely limited especially in new buildings.

3 Case Study: Sheet Metal Center

3.1 Background of the study

The new testing hall for HAMK Sheet Metal Center (SMC) was built in 2015. The building is near-zero-energy building based on different technologies such as compact envelope, energy saving windows, effective heat recovery in air handling units, building automation and renewable energy sources. The renewable energy consists of the solar and geothermal energy units. The geothermal part is the main heat supplier and it includes energy piles and heat wells. The solar energy units are used to fill energy piles with thermal energy.

Fig 1. Sheet Metal Center testing hall

The behavior of the heating and cooling system depends primarily on the season, hence it is possible to define two main operational modes: heating season and cooling season. The operation of the system during one of the modes can be described by thermal energy flows between different parts shown in the Fig. 2. During heating season both energy piles and heat well must provide energy to the heat pump as long as their outlet liquid temperatures are not bellow certain threshold in order to prevent the formation of ice. There is no heating demand in the system during the cooling season, so heat pump doesn't operate, and energy piles are separated into a separate loop to fill them with heat from solar collectors and air handling unit exhaust air via heat exchanger; heat well is used as a heat sink for the building.

Although, the state of the system is very clear and easy to monitor with normal means during summer or winter it is not the case during autumn, spring or maintenance when it can oscillate between two states, what can lead to different anomalies and faults. The designed monitoring method addresses this problem.

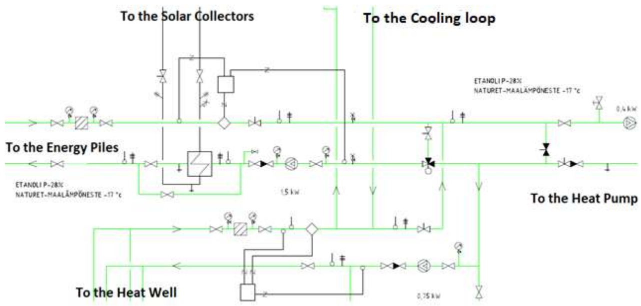

Fig. 2. Part of the heating system under the study

3.2 Data Selection

More than 130 process variables are monitored for the SMC building. Therefore, feature selection is required to select only relevant variable, otherwise the resulting model could be irrelevant to the area of interest, or worse, it is possible that no reasonable output could be derived. The feature selection for the designed model was based on physical locality principle, as a result, only measurements from the sensors close to the energy piles, heat well, heat pump and cooling loop connection were used. The final measurement list is shown in Table 1.

Measurement
Sun-heated glycol return temperature
Sun-heated glycol temperature at the top of the storage tank
Sun-heated glycol incoming temperature (meter data)
Sun-heated glycol return temperature (meter data)
Well pipes outcoming ethanol temperature
Well pipes incoming ethanol temperature
Well pipes ethanol flow rate
Well pipes incoming ethanol temperature (meter data)
Well pipes return ethanol temperature (meter data)
Well pipes outcoming ethanol temperature
Well pipes incoming ethanol temperature
Ethanol temperature after well pipes and roof radiant heaters
Ceiling heating outcoming ethanol temperature
Ceiling heating incoming ethanol temperature
Energy piles ethanol flow rate
Energy piles incoming ethanol temperature
Energy piles return ethanol temperature
Ethanol temperature before heat-exchanger with sun-heated glycol
Well pipes ethanol mixing valve

Table 1. List of the measurements used for analysis

3.3. Modelling technique selection

For this case study traditional principal component analysis was chosen instead of ANNs or SVM/SVR for following reasons:

1. Authors had access to limited amount of data: HAMK Sheet Metal Center is relatively new building and amount of available measurements is less than one year and a half, what limits the usefulness of ANNs.
2. There are gaps in available data due to various reasons, the ability to evaluate autocorrelation of the studied system. PCA considers every measurement point as a separate one, and thus its accuracy does not suffer because of the gaps.
3. SVM/SVR are very hard to use for process

monitoring and fault detection in case the faults themselves are not strictly defined as SVM/SVR are supervised learning algorithms.
4. PCA is well supported by such metrics as SPE and Hotelling's T2 statistics, which make fault detection a lot easier.

3.4. Modelling

The dimensionality of the selected dataset strongly suggests the use of some dimensionality reduction techniques such as principal component analysis (PCA). The main idea of PCA is transforming high dimensional datasets by projecting points onto a new lower dimensional space. The resulted uncorrelated variables – Principal components can be regarded as a linear combination of the original variables. PCA is very sensitive to the scaling of the variables as the method is based on calculation of the covariance matrix, hence it requires standardization of the data, so zero mean – unit variance scaling was applied. [10]

PCA is one of the best techniques for multivariate statistical analysis and it was chosen as it doesn't require much of prior knowledge about the process, which generated data. Another concern is inconsistency in data and missing values: traditional PCA does not consider autocorrelation, therefore it is not sensitive to gaps in training data unlike such methods as Dynamic PCA. [20]

3.5. Results

3.5.1. Interpretation and verification of principal components.

The resulting principal components were analyzed based on a correlation matrix for the principal components and original parameters. The results are shown in Table 2.

Principal Component	Explained variance	Meaning
1	64.14%	Heat flow into the energy piles and heat well
2	20.12%	Heat flow from the heat well into the heat pump and cooling system
3	6.26%	Random oscillations of the parameters
4	4.50%	

Table 2. Explained variance and meaning behind each principal component

The assumptions made in the Table 2 were further confirmed by applying k-means clustering to the data obtained from principal component analysis, the result of which is shown in the Fig 3. As it can be seen from the Table 2, most of the variance is explained by first two principal components (PCs) and the difference between them is essentially the heat flow direction, meaning that the simultaneous changes in both PCs can only be explained by changing the operational mode from heating to cooling or vice versa, hence the result of clustering should display the state of the heat pump: if it is on or off.

K-means clustering was performed on obtained four principal components, which cover three different operation periods:
1. From 23.10.2018 to 05.11.2018 — normal winter operation before heat pump maintenance.
2. From 05.11.2018 to 20.11.2018 — heat pump maintenance.
3. From 20.11.2018 to 01.02.2019 — normal winter operation after heat pump maintenance.

Labels obtained from clustering can be explained as follows:
- 0 – The heat pump is on during the third period.
- 1 – The heat pump is on during first period or off during the first period.
- 2 – The heat pump is under maintenance.

Class 1 brings some confusement as it can mean both on and off states of the heat pump depending on the period of time it belongs to. The reason for this is that obtained clusters are based on the heat flow between energy piles, heat well and heat pump and not heat pump production, and the flows changed a lot after the heat pump maintenance. However, heat pump production is strongly correlated to the heat flows in the system. In general the cluster can represent if energy flows into the ground, into the heat pump or if something anomalous, e.g. maintenance, happens. Oscillation of the clusters' labels are explained by the fact that the heat pump is currently used on/off controller and so it changes its state frequently. Accuracy metric was defined based on previously defined meaning of the labels and heat pump production level and essentially represent how accurately clustering is able to differentiate between on/off and maintenance states of the heat pump. The label was correct in 77.36% of cases.

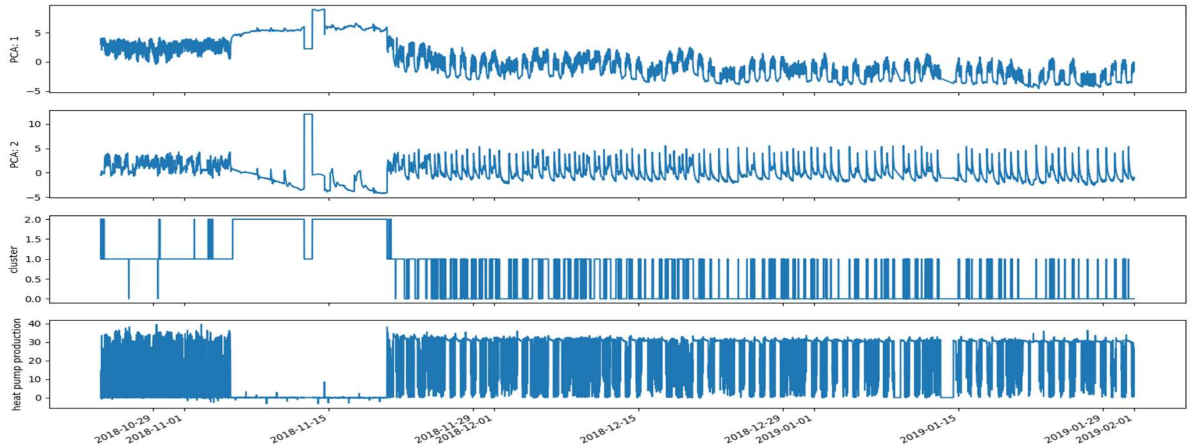

Fig 3. Clustering and comparison between clustering results and real measurements from the heat pump.

3.3.2. Fault detection with Hotelling's T^2 and SPE statistics.

Process monitoring and fault detection with principal component analysis is majorly based on T^2 and SPE metrics. Both these metrics can be used in order to find outliers in respect to original training set. Big SPE or T^2 value does not necessary mean that the system is in faulty state, however, if these values are abnormally high over extended periods of time then it might be worth some investigation.

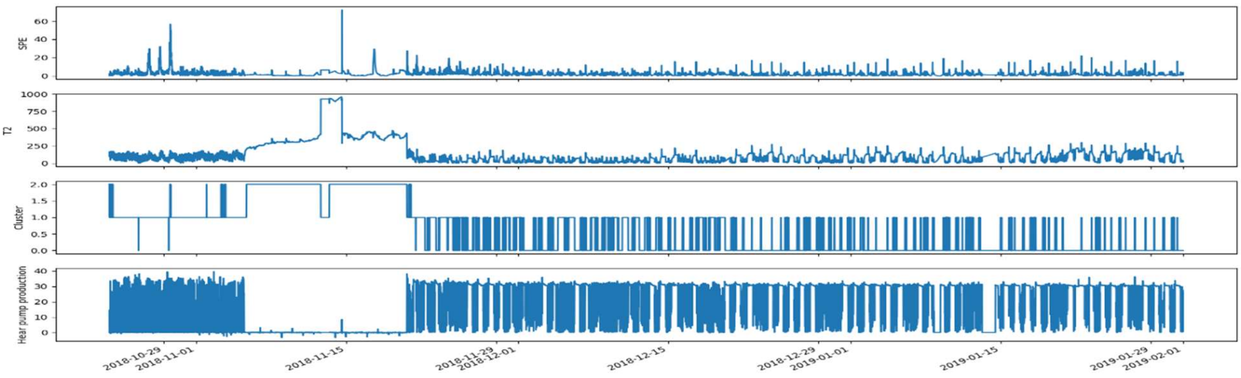

Fig 4. Hotelling's T^2 and SPE metrics.

Result of using T^2 and SPE metrics is shown on Fig 4. Most noticeable outliers happened during maintenance break. On 15.11.2018 heat pump was on for a very short period of time, when it was supposed to be off, what resulted in SPE spike. The second anomaly is covered by T^2: maintenance period itself can be considered as a deviation from the normal operation and thus it causes growth in the statistic. The squared spike is explained by missing measurements, which caused even further deviation from normal values.

4 Conclusion

In this paper a data analyzing method based on principal component analysis was developed for a geothermal energy storage, allowing to easily monitor the current state of the geothermal energy storage in order to schedule maintenance efficiently by adapting the key principles of the Condition Based Maintenance without high investment costs. However, due to the indirect nature of the measurements used for analysis, the developed tool can only be used as a supplementary tool due to its inability to accurately show the exact faulty or degrading part as the model evaluates performance of the system as a whole. Nevertheless, the designed model produces reliable results which can be used to better pinpoint time of the fault occurring.

References

[1] Ahmed, M., Baqqar, M., Gu, F., & Ball, A. D. (2012). Fault detection and diagnosis using Principal Component Analysis of vibration data from a reciprocating compressor. In *Proceedings of 2012 UKACC International Conference on Control* (pp. 461–466). IEEE. https://doi.org/10.1109/CONTROL.2012.6334674

[2] Chitrakar, R., & Chuanhe, H. (2012). Anomaly detection using Support Vector Machine classification with k-Medoids clustering. In *2012 Third Asian Himalayas International Conference on Internet* (pp. 1–5). IEEE. https://doi.org/10.1109/AHICI.2012.6408446

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

[4] de Wilde, P., Martinez-Ortiz, C., Pearson, D., Beynon, I., Beck, M., & Barlow, N. (2013). Building simulation approaches for the training of automated data analysis tools in building energy management. *Advanced Engineering Informatics*, *27*(4), 457–465. https://doi.org/10.1016/J.AEI.2013.05.001

[5] Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, *37*(5), 545–553. https://doi.org/10.1016/J.ENBUILD.2004.09.009

[6] Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, *40*(5), 356–362. https://doi.org/10.1016/J.ADVENGSOFT.2008.05.003

[7] Garcia-Alvarez, D. (2009). FAULT DETECTION USING PRINCIPAL COMPONENT ANALYSIS (PCA) IN A WASTEWATER TREATMENT PLANT (WWTP).

[8] Ghasemi, A., Yacout, S., & Ouali, M.-S. (2010). Parameter Estimation Methods for Condition-Based Maintenance With Indirect Observations. *IEEE Transactions on Reliability*, *59*(2), 426–439. https://doi.org/10.1109/TR.2010.2048736

[9] Heller, K., Svore, K., Keromytis, A. D., & Stolfo, S. (2003). One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses, 2–9. https://doi.org/10.7916/D84B39Q0

[10] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

[11] Krenek, J., Kuca, K., Blazek, P., Krejcar, O., & Jun, D. (2016). Application of Artificial Neural Networks in Condition Based Predictive Maintenance (pp. 75–86). https://doi.org/10.1007/978-3-319-31277-4_7

[12] Kumar, R., Aggarwal, R. K., & Sharma, J. D. (2013). Energy analysis of a building using artificial neural network: A review. *Energy and Buildings*, *65*, 352–358. https://doi.org/10.1016/J.ENBUILD.2013.06.007

[13] Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. *Energy Conversion and Management*, *50*(1), 90–96. https://doi.org/10.1016/J.ENCONMAN.2008.08.033

[14] Martón, I., Sánchez, A., Carlos, S., & Martorell, S. (2013). Application of Data Driven Methods for Condition Monitoring Maintenance. In *Chemical Engineering Transactions* (Vol. 33, pp. 301–306). https://doi.org/10.3303/CET1333051

[15] Mnassri, B., Adel, E. M. El, Ananou, B., & Ouladsine, M. (2009). Fault Detection and Diagnosis Based on PCA and a New Contribution Plot. *IFAC Proceedings Volumes*, *42*(8), 834–839. https://doi.org/10.3182/20090630-4-ES-2003.00137

[16] Mujica, L., Rodellar, J., Fernández, A., & Güemes, A. (2011). Q-statistic and T2-statistic PCA-based measures for damage assessment in structures. *Structural Health Monitoring: An International Journal*, *10*(5), 539–553. https://doi.org/10.1177/1475921710388972

[17] Rastegari, A., & Bengtsson, M. (2015). Cost effectiveness of condition based maintenance in manufacturing. In *2015 Annual Reliability and Maintainability Symposium (RAMS)* (pp. 1–6). IEEE. https://doi.org/10.1109/RAMS.2015.7105079

[18] Srivastava, N., & Mondal, S. (2013). *Maintenance Practices in Indian Manufacturing Industries*. *International Journal of Computer Science & Management Studies* (Vol. 13).

[19] Stenholm, A., & Andersson, D. (2014). On

condition-based maintenance for machine components. Retrieved from https://lup.lub.lu.se/student-papers/search/publication/8593562

[20] Vanhatalo, E., Kulahci, M., & Bergquist, B. (2017). On the structure of dynamic principal component analysis used in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, *167*, 1–11. https://doi.org/10.1016/J.CHEMOLAB.2017.05.016

[21] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9