eSim 2018
«BUILDING SIMULATION TO SUPPORT BUILDING SUSTAINABILITY»

IBPSA Canada
The Canadian Chapter of the International Building Performance Simulation Association
Le chapitre canadien de International Building Performance Simulation Association

Principal Component Analysis for the ongoing commissioning of northern houses

Behrad Bezyan, Radu Zmeureanu

Centre for Net-Zero Energy Buildings Studies, Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, Canada

***Abstract*:** *This paper presents the application of the Principal Component Analysis (PCA) for defining the threshold of normal operation of the heating system in two houses, detecting the outliers in the PC-based space of the heating system operation, and identifying those system variables which are the source of outliers. This case study uses measurements collected in December 2014 in two houses of Inuvik, NWT, as the training data set, which is then applied to measurements of February 2015 as the application data set. The temperature of supply and return water temperature for heating one house are the major sources of outliers identified from data of February 2015. The identification by the PCA of variables with abnormal values is validated by using of a modified data set.*

Keywords: *Principal Component Analysis, outlier detection, heating, ongoing commissioning, northern houses*

INTRODUCTION

Building commissioning is a process of measuring and evaluation of the energy performance of a building and HVAC system with respect to the design intent and standard performance, Roth et al. (2008). Katipamula and Brambley (2005) showed that 15% to 30% of the energy use in buildings is wasted due to the degradation of HVAC equipment, unsuitable scheduled maintenance and systems control. A high proportion of the energy waste can be avoided with proper maintenance, by using automated fault detection and diagnosis (FDD). Several studies focused on the FDD of HVAC equipment by using different data driven models, Jia and Reddy (2003). One practical problem related to the use of accurate and time-efficient FDD techniques comes from the large number of measurements of several variables over extended intervals of time, many of those variables being correlated.

Principal Component Analysis (PCA) method is a multivariate technique, used in other fields such as Draper and Smith (1981), Hadley and Tomich (1986), that can mitigate this problem by transforming the original data set of j inter-correlated variables into a smaller data set of k independent new variables, where k < j, known as Principal Components (PCs).

The literature review presented by Cotrufo and Zmeureanu (2016) on the use of PCA for FDDs applications is expanded with the papers presented in this section. Xu et al. (2008) concluded that the PCA gives useful residuals for sensor-fault detection, diagnosis and estimation. However, the conventional PCA-based strategy that directly employs raw measurements is less efficient because of noises and dynamics embodied in the data, than the proposed strategy using wavelet analysis and PCA.

Hu et al. (2012) developed a self-Adaptive Principal Component Analysis (APCA) method that can remove erroneous temperature measurements with absolute magnitude less than 1ºC, which increases the efficiency of fault detection.

Cotrufo and Zmeureanu (2016) proposed the use of PCA method for the fault detection of sensors and degradation in equipment performance of a system.

Gajjar et al. (2016) recognized that the PCA was widely used for process fault detection, however the interpretation of the principal components (PCs) is a challenging task since each PC is a linear combination of the original variables. They applied the sparse-PCA by restricting some PC loadings to zero, which results in a clearer interpretation of results.

Beghi et al. (2016) used the PCA to identify anomalies from normal operation variability and isolated variables related to faults, without having any prior knowledge about abnormal measurements. They used Statistical Process Monitoring (SPM) approach to monitor the behavior of the system, and a simple diagnosis table, based on experts' diagnosis rules, to identify the shifts from the nominal working conditions.

Hu et al. (2016a) estimated the undetectable boundary of each sensor assigned in PCA model, and the fault detection efficiency of eight sensors installed in a chiller, for different severe levels of faults. Hu et al. (2016b) started from the observation that the results of any data-driven method are highly dependent on the quality of the training data. They developed a statistical training data cleaning strategy for

ISBN 978-2-921145-88-6

PCA-based for FDD & Reconstruction method for chiller sensors. The method called SPCA uses the Euclidean distance to find out outliers in the measurement data set.

Guo et al. (2017a) developed an enhanced sensor fault detection and diagnosis method for the variable refrigerant flow (VRF) system based on Savitzky-Golay (SG) method of the PCA. The SG method is used for data smoothing. They concluded that the SG-PCA method is efficient and reliable for FDD under large fluctuations in the VRF operation. Guo et al. (2017b) proposed a modularized PCA method for FDD of the VRF system with different numbers of indoor units (IDUs), and which identifies which IDU is faulty.

The previous studies concluded that different PCA-based methods can detect faulty measurements or performance degradation of the HVAC system. The identification of faulty sensors, however, still requires additional research. This paper expands the approach of Cotrufo and Zmeureanu (2016) for the use of PCA-based method for the identifications of the variables which cause the abnormal performance of space heating. The case study uses measurements of two houses, recently built in Inuvik, NWT, Canada. In the context of this paper, the term "abnormal performance" is not limited to faults; it might be the result of changes of the operation conditions.

PRINCIPAL COMPONENT ANALYSIS (PCA)

The method consists in the transformation of observations (measurements) of j-variables from the Building Automation System (BAS) into a reduced set of k-variables (k < j), which are known as Principal Components (PC), Cotrufo and Zmeureanu (2016). In other words, the observations of j-variables are projected into a k-dimensional PC-based space. The transformed observations in the PC-based space are named scores.

The matrix of training data set X_{tr} (i;j) is composed of i observations for each j variable. The data normalization is performed by using (1):

$$zX_{j,tr} = \frac{X_{j,tr} - \mu_{j,tr}}{\sigma_{j,tr}} \quad (1)$$

where, $zX_{j,tr}$ is the j-column of the normalized training data set; $X_{j,tr}$ is the j-column of the original training data set; $\mu_{j,tr}$ is the average value of the j-column of the original training data set; and $\sigma_{j,tr}$ is the standard deviation of the j-column of the original training data.

The matrix (i, j) of normalized values is then transformed into the matrix (j × j) of coefficients (Q), by using the PCA transformation available in Matlab (2017). The first column of matrix Q, corresponds to the first principal component (PC), the second column corresponds to the second PC, and so on. The first row corresponds to the first variable listed in the original training data set, the second row corresponds to the second variable, and so on.

The matrix (F_{tr}) of the projection of original measurements in the PC-based space, called scores, is created in (2).

$$F_{tr} = zX_{j,tr} \cdot Q \quad (2)$$

The scores corresponding to the normal operation conditions form a cloud of points, which can be surrounded by a threshold or frontier. Different 2D, 3D or n-dimensional models can be used to analytically define the threshold. This study uses an ellipsoid threshold model (3). Those scores outside the threshold correspond to abnormal performance (4).

$$\sum_{j=1}^{k} \frac{{f_{ij}}^2}{(sx_j)^2} = 1 \quad (3)$$

$$\sum_{j=1}^{k} \frac{{f_{ij}}^2}{(sx_j)^2} > 1 \quad (4)$$

where, f_{ij} is the score of the i-observation along the j-principal direction; sx_j is the ellipsoid semi-axis along the j-principal direction; $sx_j = 2 \cdot \sigma_j$; σ_j is the standard deviation of the scores along the j-principal direction.

CASE STUDY

This case study uses the measurements from October 1, 2014 to September 30, 2015 in two semi-detached houses A and B of Inuvik, NWT, Canada at 68.3°N latitude and 9,769 HDD (°C-day). These houses are accounted as sustainable houses which have lower energy demand compared with other conventional houses at warmer locations and lower latitudes. The design thermal resistance of different components of these houses exceed the minimum requirement of the Model National Energy Code of Canada for Houses (MNECH-1997). Thermal resistance of exterior walls in these houses is 8.1 m^2K/W compared with 4.75 m^2K/W in MNECH, 14.1 m2K/W for roof compared with 10.6 m^2K/W in MNECH, and 9.3 m^2K/W for floors compared with 8.1 m^2K/W in MNECH. The air infiltration rate at 50 Pa pressure difference is about 50% of the maximum value of 4.55 ACH required by MNECH.

Most measurements are recorded at 1-minute time step, from which the hourly and daily values are calculated to eliminate the noise in data.

ISBN 978-2-921145-88-6

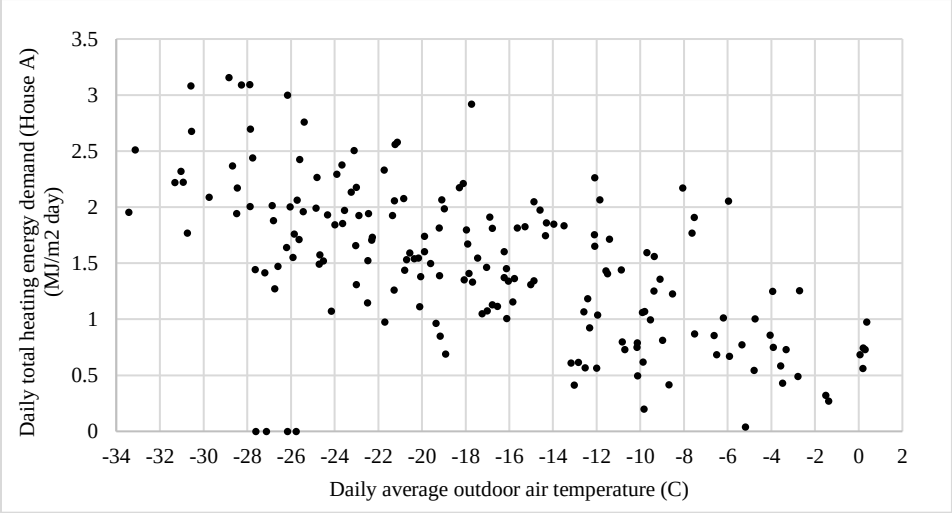

Figure 1: Daily heating energy demand of house A versus the daily average outdoor air temperature, from October 1, 2014 to March 31, 2015

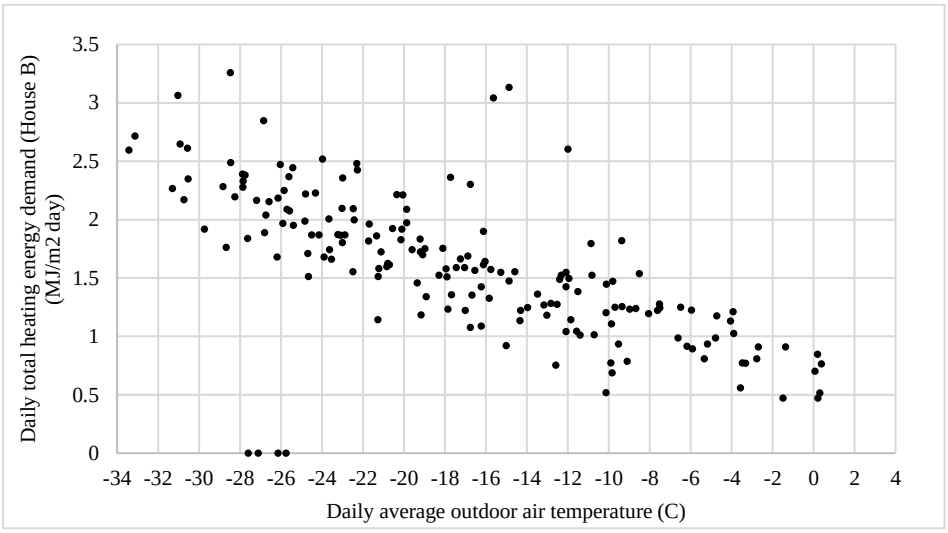

Figure 2: Daily heating energy demand of house B versus the daily average outdoor air temperature, from October 1, 2014 to March 31, 2015

The annual heating energy demand is 98.1 kWh/(m^2 year) for house A, and 101.7 kWh/(m^2 year) for house B, with the average of 99.9 kWh/(m^2 year). The total energy demand for space heating and domestic hot water is 122.4 kWh/(m^2 year), which is satisfied by a natural gas-fired boiler. The annual measured natural gas energy use is 178.2 kWh/(m^2 year).

The daily space heating energy demand in house A from October 2014 to March, 2015 has a larger dispersion (Figure 1) compared with house B (Figure 2).

Since the focus of this paper is the presentation of PCA method, and not on the analysis of the annual energy performance, the authors selected December 2014 as the reference month of the heating season. They used the measurements of December 2014 (Table 1) related to the space heating for the training of the ellipsoid threshold model. The variables that might affect the heating energy demand of February 2015 compared with the reference month of December 2014 are then identified.

ISBN 978-2-921145-88-6

Table 1: Measured variables of space heating for houses A and B

Variables	Unit	Symbol
Water flow rate (House A)	L/min	F_{hA}
Water flow rate (House B)	L/min	F_{hB}
Supply water temperature (House A)	°C	$T_{supplyA}$
Supply water temperature (House B)	°C	$T_{supplyB}$
Return water temperature (House A)	°C	$T_{returnA}$
Return water temperature (House B)	°C	$T_{returnB}$

Transformation of observations

After the data normalization of the hourly values, a Q matrix for space heating is generated (5).

$$Q = \begin{pmatrix} 0.4540 & -0.1762 & -0.3506 & 0.6398 & -0.4686 & -0.1053 \\ -0.4144 & 0.3625 & 0.3404 & 0.7311 & 0.2131 & -0.0332 \\ 0.4655 & 0.0184 & 0.5559 & -0.0627 & 0.0952 & -0.6789 \\ 0.4716 & -0.0202 & 0.4949 & 0.0714 & 0.0287 & 0.7255 \\ 0.4178 & 0.3690 & -0.4548 & 0.0648 & 0.6914 & 0.0152 \\ 0.0920 & 0.8371 & -0.0211 & -0.2074 & -0.4971 & 0.0180 \end{pmatrix} \quad (5)$$

where the rows correspond to the variables listed in Table 1, and the columns correspond to each principal component. For instance, $Q_{11} = 0.4540$ is the coefficient of water flow rate for space heating of house A that is used by the first principal component.

Previous studies by Cotrufo and Zmeureanu (2016), Morrison and Donald (1976), Ladd and Driscoll (1980), and Jolliffe (1986) suggested that the number of PCs should be selected in such a way to explain at least 75% to 90% of minimum cumulative variance in the initial data set. This study uses the first two PCs (k = 2) that contain about 89% of cumulative variance for the space heating (Figure 3), and also facilitates the 2D graphical representation of the PC-based space.

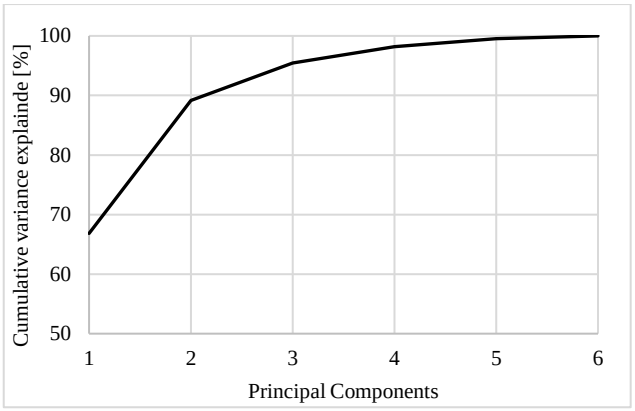

Figure 3: Cumulative variance in the initial data set of December 2014 versus the number of PCs for space heating

Ellipsoid threshold model

Figure 4 shows the scores distribution and the threshold model in the PC-based space, for the space heating of houses A and B, by using the first two PCs (PC#1 and PC#2).

The scores inside the ellipsoid indicate the normal operation, while those outside the ellipsoid are identified as outliers that correspond to abnormal performance. Out of total number of 406 scores for space heating, there are 56 outliers or about 14% in the PC#1-PC#2 space.

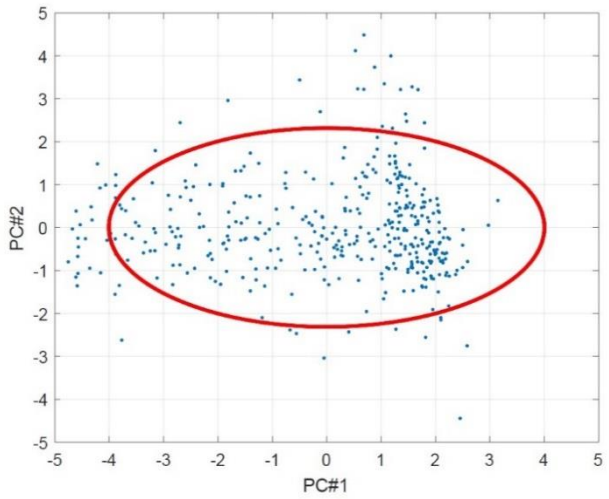

Figure 4: Scores distribution in the PCs-based space (PC#1 and PC#2) for space heating of houses A and B in December 2014

Variables identification

Once the outliers are identified in the PC-based space, the next phase is the identification of those variables from Table 1 that might generate the outliers.

ISBN 978-2-921145-88-6

For each variable in Table 1 the following steps are undertaken, using as an example, $T_{supplyB}$, which is the fifth variable in Table 1 and Q matrix (5):

(1) In the PC-based space (PC#1 – PC#2) (Figure 5), a line is drawn through the origin and the point P of coordinates; Q_{51} = 0.4178 (i.e., for PC#1) and Q_{52} = 0.3690 (for PC#2);
(2) The Euclidean distance between each outlier (e.g., point S) and the zero-value of $T_{supplyB}$ axis is calculated (6).
(3) The Euclidian distance is then calculated for each outlier with respect to all other variables (Table 1) for space heating of houses A and B.
(4) The variable with the highest Euclidean distance for each outlier indicate the variable that might have the highest impact on the abnormal performance.

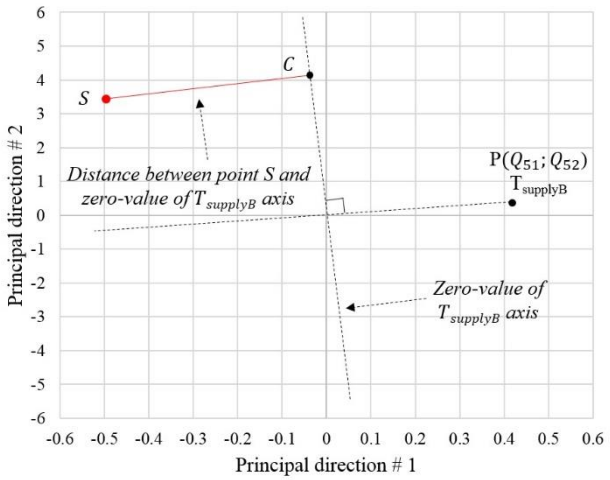

Figure 5: Euclidean distance of the outlier (S) from the zero-value of $T_{supplyB}$ axis in the PC#1-PC#2 space for space heating

The Euclidean distance $\overline{SC}$ of the outlier S is calculated as follows;

$$\overline{SC} = \sqrt{(x_s - x_c)^2 + (y_s - y_c)^2} \quad (6)$$

Where:

$$x_C = \left(\frac{y_p}{x_p}.\ x_s - y_s\right).\ \frac{x_p . y_p}{x_p^2 + y_p^2} \quad (7)$$

$$y_C = -\frac{x_p}{y_p}.x_c \quad (8)$$

where, $\overline{SC}$ is the Euclidean distance, x_C and y_C are the coordinates of the projection of outlier S on the zero-value axis of $T_{supplyB;}$ x_S and y_S are the coordinates of the outlier S in the PC#1 - PC#2 space; $x_P = Q_{5,1}$ and $y_P = Q_{5,2}$ are the coordinates of point P from the Q matrix.

The number of occurrences of the first highest Euclidean distance for each variable (Table 2) reveal that the measurements of the temperature of supply water for space heating of house A ($T_{supplyA}$), and of the return temperature of water for space heating for house B ($T_{returnB}$) are responsible for about 27% and 29% of outliers.

Table 2: Variables with the highest impact on the outliers for houses A and B in December 2014

Variables	**December 2014**	
	1st highest Euclidean distance	
F_{hA}	6	10.7%
F_{hB}	8	14.3%
$T_{supplyA}$	15	26.8%
$T_{returnA}$	5	8.9%
$T_{supplyB}$	6	10.7%
$T_{returnB}$	16	28.6%
Total outliers	56	100%

Application of the trained ellipsoid threshold model on new data set

The trained ellipsoid threshold model developed with data of December 2014, considered to be the reference month, is now applied to measurements of February 2015, called the application data set. The normalized values of the new data set are calculated with (9).

$$zX_{j.ap} = \frac{X_{j.ap} - \mu_{j,tr}}{\sigma_{j,tr}} \quad (9)$$

where, $zX_{j.ap}$ is the j-column of the normalized application data set, $X_{j.ap}$ is the j-column of the new data set, $\mu_{j,tr}$ is the average value of the j-column of the training data set of December 2014, and $\sigma_{j,tr}$ is the standard deviation of the j-column of the training data set of December 2014.

The matrix (F_{ap}) of scores is calculated with (10) and displayed in Figure 6.

$$F_{ap} = zX_{ap}.Q \quad (10)$$

where: Q matrix is given in Equation 5 for the training data set of December 2014.

ISBN 978-2-921145-88-6

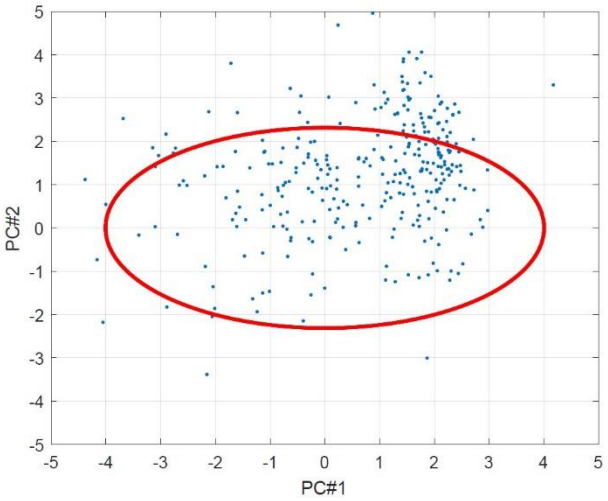

Figure 6: Scores distribution in the PCs-based space (PC#1 and PC#2) for space heating of houses A and B in February 2015, compared with the trained ellipsoid threshold model from data of December 2014

Overall the measurements of February 2015 led to 94 outliers compared with only 56 in December 2014. When the first Euclidean distance is considered (Table 3), the supply and return water temperature for space heating of house B ($T_{supplyB}$) and ($T_{returnB}$) are responsible for about 46% and 37% of outliers, respectively. The outliers could be generated by faults or by changes of the operation conditions due to changes of thermostat set point, or changes due to number and activities of people inside the house.

Table 3: Variables with the highest impact of the outliers for houses A and B in February 2015

Variables	February 2015	
	1st highest Euclidean distance	
F_{hA}	3	3%
F_{hB}	11	12%
$T_{supplyA}$	1	1%
$T_{returnA}$	1	1%
$T_{supplyB}$	43	46%
$T_{returnB}$	35	37%
Total outliers	94	100%

DISCUSSION

The use of measurements from two identical houses enables the additional verification of method proposed by Cotrufo and Zmeureanu (2016). The water temperatures $T_{supplyB}$ and $T_{returnB}$ are identified by the PCA method as the variables that generated the outliers in the PC-based space in February 2015. The outliers identified by the PCA method are displayed on the graphs of hourly supply water temperature (Figure 7) and return water temperature (Figure 8). Those outliers are displayed at the border of the data clouds. The use of 2nd highest Euclidian distance in the PCA method might reveal other abnormal values.

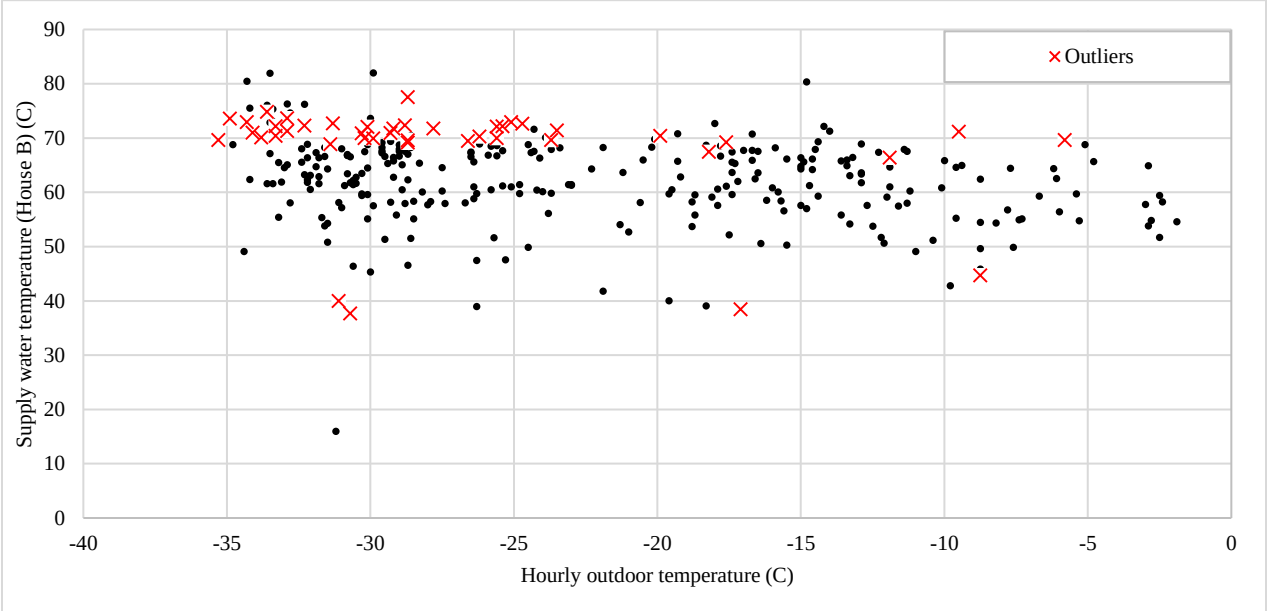

Figure 7: Hourly supply water temperature for house B versus hourly outdoor temperature in February 2015

ISBN 978-2-921145-88-6

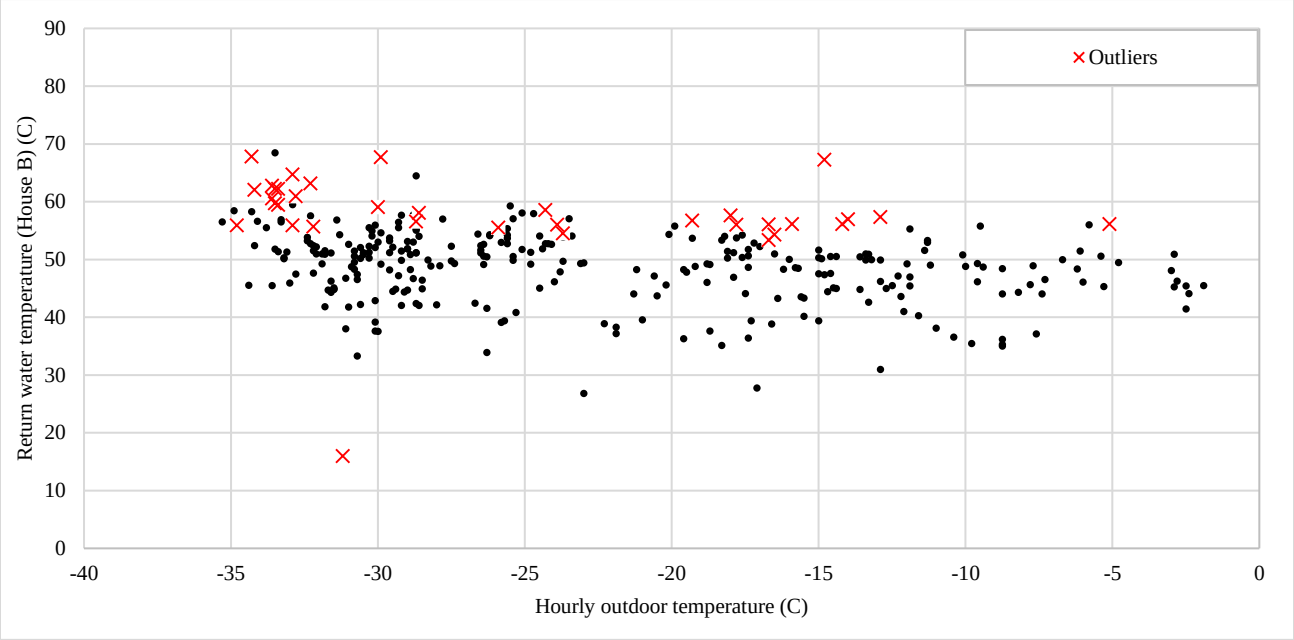

Figure 8: Hourly return water temperature for house B versus hourly outdoor temperature in February 2015

To validate the PCA results, a modified data set is derived in which the abnormal values of those two water temperatures, $T_{supplyB}$ and $T_{returnB}$ of February 2015, which are outside the range given by (11) are removed from the data file. Then the modified application data set of February 2015 is used by the PCA method.

$$\mu_{j,tr} - 2 \cdot \sigma_{j,tr} < X_{j,app} < \mu_{j,tr} + 2 \cdot \sigma_{j,tr} \quad (11)$$

As a result of removing the abnormal values of $T_{supplyB}$ and $T_{returnB}$ the number of outliers is reduced in the PC#1-PC#2 based space (Table 4). This result proves that the two temperatures, $T_{supplyB}$ and $T_{returnB}$ are the cause of the outliers identified by the PCA method.

Table 4: Number of outliers for each variable in February 2015 with original and modified data sets

Variables	1st highest Euclidean distance	
	Original data set	Modified data set
F_{hA}	3	3
F_{hB}	11	9
$T_{supplyA}$	1	0
$T_{returnA}$	1	1
$T_{supplyB}$	43	5
$T_{returnB}$	35	9
Total outliers	94	27

CONCLUSION

The PCA-based method for the detection of outliers, and the identification of variables was applied in this study to the measurements from two semi-detached houses, recently built in Inuvik, NWT, Canada. Measurements of December 2014 related to the space heating are used as reference values for the training of the ellipsoid threshold model. The supply and return water temperatures for house B were identified as the main sources of the outliers of the heating energy demand in February 2015. The identification of variables with abnormal values were compared with two different approaches: the graphical representation of hourly values, and the use of a modified data set. The two approaches identified the water temperatures, $T_{supplyB}$ and $T_{returnB}$ as the main sources of outliers in February 2015 by the PCA method.

Future work would focus on the verification of the proposed method for complex HVAC systems with many correlated physical variables, and the comparison of the identified faults with physical faults.

The PCA-based method for fault detection and identification should be implemented in the BAS for the ongoing commissioning of heating systems, to help the building operator to take timing actions.

ACKNOWLEDGEMENT

The authors acknowledge the financial support from Natural Sciences and Engineering Research Council of Canada and the Faculty of Engineering and Computer Science of Concordia University. The authors also

acknowledge the collaboration of L. Robinson, S. Outlet, N. Walker, L. Azzolini from Arctic Energy Alliance, J. M. Robson from Northwest Territories Housing Corporation, and Aurora Research Institute.

REFERENCES

Beghi, A., R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, F. Simmini. 2016. Data-driven Fault Detection and Diagnosis for HVAC water chillers. *Control Engineering Practice.* 53, 79-91.

Cotrufo, N. and R. Zmeureanu. 2016. PCA-based method of soft fault detection and identification for theongoing commissioning of chillers. *Energy and Buildings.* 130, 443-452.

Draper, N.R. and H. Smith. 1981. Applied Regression Analysis, 2nd ed. *John Wiley & Sons, Inc*. New York, USA.

Gajjar, S., M. Kulahci, A. Palazoglu. 2016. Use of Sparse Principal Component Analysis (SPCA) for Fault Detection. *IFAC-PapersOnLine.*49(7), 693-698.

Guo, Y., G. Li, H. Chen, Y. Hu, H. Li, J. Liu, M. Hu, W. Hu. 2017b. Modularized PCA method combined with expert-based multivariate decoupling for FDD in VRF systems including indoor unit faults. *Applied Thermal Engineering.* 115, 744-755.

Guo, Y., G. Li, H. Chen, Y. Hu, H. Li, L. Xing, W. Hu. 2017a. An enhanced PCA method with Savitzky-Golay method for VRF system sensor fault detection and diagnosis. *Energy and Buildings.* 142, 167-178.

Hadley, D.L. and S.D. Tomich. 1986. Multivariate statistical assessment of meteorological influence on residential space heating. *ACEEE* 9.

Hu, Y., G. Li, H. Chen, H. Li, J. Liu. 2016a. Sensitivity analysis for PCA-based chiller sensor fault detection. *International journal of refrigeration*. 63, 133-143.

Hu, Y., H. Chen, G. Li, H. Li, R. Xu, J. Li. 2016b. A statistical training data cleaning strategy for the PCA-based chillersensor fault detection, diagnosis and data reconstruction method. *Energy and Buildings.*112, 270-278.

Hu, Y., H. Chen, J. Xie, X. Yang, C. Zhou. 2012. Chiller sensor fault detection using a self-Adaptive Principal Component Analysis method. *Energy and Buildings*. 54, 252-258.

Jia, Y. and T.A. Reddy. 2003. Characteristic Physical Parameter Approach to Modeling Chillers Suitable for Fault Detection, Diagnosis, and Evaluation. *Solar Energy Engineering* 125(3), 258-265.

Jolliffe, I.T. 1986. Principal Component Analysis, *Springer*. New York, USA.

Katipamula, S. and M.R. Brambley. 2005. Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I. *HVAC & R Research* 11(1), 3-26.

Ladd, J.W. and D.M. Driscoll. 1980. A comparison of objective and subjective means of weather typing: an example from west Texas. *American Meteorological Society.*19, 691-704.

Matlab. Mathworks/Statistics and Machine Learning Toolbox. 2017: http://www.mathworks.com/help/stats/pca.html?searchHighlight=PCA&s_tid=doc_srchtitle.

Morrison, P. and F. Donald. 1976. Multivariate statistical method, 2nd ed., New York, USA.

National Research Council of Canada. 1997. Model National Energy Code of Canada for Buildings. Ottawa, Canada.

Roth, K., D. Westphalen, J. Brodrick. 2008. Ongoing Commissioning. *ASHRAE Journal* 50(3), 66-71.

Xu, X., F. Xiao, S. Wang. 2008. Enhanced chiller sensor fault detection, diagnosis and estimation using wavelet analysis and principal component analysis methods. *Applied Thermal Engineering* 28, 226-237.

ISBN 978-2-921145-88-6