World Academy of Science, Engineering and Technology
International Journal of Architectural and Environmental Engineering
Vol:13, No:1, 2019

# Automatic Thresholding for Data Gap Detection for a Set of Sensors in Instrumented Buildings

Houda Najeh, Stéphane Ploix, Mahendra Pratap Singh, Karim Chabir, Mohamed Naceur Abdelkrim

Open Science Index, Architectural and Environmental Engineering Vol:13, No:1, 2019 publications.waset.org/10009912.pdf

***Abstract*****—**Building systems are highly vulnerable to different kinds of faults and failures. In fact, various faults, failures and human behaviors could affect the building performance. This paper tackles the detection of unreliable sensors in buildings. Different literature surveys on diagnosis techniques for sensor grids in buildings have been published but all of them treat only bias and outliers. Occurences of data gaps have also not been given an adequate span of attention in the academia.

The proposed methodology comprises the automatic thresholding for data gap detection for a set of heterogeneous sensors in instrumented buildings. Sensor measurements are considered to be regular time series. However, in reality, sensor values are not uniformly sampled. So, the issue to solve is from which delay each sensor become faulty?

The use of time series is required for detection of abnormalities on the delays. The efficiency of the method is evaluated on measurements obtained from a real power plant: an office at Grenoble Institute of technology equipped by 30 sensors.

***Keywords*****—**Building system, time series, diagnosis, outliers, delay, data gap.

## I. Introduction

FAULT detection and diagnosis is well proven and known tool for several industrial process like aerospace, automotive, nuclear and process industry and there are various techniques available, considering the kind of applications [5], [11], [12]. Over the past few years growing research interests have been found to apply FDD techniques for buildings [6], [7].

Nowadays, buildings are considered as highly dynamics and complex systems as well. They include HVAC systems, sophisticated controllers, a large number of sensors and energy management systems. Various faults, failures and unplanned events could cause a discrepancy in building performance and a poorly maintained building causes discomfort to occupants. Sensor measurements are often inaccurate and many external factors can interfere with the sensing device. Then, it is important to ensure the accuracy of sensor data before it is used in a decision making process [17].

Recent literature contributes methods for the detection and classification of sensor data faults [13], [14]. Fault classification techniques vary: several existing fault taxonomies use different criteria for categorising a fault. For exemple, [10] gives extensive taxonomies of data faults that include a definition, the cause of the fault, its duration and its impact onto sensed data. However, most researchers in the literature are interested only by the following known fault types: drift, outliers and bias. Occurences of data gap faults have also not been given an adequate span of attention in the academia. Data gap means an abnormal change in the data delays sending by a sensor.

Conventional fault diagnosis and classification methods usually implement pretreatments to decrease noise and extract some time domain or frequency domain features from raw time series sensor data. Then, some classifiers are utilized to make diagnosis. However, these conventional fault diagnosis approaches do not solve automatic thresholding for heterogeneous sensors and they do not consider the non regular samples of data time series.

Different diagnosis techniques for sensor grids in buildings have been published in the literature. Signal based FDD methods mainly use signals and time series which are obtained from measurements [16]. Clustering is also considered as an important application area for many fields including data mining and statistical data analysis [3]. Clustering has been formulated in various ways in diagnosis of sensor faults [9]. However, all of these methods treat only bias and outliers type faults [8]. Also, none of these methods treat the automatic thresholding for heterogeneous sensors which can be very helpful to building managers.

This paper tackles this issue and focuses on developing method for automatic thresholding for automatic data gap detection for heterogeneous sensors in instrumented building. In fact, if the threshold is too high, it may lead to non-detection: the situation is assumed to be normal even though it is not. On the contrary, a threshold that is too low will cause false alarms: the situation is supposed to be abnormal when it is not. Thus, there is the problem of determining from which deviation a fault can potentially be considered. The problem is to find an optimal time threshold that will be the ideal compromise between a non-detection rate and a false alarm rate. An algorithm based on time series has been adapted to an office setting which is a sensor test bed with a large number of ENOCEAN sensors.

This paper is organized as follow: Section II presents the problem statement of automatic thresholding for heterogeneous sensors (i.e, occurence of data gaps). Section III presents the application of time series in fault diagnosis. Section IV discusses the proposed algorithm for automatic thresholding for data gap detection for a set of sensors in instrumented buildings and Section V analyses the simulation results for an

Houda Najeh, Stéphane Ploix and Mahendra Pratap Singh are with University of Grenoble Alpes, G-SCOP lab, CNRS UMR 5272, Grenoble, France (e-mail: houda.najeh@grenoble-inp.fr, stephane.ploix@grenoble-inp.fr, talk2mpsingh@gmail.com).

Karim Chabir and Mohamed Naceur Abdelkrim are with University of Gabes, National Engineering School of Gabes, Tunisia, Research lab. Modeling, Analysis and Control of Systems (MACS) LR16ES22 (e-mail: karim.chabir@yahoo.fr, naceur.abdelkrim@enig.rnu.tn).

Open Science Index, Architectural and Environmental Engineering Vol:13, No:1, 2019 publications.waset.org/10009912.pdf

office at Grenoble Institute of Technology. Finally, concluding remarks and future works are given in Section VI.

## II. Problem Statement

### A. Test Bed

The testbed is an office in Grenoble Institute of Technology, which accommodates a professor and 3 PhD students. The office has frequent visitors with a lot of meetings and presentations all through the week. The setup for the sensor network includes (see Fig. 1):

- 2 video cameras for recording real occupancy and activities.
- 2 luminosity sensors with different sensitivities
- 4 indoor temperature, for the office and the bordering corridor
- 2 COV+CO2 concentration sensors for office and corridor
- 1 relative humidity sensor
- 4 door and window contact sensors
- 1 motion detector
- 1 binaural microphone for acoustic recordings
- 5 power meters
- outdoor temperature, nebulousness, relative humidity, wind speed and direction, . . . from weather forecasting services
- a centralized database with a web-application for retrieving raw data from different sources continuously

When smart buildings are expanded to make use of multiple sensors, the possibility of time delayed data becomes a reality. Sensor measurements are considered to be statistical time series. When the sensors are in ok states, the delay data have a distributions corresponding to the normal mode of operation and theses distributions change when the sensor is faulty.

Sensor values are not necessarily uniformly sampled. While after pre-processing the sensors report values regularly, reality shows that quite many values are missing. The gaps that as a result exist, are sometimes too small to be visible on a graph.

In general, there is no regularly delayed data for a variable. Delays depend not only on type of the sensor but also on the measured values. The question that arises is from which delay can we say that the sensor become fauly? Automatic thresholding for data gap detection for heterogeneous sensors is a feasible paradigm for the instrumented residential environment.

## III. Time Series Applications in Fault Diagnosis

Over the last two decades, advances in data manipulation techniques and sensor technology have contributed to the widespread application of signal processing concepts for fault diagnosis. However, most of these diagnoses, or data, don't have the level of redundancy presented in the model-based methods. Among other things, it is important to consider the importance of this method. In addition, the data-based method describes the dynamics of the system with an increased level of redundancy.

The strength of time series techniques is their capability to represent an observation in a time invariant parameterized structure. Such invariance creates the possibility to predict the data and, hence offers insight to the sensor behavior. Moreover, their applicability adds to their adequacy for fault diagnosis. These facts represent the driving motivation in pursuing their application for diagnostics.

Under certain normal conditions, sensor measurements have a typical spectrum frequency and any deviation from the frequent characteristics of a signal is connected to an anomaly. The application of a decision procedure makes it possible to detect and locate the faulty sensor. Among the decision procedures applied to a sample of measures are empirical test of crossing of threshold, test of variance, the test of the average [2].

Break detection is a subject related to other classical problems of signal processing, information or statistics, among which we can mention the detection of anomalies in general. The detection of breaks itself corresponds to several problems : detect changes in signal characteristics and locate them, and then be able to analyze segments of the time series individually. The information extracted from it makes it possible to propose a diagnosis. Reference works detail the various existing approaches, the reader can particularly refer to [1] for the detection of a single break mainly in an online context and [4] for a set of parametric methods and their applications.

In the time domain, a main trend exists consisting of constructing a parametric time series model for the observed data considering a fault-free situation. Then, an analysis is performed to the observed residuals between the model prediction and the actual observation in an approach very similar to that pursued in model-based techniques. A wide variety of residual analysis methods have been applied. For example, [15] compared the prediction error of an ARMA model of the observed multivariate signal in both fault-free and fault situation as a decision base to detect sensor faults.

Online approaches are also present in the literature [1]. In on-line approach, samples are taken sequentially and decisions are made based on the observations up to the current time.

If the decision is made from the values of observations directly, then when the observations X(t)= $[x_1(t),\ x_2(t),\ ...,\ x_n(t)]\ '\ \in R_n$ where $R_n$ is the so called stopping region, it is concluded that there is a change in the process.

More often g(t), a function of observation x(t) is designed and the decision is made according to the comparison of g(t) with some threshold value c. This idea can be translated into a stopping rule problem with a standard form

$\{\ \tau$=inf t $\geq$ 1; g(t)$\geq$ c }

$\tau$ is the greatest lower bound, i.e., the first time when g(t ) is greater than c.

The main objective of this work is a time series, consisting of a finite number of successive observations i.e delay data. It is formalized by the vector X = (X1, ..., Xn). In the absence of fault, X is stationary i.e time samples are regular. A break is defined by a change in the sending of data by a sensor. The solution of this issue is discussed in the following section.

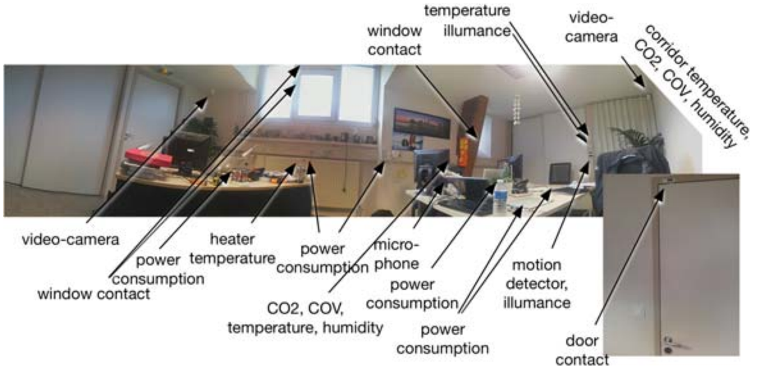

Fig. 1 Test bench

## IV. Proposed Algorithm

The proposed algorithm is a way of describing a data point in terms of its relationship to the average difference between two consecutive data points for the whole data set and standard deviation of the difference between two consecutive data points for the whole data set.

This method was based on the principle that the variation of measurements (i.e delays) should smoothly vary and follow a uniform distribution. The proposed cutoff was the outlier fence, which is defined by the average difference between two consecutive data added by a standard deviation of the difference of the time serie measurements.

The aim of this paper is to process time series of data representing time samples. Outliers on the delay have to be detected. They are defined as data points, which, in the contact of previous and future data points, seem highly improbable.

In the case of normally distributed time samples, it is assumed that, at a given moment, the difference between the current and previous data point i.e. data sent by a sensor (see (1)) is equal to the current and next data point (see (2)).

$$p_{diff_k} = x_k - x_{k-1} \tag{1}$$

$$f_{diff_k} = x_{k+1} - x_k \tag{2}$$

Then, a rule has been fixed at a fixed threshold and is equal to $m_{\Delta x} + \lambda\sigma_{\Delta x}$, with $\lambda$ a configurable parameter. If no outliers is found, reducing the value of $\lambda$ for testing is required.

Since, the majority of points in a distribution are within "$\lambda\sigma_{\Delta x}$" deviations of the average difference. The decision is "abnormal delay" when the value of delay is this threshold, otherwise the decision is "normal case".

These points have been detected as the ones that follow the equations simultaneously:

$$|p_{diff_k}| > m_{\Delta x} + \lambda\sigma_{\Delta x} \tag{3}$$

$$|f_{diff_k}| > m_{\Delta x} + \lambda\sigma_{\Delta x} \tag{4}$$

$$p_{diff_k} . f_{diff_k} < 0 \tag{5}$$

where:

- $x_k$: value of the variable at time sample
- $p_{diff_k}$: difference between the current and previous data point
- $f_{diff_k}$: difference between the current and next data point
- $m_{\Delta x}$: average difference between two consecutive data points for the whole dataset
- $\sigma_{\Delta x}$: standard deviation of the difference between two consecutive data points for the whole dataset
- $\lambda$: configurable parameter

## V. Simulation Results

A data set covering 1 month from 01-March-2016 have been used to detect the abnormalities on the delay and subsequently the data gaps from raw measurements of sensors.

The following figures show the detection of outlier with $\lambda = 5$ (Fig. 2) as well as the data gaps (Fig. 3) for the Toffice-wall sensor installed in the H358 office in Grenoble INP.

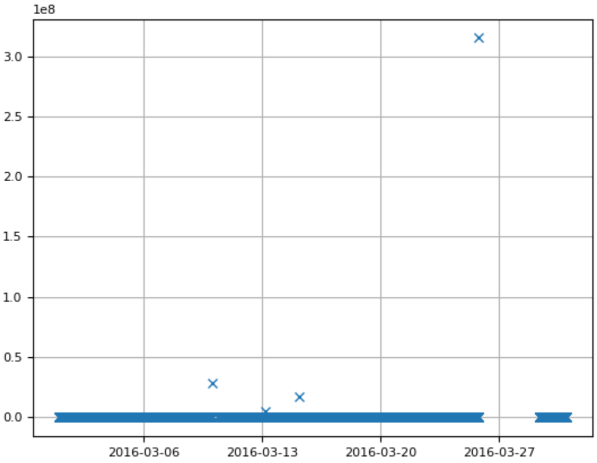

Fig. 2 Detection of outliers

Each outlier corresponds to a non-healthy period.

The data gaps are detected in the following intervals:

[((2016, 3, 10, 2, 19, 17), (2016, 3, 10, 10, 11, 28)),
((2016, 3, 15, 5, 24, 56), (2016, 3, 15, 10, 6, 6)),
((2016, 3, 25, 18, 1, 6), (2016, 3, 29, 9, 33, 12))]

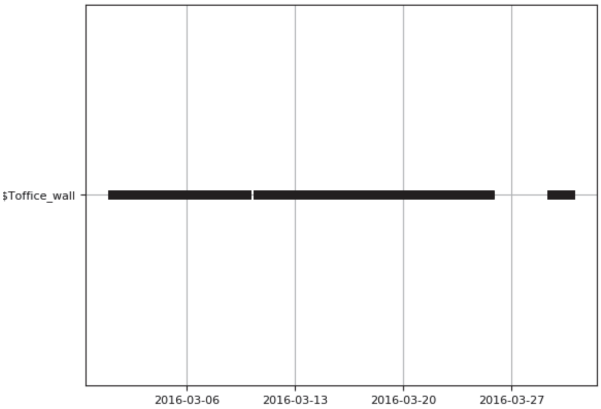

Fig. 3 Detection of data gaps

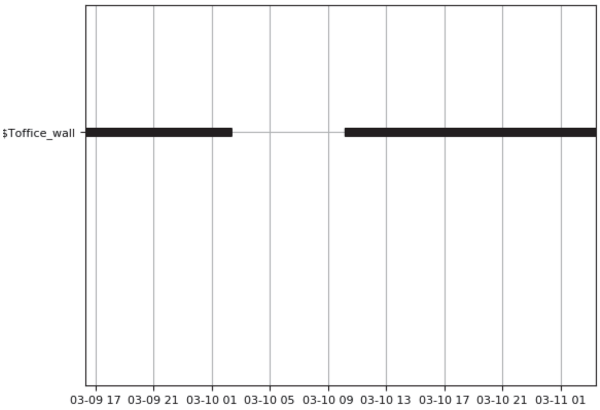

Fig. 4 Interval 1

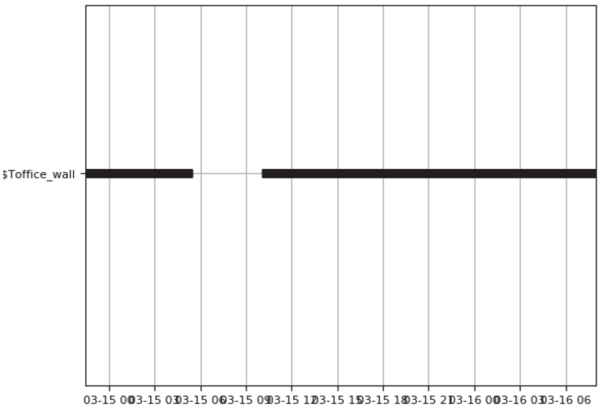

Fig. 5 Interval 2

A zoom on the result of the figure (Fig. 3) shows the accuracy of these intervals (see Figs. 4-6)

The configurable parameter $\lambda$ has an effect on the outlier detection rate and consequently on the determination of unsound periods of sensors (see Table I).

**Interpretation**:

A well-known advantage of the proposed method is robustness in the presence of outliers and more precisely to the different $\lambda$ configurable parameter change profiles. For example, consider the calculation of the average difference between two consecutive data points for the data set. If the standard deviation of the difference between two consecutive data point is linked with a very high value of $\lambda$, the average difference is largely changed. On the other hand, it can be quite substantially modified for a low value of $\lambda$.

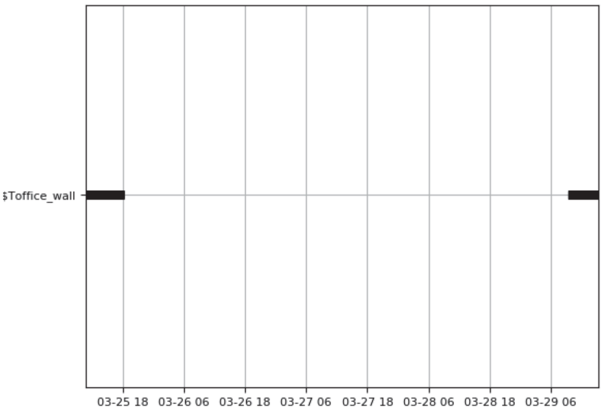

Fig. 6 Interval 3

TABLE I
EFFECT OF $\lambda$ ON THE OUTLIER DETECTION RATE

| $\lambda$ | detection rate | intervals |
|---|---|---|
| 12 | 1 | ((2016, 3, 25, 18, 1, 6), (2016, 3, 29, 9, 33, 12)) |
| 5 | 3 | ((2016, 3, 10, 2, 19, 17), (2016, 3, 10, 10, 11, 28)),<br>((2016, 3, 15, 5, 24, 56), (2016, 3, 15, 10, 6, 6)),<br>((2016, 3, 25, 18, 1, 6), (2016, 3, 29, 9, 33, 12)) |
| 0.3 | 4 | ((2016, 3, 10, 2, 19, 17), (2016, 3, 10, 10, 11, 28)),<br>((2016, 3, 13, 5, 29, 42), (2016, 3, 13, 6, 55, 42)),<br>((2016, 3, 15, 5, 24, 56), (2016, 3, 15, 10, 6, 6)),<br>((2016, 3, 25, 18, 1, 6), (2016, 3, 29, 9, 33, 12)) |

Consider scenarios 2 and 3 in which the data (whose distribution has been described by $\lambda = 5$ and $\lambda = 0.3$ respectively) are progressively contaminated by an increasing amount of outliers. The number of outliers on the delay for scenario 2 is 3 while that of scenario 3 is 4. The limitation of this method is that we can not conclude on the optimal $\lambda$ parameter for the detection of outliers.

The following figure (Fig. 7) shows the evolution of data gaps for the sensor grid in the H358 office during the year 2016.

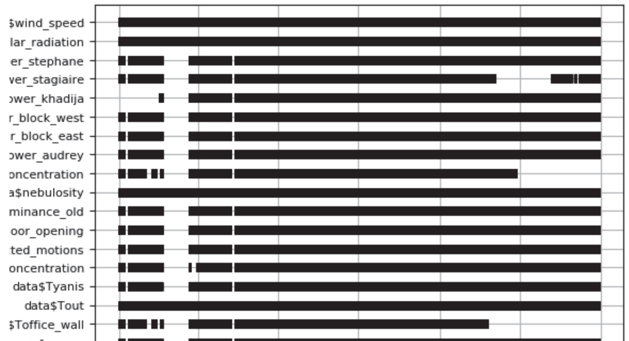

Fig. 7 Evolution of raw measurements

## VI. Conclusion

With the emergence of new building issues, diagnosis is become more and more complex and new methods for design are required. This paper presents a method for automatic thresholding for data gap detection for heterogeneous sensors in instrumented buildings. The solution is based on the use of time series for detection of abnormalities on the delay. This method has been applied to a small selection of sensors.

The limitation of the propsed method is that, it depends on the configurable parameter $\lambda$. Moreover, proposed methodology could provide the first explanation of automatic thresholding for data gap detection for hererogeneous sensors to help building reserachers.

Future works will be around the development of an off-line algorithm for the determination of an optimal configurable parameter $\lambda$. More improvements could be made also for testing the efficiency of this method on a bigger selection of sensors, so a large time series.

## Acknowledgment

The authors would like to thank the Region Rhône-Alpes for financial support of this research work.

## References

[1] Basseville, M., Nikiforov, I. V., et al. (1993). Detection of abrupt changes: theory and application, volume 104. Prentice Hall Englewood Cliffs.

[2] Basseville, M. (1988). Detecting changes in signals and systems: a survey. Automatica, 24, 309-326.

[3] Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer, Berlin, Heidelberg.

[4] Chen, J., & Gupta, A. K. (2012). Parametric statistical change point analysis: With applications to genetics, medicine, and finance. Basel, Switzerland: Springer Science+Business Media, LLC.

[5] Giap, Q.-H., Ploix, S., and Flaus, J.-M. (2009). Managing Diagnosis Processes with Interactive Decompositions. Milan, Italy.

[6] Greiner, R., Smith, B. A., and Wilkerson, R. W. (1989). A correction to the algorithm in reiter's theory of diagnosis. Artificial Intelligence, 41(1), 79-88.

[7] Guannan Li, Yunpeng Hu, Huanxin Chen, Haorong Li, Min Hu, Yabin Guo, Shubiao Shi, Wenju Hu (2016). A Sensor Fault Detection and Diagnosis Strategy for Screw Chiller System Using Support Vector Data Description-based D-statistic and DV-contribution plots. Energy and Buildings.

[8] Llanos, C. E., Sanchéz, M. C., & Maronna, R. A. (2017). A robust methodology for the sensor fault detection and classification of systematic observation errors. In Computer Aided Chemical Engineering (Vol. 40, pp. 1525-1530). Elsevier.

[9] Li, G., & Hu, Y. (2018). Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. Energy and Buildings.

[10] Ni, K., Ramanathan, N., Chehade, M. N. H., Balzano, L., Nair, S., Zahedi, S., ... & Srivastava, M. (2009). Sensor network data fault types. ACM Transactions on Sensor Networks (TOSN), 5(3), 25.

[11] Pomorski, D., Perche, P., (2001). Inductive learning of decision trees: application to fault isolation of an induction motor. Eng. Appl. Artif. Intell. 14, 155-166 .

[12] Ploix, S. (2009). Des systèmes automatisés aux systémes coopérants application. au diagnostic et à la gestion énergétique.

[13] Ren, J. Y., Chen, C. Z., He, B., & Wang, B. (2008). Application of SiC and SiC/Al to TMA optical remote sensor. Optics and Precision Engineering, 16(12), 2537-2543.

[14] Shi, L., Cheng, P., & Chen, J. (2011). Sensor data scheduling for optimal state estimation with communication energy constraint. Automatica, 47(8), 1693-1698.

[15] Upadhyaya, S.K., Rand, R.H. and Cooke, J.R., 1983. A mathematical model of the effects of CO 2 on stomatal dynamics. J. Theor. Biol., 101: 415-440.

[16] Zhang, R., Peng, Z., Wu, L., Yao, B., & Guan, Y. (2017). Fault diagnosis from raw sensor data using deep neural networks considering temporal coherence. Sensors, 17(3), 549.

[17] Zhang, Y., Meratnia, N., & Havinga, P. J. (2010). Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 12(2), 159-170.

**Houda Najeh** received Eng. degree in electrical-automatic engineering from the National Engineering School of Gabes (ENIG), Tunisia in 2015. Since that, she is a PhD student in the research laboratory for Sciences for the Production, Optimization and Production of Grenoble (G-SCOP) and the research laboratory of Modelling, Analysis and Control Systems of the National Engineering School of Gabes (MACS), Tunisia. Her research interests include fault detection and diagnosis in building systems.

**Stéphane Ploix** is professor at the Grenoble Institute of Technology in the GSCOP lab. After an engineer degree in mechanics and electricity, in 1998 he obtained a Ph.D. from Institute National Polytechnique de Lorraine in control engineering and signal processing and the HDR degrees in 2009. He is a specialist in supervision, monitoring and diagnosis, and his studies focus on human-machine cooperative mechanisms. He is involved in different industrial projects dealing with the supervision of distributed plants, the diagnosis of human skills, iterative diagnosis tool for companies and power management in building.

**Mahendra Pratap Singh** Dr. Mahendra Singh earned his Ph.D. degree from University of Grenoble, France where he worked on the development of Reactive building management with the key focus on fault diagnosis. He is currently a post-doctoral researcher at Maersk Mc-Kinney Moller, University of Southern Denmark

**Karim Chabir** received the B.Eng. degree in electrical engineering and automatic engineering from The Higher School of Sciences and Technology of Tunis (ESSTT), Tunisia in 2003, the M. Sc. degree in automatic and intelligent techniques from the National Engineering School of Gabes, Tunisia in 2006, and the Ph.D. degree in automatic control from Henri Poincare University, France in 2011. The research works were carried out at the Research Centre for Automatic Control of Nancy (CRAN) and at the Research Unit of Modelling, Analysis and Control Systems of the National Engineering School of Gabes. He was a member of the dependability and system diagnosis group (SURFDIAG). He was a secondary school teacher of Gabes from 2003 to 2007, where he was also an assistant professor in the Faculty of Science of Gabes from 2007 to 2011. He is now assistant professor at the National Engineering School of Gabes (ENIG), Tunisia. His research interests include model-based fault diagnosis and fault-tolerant control.

**Mohamed Naceur Abdelkrim** received the B. Sc. degree in electrical construction in 1980, and the M. Sc. degree in electrical construction in 1981 from the High Normal School of Technical Education of Tunis, Tunisia. He also received the Ph.D. degree in automatic control from the National School of Engineers of Tunis, Tunisia in 2003. He began teaching in 1981 at the National School of Engineers of Tunis and since 2003, he has been a professor of automatic control at the National School of Engineers of Gabes, Tunisia. He is currently the head of the research unit on Modeling, Analysis and Control of Systems (MACS), Tunisia. His research interests include diagnosis, optimal control, robust control and robotics.